INTERNATIONAL LITERACY ASSOCIATION

Writing Machines: Formative Assessment in the Age of Big Data

T. Philip Nichols, Robert Jean LeBlanc, David Slomp

It was a decade ago, in a professional development workshop, when Phil (first author) first encountered automated writing assessment in classrooms. A middle school language arts teacher, fresh from a National Writing Project summer institute, he sat in stunned silence as the presenter touted software solutions to streamline essay grading by offloading the task to a computer. Phil knew such technologies existed, of course, but he associated them with large-scale testing, not the day-to-day rhythms of classrooms. Applying the techniques of the former to the latter felt wrong—an abdication of some fundamental principle of what it means to teach writing. A computer could analyze students' words, he thought, but it couldn't replicate the personalized response of a caring, human reader.

Or could it? In the years since Phil's uneasy introduction to automated assessment, a lucrative industry has emerged, premised on this possibility. Today's software solutions—platforms such as WriteToLearn (Pearson), Redbird (McGraw-Hill), and Writing Mentor (ETS)—no longer promise just to streamline assessment but to provide adaptive feedback on students' in-process work. Unlike the machine scoring of the past, these platforms use natural language-processing technologies to collect and analyze vast stores of information, or big data, to deliver immediate, personalized responses to student writing.

On paper, these adaptive technologies appear to address Phil's, and many other teachers', reservations about automated assessment. The National Council of Teachers of English position statement on machine scoring (NCTE Task Force on Writing Assessment, 2013), for example, warns against devices which encourage an impersonal view of writing that privileges rigid summative assessment over formative feedback. Yet, today's technologies do the opposite, prioritizing data-rich responsiveness on early drafts of student writing. Even more, in doing so, today's technologies can free educators to devote more time to forms of student support that can't be ceded to machines—thus, they claim, making writing instruction even more humanizing and personal than before.

The promises of adaptive writing assessment are alluring, but as with any technological solution, its potential must be weighed against its realities. In what follows, we offer a critical examination of one such platform, Turnitin's Revision Assistant. We suggest that the adaptiveness of this and similar technologies is more complicated than it appears and that its impacts raise important questions for educators, not only about the purpose and form of writing assessment itself but also about the ethics and potential problems of introducing big-data technologies into everyday instruction.

Disrupting Writing Instruction

Once termed "the neglected *R*," writing has emerged over the last decade as a central concern for education policy and technological innovation, leading to massive investments to address what some have called a burgeoning crisis in writing achievement (Graham, 2019; National Commission on Writing in America's Schools and Colleges, 2003). Adaptive technology developers have been key recipients of this largesse. Because writing instruction is complex and time-consuming, investors have suggested that it is ripe for technological disruption: By automating its most inefficient facets, the argument goes, students will get more feedback on more writing, and teachers can devote their time to personalized lessons and conferences rather than marking up papers.

T. PHILIP NICHOLS is an assistant professor in the Department of Curriculum and Instruction at Baylor University, Waco, TX, USA; email phil_nichols@baylor.edu.

ROBERT JEAN LEBLANC is an assistant professor in the Faculty of Education at the University of Lethbridge, AB, Canada; email robert.leblanc@uleth.ca.

doi: 10.1002/jaal.1160

The market for such writing assessment platforms is competitive. Whereas some focus on particular facets of composition, such as sentence complexity (Hemingway) or grammar (NoRedInk), others have grander ambitions to fundamentally transform how writing is taught and assessed in schools. We focus here on one platform in this latter category, Revision Assistant, which was acquired by Turnitin in 2015 and now reaches millions of students in over 10,000 K–16 institutions worldwide.

Revision Assistant

Developed through federal and philanthropic grants from the Institute of Education Sciences and the Bill & Melinda Gates Foundation, Revision Assistant uses machine-learning algorithms and natural language processing to identify and re-create the features of formative feedback that teachers usually give on drafts of student work. How it does this is, admittedly, complex. To break down the moving pieces, we outline in Figure 1 the platform's hypothesized theory of action as we have constructed it through a close reading of the promotional and academic literature published by Turnitin.

Generating adaptive feedback starts long before teachers and students ever use the platform in a classroom. The process begins with "training data" being fed into Revision Assistant's algorithm—effectively teaching the machine how to behave (A–D). This is why such platforms are sometimes called machine-learning or artificial intelligence technologies.

As a writing-oriented platform, teaching the algorithm means providing it with a set of essays (E) written in response to a particular prompt (C), along with human-produced feedback on those samples (D). As the algorithm processes this information, it learns to adapt itself to a wide range of possible responses related to, for instance, development, organization, language, and clarity. Once trained, the data set and its associated prompt are added to Revision Assistant's database of writing assignments (A).

This brings us to the classroom. After selecting an assignment from its database (F), students can compose directly in Revision Assistant's word-processing interface (G). As they write, the algorithm analyzes their work in relation to its training data to determine what feedback would be most relevant (H–J). For instance, when students complete a draft, they may be given highlighted suggestions for sections in need of stronger development or better organization (see Figure 2)—responses learned from previous human feedback embedded in the platform's training data. Students then return to these sections for revision and repeat the process until the assignment is ready to be submitted to the teacher.

The theory is if students revise multiple times before a teacher interacts with their work, then educators can focus on more substantive suggestions (K). In effect, the platform outsources the early, iterative feedback to the algorithms—pushing students to do more writing (M and N), while directing teachers' energies toward forms of support that humans are better at providing (L and O). The long-term goal is to reinforce the value of revision, leading to improved outcomes for students (R–T).

Importantly, this forms a loop. Data generated from students' writing (R) can theoretically be fed back into this system (C). In other words, the entire process can be thought of as in motion: At no point is there a stable, aggregated essay against which student work is being compared; the data generated in classrooms can be cycled back into the training data, further refining how the algorithm responds to students' writing (A and B).

Adaptive Writing Assessment: Promises and Realities

On the surface, platforms like Revision Assistant address some real, on-the-ground needs. For overworked English teachers, often reviewing the work of hundreds of different students each semester and facing overwhelming pressures to improve writing achievement, the promises of adaptive assessment technologies are enticing. Yet, the spread of these platforms also raises critical questions about the means and ends of writing, assessment, and technology in schools. In what remains, we examine the realities of adaptive writing assessment and highlight three key concerns: assessment quality, platform imperatives, and the ethics of data enrollment.

Assessment Quality

There are three types of evidence commonly used to determine the quality of an assessment: validity, reliability, and fairness. Validity asks, Are the interpretations and uses drawn from assessment data appropriate and justified? Central to validity is the concern for construct representation: Does this assessment measure the knowledge, skills, and dispositions that it's supposed to? Writing ability, the construct measured in all writing assessments, is notoriously complex, as numerous rival models describe the knowledge, skills, and dispositions deemed essential to writing. These range from the simple (e.g., writing as the instrumental ability to encode words or express ideas) to the complex (e.g., writing as a metacognitive, problem-solving process).

Figure 1
Revision Assistant's Theory of Action

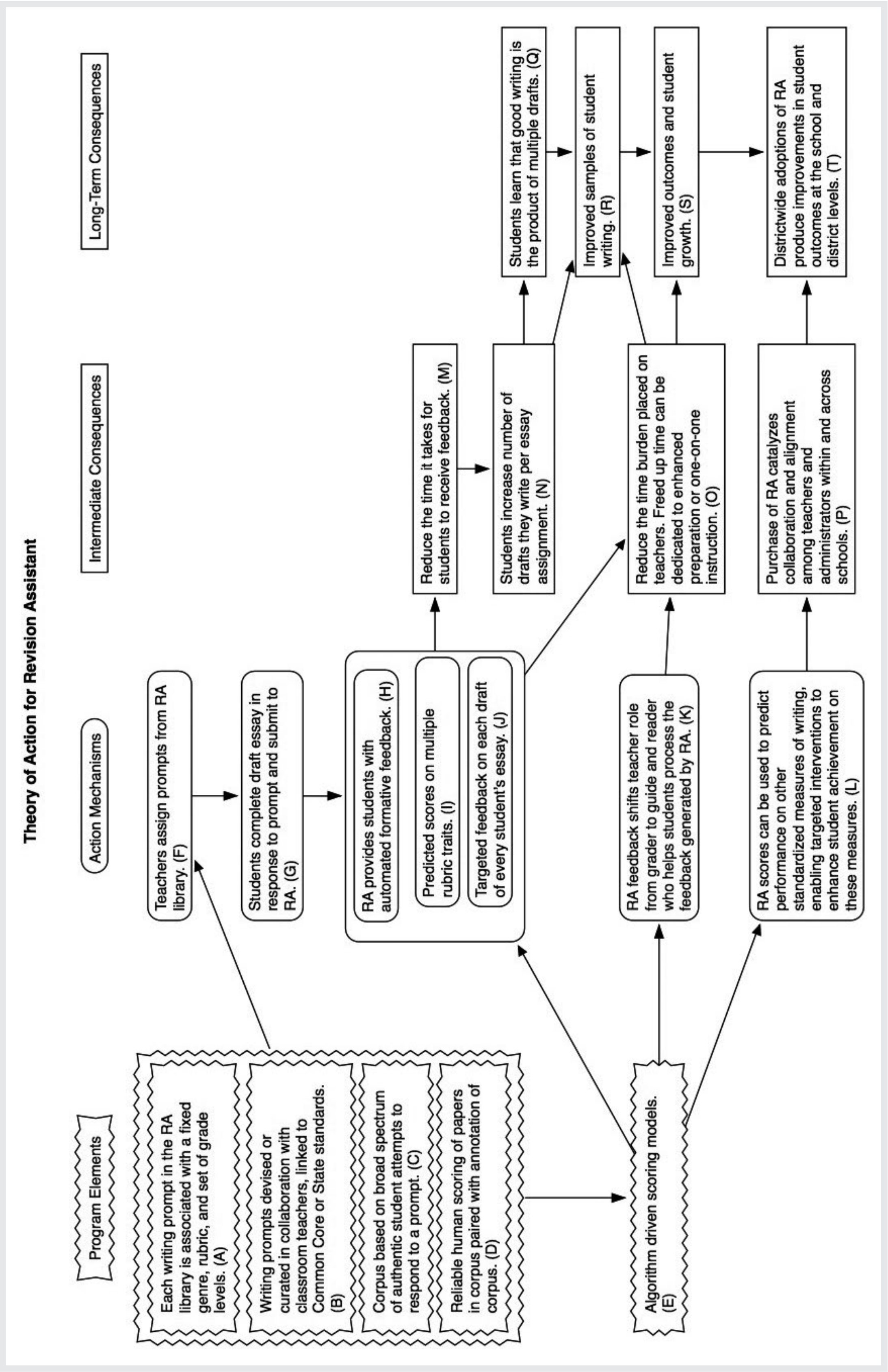
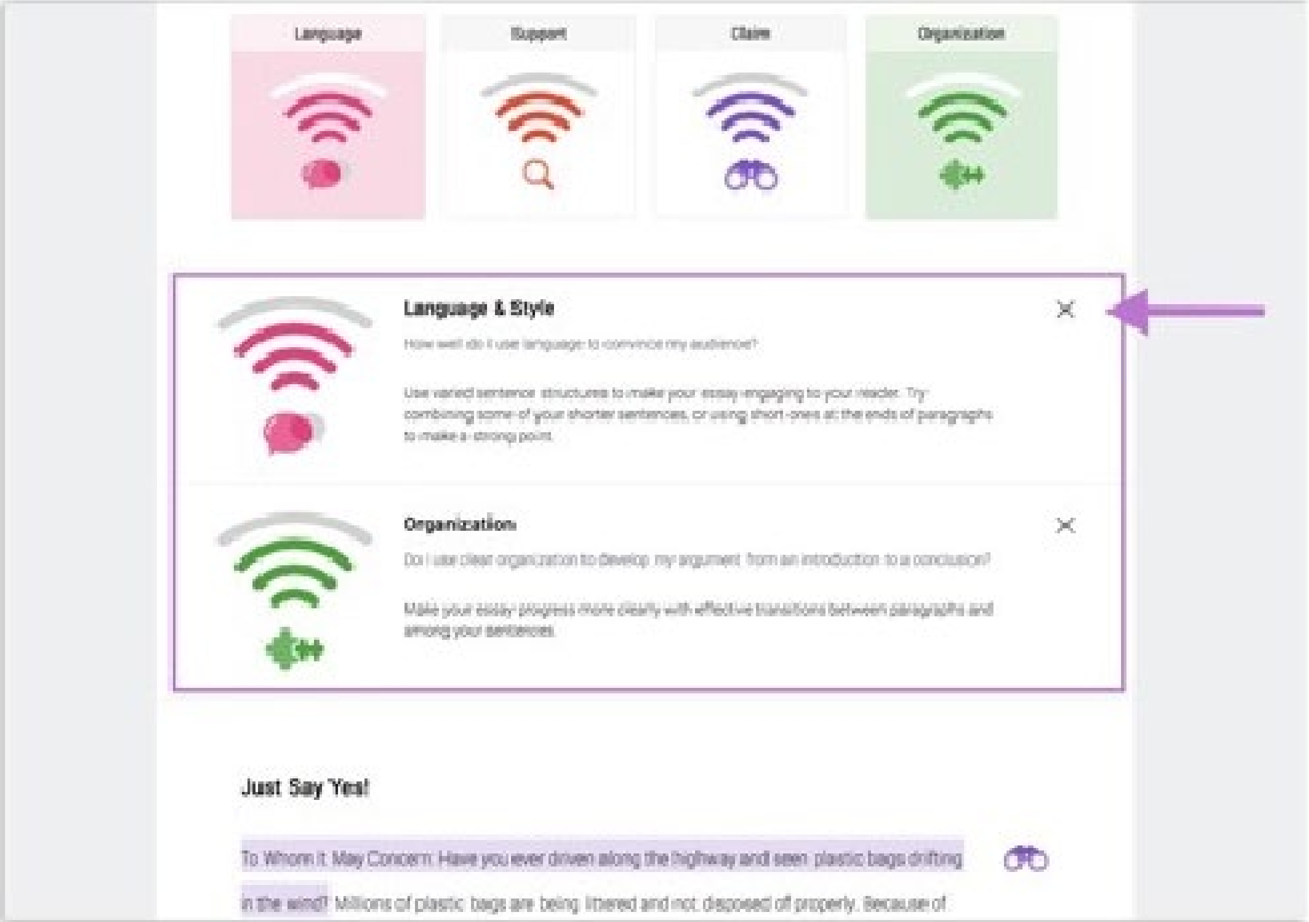

Theory of Action for Revision Assistant
Program Elements
Action Mechanisms
Intermediate Consequences
Long-Term Consequences
Each writing prompt in the RA library is associated with a fixed genre, rubric, and set of grade levels. (A)
Writing prompts devised or curated in collaboration with classroom teachers, linked to Common Core or State standards. (B)
Corpus based on broad spectrum of authentic student attempts to respond to a prompt. (C)
Reliable human scoring of papers in corpus paired with annotation of corpus. (D)
Algorithm driven scoring models. (E)
Teachers assign prompts from RA library. (F)
Students complete draft essay in response to prompt and submit to RA. (G)
RA provides students with automated formative feedback. (H)
Predicted scores on multiple rubric traits. (I)
Targeted feedback on each draft of every student's essay. (J)
RA feedback shifts teacher role from grader to guide and reader who helps students process the feedback generated by RA. (K)
RA scores can be used to predict performance on other standardized measures of writing, enabling targeted interventions to enhance student achievement on these measures. (L)
Reduce the time it takes for students to receive feedback. (M)
Students increase number of drafts they write per essay assignment. (N)
Reduce the time burden placed on teachers. Freed up time can be dedicated to enhanced preparation or one-on-one instruction. (O)
Purchase of RA catalyzes collaboration and alignment among teachers and administrators within and across schools. (P)
Students learn that good writing is the product of multiple drafts. (Q)
Improved samples of student writing. (R)
Improved outcomes and student growth. (S)
Districtwide adoptions of RA produce improvements in student outcomes at the school and district levels. (T)

Figure 2
Revision Assistant's Student View



Language & Style

How well do I use language to convince my audience?

Use varied sentence structures to make your essay engaging to your reader. Try combining some of your shorter sentences, or using short ones at the ends of paragraphs to make a strong point.

Organization

Do I use clear organization to develop my argument from an introduction to a conclusion?

Make your essay progress more clearly with effective transitions between paragraphs and among your sentences.

Just Say Yes!

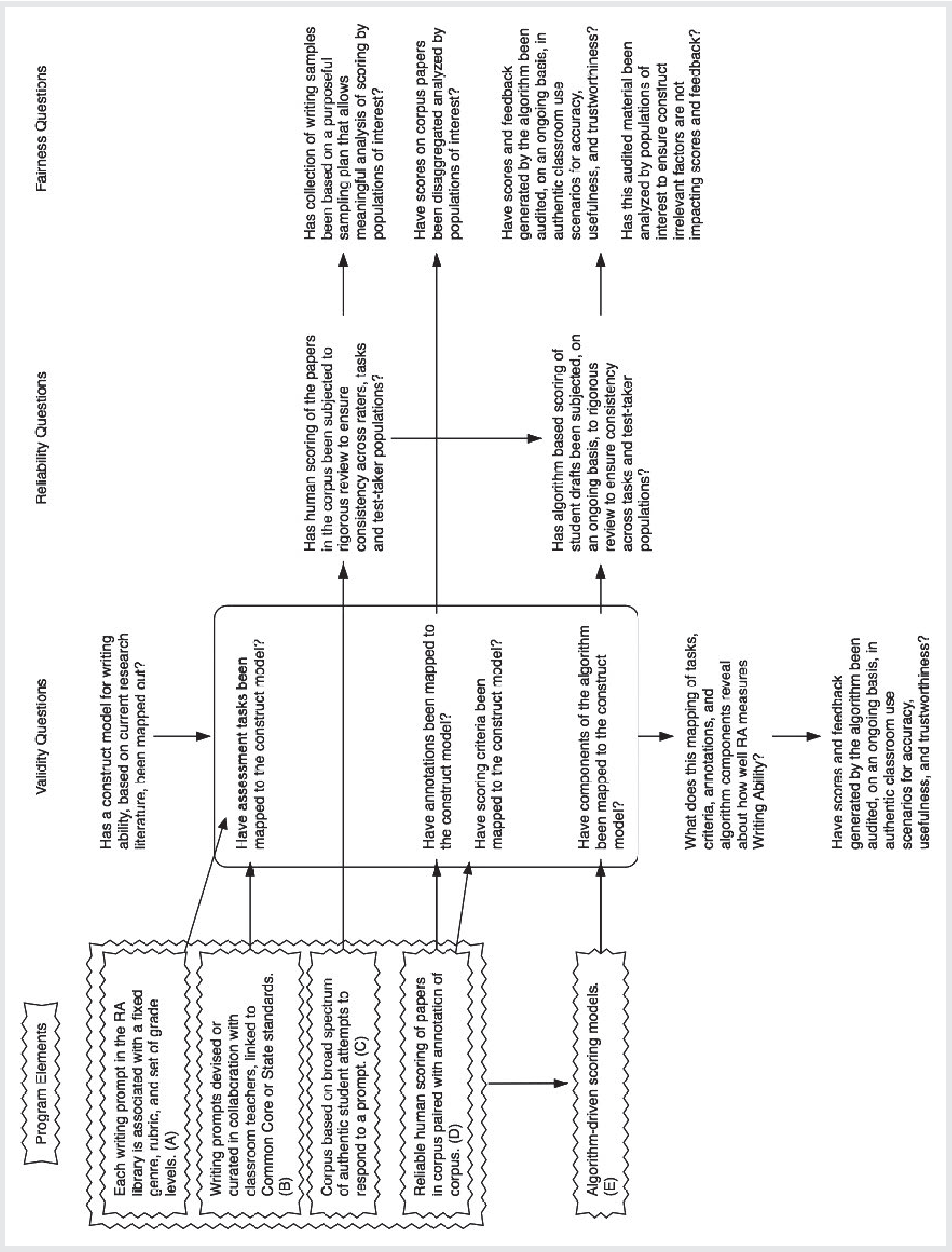
To Whom it May Concern: Have you ever driven along the highway and seen plastic bags drifting in the wind? Millions of plastic bags are being littered and not disposed of properly. Because of

Note. The color figure can be viewed in the online version of this article at http://ila.onlinelibrary.wiley.com.

The validity questions posed in Figure 3 highlight foundational design issues in the Revision Assistant platform. Researchers at Turnitin (West-Smith, Butler, & Mayfield, 2018) suggested that platforms like Revision Assistant do not require explicit construct validation before being implemented in the classroom. They proposed, instead, that validation focuses on alignment with curriculum, diversity of students from whom sample papers are collected, and reliability of the process of annotating these sample papers. Although all of these are important sources of validity evidence, they are essentially meaningless without there first being clear evidence that this platform and its processes reflect a broad and rich understanding of the knowledge, skills, and dispositions that shape one's ability to write. Without this explicit construct, validity evidence, it is impossible for educators to know how well the platform actually measures writing ability, making it difficult, if not impossible, to trust the feedback and scoring provided by Revision Assistant.

A selling feature of automated systems like Revision Assistant is their consistency in scoring. Computers, it is assumed, can filter out the biases and other human factors (e.g., fatigue, boredom, engagement) that create variability when scoring writing. Importantly, for adaptive technologies like Revision Assistant, measurement is not only what comes out of its algorithm but also what goes into it: evidence of reliability drawn from the training data and human scoring used to teach it. Revision Assistant can only assess in the ways its training data were assessed. As Figure 3 points out, if the reliability of human scorers, or the rubrics they use, are not rigorously questioned, their biases become baked into the algorithm that runs Revision Assistant, ultimately impairing the quality of the information provided to students.

Figure 3
Validity, Reliability, and Fairness Questions


Program Elements
Validity Questions
Reliability Questions
Fairness Questions
Each writing prompt in the RA library is associated with a fixed genre, rubric, and set of grade levels. (A)
Writing prompts devised or curated in collaboration with classroom teachers, linked to Common Core or State standards. (B)
Corpus based on broad spectrum of authentic student attempts to respond to a prompt. (C)
Reliable human scoring of papers in corpus paired with annotation of corpus. (D)
Algorithm-driven scoring models. (E)
Has a construct model for writing ability, based on current research literature, been mapped out?
Have assessment tasks been mapped to the construct model?
Have annotations been mapped to the construct model?
Have scoring criteria been mapped to the construct model?
Have components of the algorithm been mapped to the construct model?
What does this mapping of tasks, criteria, annotations, and algorithm components reveal about how well RA measures Writing Ability?
Have scores and feedback generated by the algorithm been audited, on an ongoing basis, in authentic classroom use scenarios for accuracy, usefulness, and trustworthiness?
Has human scoring of the papers in the corpus been subjected to rigorous review to ensure consistency across raters, tasks and test-taker populations?
Has algorithm based scoring of student drafts been subjected, on an ongoing basis, to rigorous review to ensure consistency across tasks and test-taker populations?
Has collection of writing samples been based on a purposeful sampling plan that allows meaningful analysis of scoring by populations of interest?
Have scores on corpus papers been disaggregated analyzed by populations of interest?
Have scores and feedback generated by the algorithm been audited, on an ongoing basis, in authentic classroom use scenarios for accuracy, usefulness, and trustworthiness?
Has this audited material been analyzed by populations of interest to ensure construct irrelevant factors are not impacting scores and feedback?

These challenges related to construct validity and training data reliability can lead to concerns about fairness. Fairness asks whether an assessment has provided evidence that the information provided is valid for all relevant subgroups and that factors irrelevant to writing ability (e.g., cultural knowledge, socioeconomic background) are minimized. Although we have highlighted how Revision Assistant's training data may be biased toward (or against) certain features of writing ability, these predispositions can have profound implications for equity, too. Scholars have demonstrated, for instance, how algorithmic bias in writing technologies often reproduce raced, classed, and gendered formations of difference in classrooms (Dixon-Román, Nichols, & Nyame-Mensah, 2020). Far from a neutral arbiter of writing, the assumptions inherited from the platform's training data and construct of writing may discriminate against nonstandard forms of English. This baked-in bias, in turn, can undercut educators' efforts to make classrooms culturally responsive and supportive of diverse student populations.

Platform Imperatives

In light of these limitations, some readers might wonder, What if the platform used more inclusive training data or was adjusted to eliminate discriminatory biases in its algorithms? Couldn't adaptive assessments, then, support the aims of equity-oriented writing instruction? The problem, however, isn't simply flawed data or algorithms, but rather that the underlying logic of the platform produces imperatives for writing that are often different from those of writing classroom teachers.

One platform imperative relates to content. Because Revision Assistant and similar platforms can only assess what they are trained to evaluate, this sense of defined targets limits what students can write about. A teacher cannot, for instance, use Revision Assistant to assign an open-ended argumentative essay, because Revision Assistant's algorithm has not been trained to score open-ended writing tasks. Even if the training data were expanded to include a wide range of culturally diverse topics, the platform would still need to steer students toward preordained themes to function.

A second imperative involves process. Although today's adaptive assessment technologies are certainly more process oriented than the brute machine scoring of the past, there is a particular logic to the process that animates these platforms. When Revision Assistant offers formative feedback, it does so by comparing students' work with an aggregate of high-scoring essays in its training data. Writing, in other words, becomes a process of iteratively molding ideas to align with a predetermined ideal essay—starkly different from the emergent practices often associated with process-based writing. Even more, this logic also leaves little room for any creative flourishes that deviate from the norms defined by the platform's training data.

Although it would be comforting if the problems of automated assessment could be addressed with unbiased training data or nondiscriminatory algorithms, these simply don't exist, partly because data and algorithms are always embedded in larger platform systems whose imperatives will be aligned with certain predetermined targets (and user purposes) more than others. The integration of such platforms in writing classrooms means educators cannot take platform developers' promises at face value; instead, educators must consider how taken-for-granted ideas—process, choice, and personalization—are redefined when they are folded into the logic of a platform.

The Ethics of Data Enrollment

The development of digital technologies and their uses in classrooms are animated by different interests—educational, social, and financial—and ethical concerns undergird all of these. In schools, one overlooked facet of educational platforms and their data practices is what Zuboff (2019) called surveillance capitalism. From a tech development standpoint, the product that matters most is not what students write but the data generated from the process; platforms participate in data monetization, and our data are repurposed to train machine-learning algorithms. When we invite these apps and platforms into classrooms, our students' intellectual labor generates value for distant companies that have no accountability to or long-term investment in education. Even when we commit to using these platforms ethically in our classrooms, we cannot avoid offering up our students' interactions as data to be used and sold.

Rather than focusing solely on our students' interaction with the writing app itself, an ecological orientation (Nichols & LeBlanc, 2020) can help in understanding how machine-learning technologies are tied into larger systems of governance, ownership, data extraction, and business. This perspective asks us to think of the platform as part of a broader network of systems and processes—scaling up beyond our classroom pedagogy to see how we may be enrolling our students and ourselves in processes beyond our immediate view. Scholars and activists have called for algorithmic transparency: to open up the black box of algorithms and show how these invisible processes are sorting and sifting students. Some have pushed further to rethink what we mean by data transparency (Amoore, 2020): not only understanding the platform's mechanisms (e.g., line-by-line code, data-harvesting practices) but also having policies that can hold companies accountable, that prevent the sale/sharing of our data, and that give local users control over what they have contributed.

Conclusion: From Machine Scoring to Performative Platforms

Adaptive writing technologies are rapidly evolving, and the focus for app developers and educators has now moved beyond high-stakes summative contexts to emerging formative assessment platforms. One-time scoring has been replaced by trait-based machine feedback—closing the loop among testing, assessment, and instruction—presenting its own unique set of challenges and apprehensions for teachers. New machine-learning platforms are performative (Nichols & LeBlanc, in press), constantly changing in response to ever-evolving relationships among the writer, training data, and the underlying algorithm: no longer just teaching to the test but the test recursively shaping writing itself.

Without reflection, the same adaptive technologies advertised to close achievement gaps and aid teachers can become levers by which those gaps are reproduced and teachers' jobs made more precarious. We may not be able to anticipate every possible outcome that new technologies make available, but it is in critical moments like the present—before such devices are fully integrated in the fabric of schools—when we can advocate for an alternate vision of writing instruction and assessment. Literacy educators can help shape this future by pushing back and asking critical questions of these technologies. As Table 1 demonstrates, the theory of action described in Figure 1 can provide a guide for framing these questions. They can help shape purchasing decision for new technologies, inform critical inquiry projects within schools and school divisions, and help temper, contextualize, and challenge expectations for what these technologies can achieve.

Despite these concerns, these emerging technologies can mobilize us to reflect on the construct of writing and the purpose of writing instruction in light of the emerging digital landscape and to set an agenda that can build toward these ends. What is the fundamental intention of our writing in classrooms, and how are we using technology to drive that (and not the other way around)? Despite what technology developers say, there is nothing inevitable about how writing instruction and assessment will take shape in the age of big data. The future of writing remains unwritten, and there is an opportunity for educators to write it ourselves before someone else (or some technology) does it for us.

REFERENCES

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Durham, NC: Duke University.

Dixon-Román, E., Nichols, T.P., & Nyame-Mensah, A. (2020). The racializing forces of/in AI educational technologies. *Learning, Media and Technology*, *45*(3), 236–250. https://doi.org/10.1080/17439884.2020.1667825

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732X18821125

National Commission on Writing in America's Schools and Colleges. (2003). *The neglected "R": The need for a writing revolution*. New York, NY: College Board.

NCTE Task Force on Writing Assessment. (2013). *NCTE position statement on machine scoring*. Retrieved from https://ncte.org/statement/machine_scoring/

Nichols, T.P., & LeBlanc, R.J. (2020). Beyond apps: Digital literacies in a platform society. *The Reading Teacher*, *74*(1), 103–109. https://doi.org/10.1002/trtr.1926

Nichols, T.P., & LeBlanc, R.J. (in press). Media pedagogy and the limits of "literacy": Ecological orientations to performative platforms. *Curriculum Inquiry*.

West-Smith, P., Butler, S., & Mayfield, E. (2018). Trustworthy automated essay scoring without explicit construct validity. In *Proceedings of the 2018 AAAI Spring Symposium Series* (Technical Report No. SS-18). New York, NY: Association for Computing Machinery.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York, NY: PublicAffairs.

The department editor welcomes reader comments.

DAVID SLOMP is an associate professor in the Faculty of Education at the University of Lethbridge, AB, Canada; email david.slomp@uleth.ca.

Table 1
Based on Theory of Action for Interrogating the Use of Revision Assistant (RA) in Literacy Classrooms

Element of the theory of action	Student-oriented question(s)	Teacher-oriented question(s)	System-oriented question(s)
Increased number of drafts (R)	▪ Is the algorithm focusing student revision on sentence-level edits at the expense of global or larger structural revisions? ▪ How divergent from the corpus must a text be before it is considered off topic or bad faith? ▪ When students submit multiple drafts for marking and feedback, are they gaming the algorithm, focusing on letting the algorithm do the work for them?	▪ What are the anticipated implications for student learning and long-term outcomes of relying on writing assignments limited to the prompts, genres, and sample texts contained in the RA database?	▪ How is ownership of student work being protected? ▪ How is Turnitin reinvesting in the educational systems that provide the data that drive their platforms?
Reduced marking burden for teachers (S)	▪ Are students learning how to independently analyze their drafts so they can perform multiple draft revisions without the guidance of the algorithm?	▪ Does use of RA change the roles teachers play in supporting students' growth and development as writers? ▪ Does implementation of RA reduce instructional time burden on teachers? ▪ If RA frees up time for teachers, how is that time being reallocated (a) by teachers and (b) by school administration?	▪ Have the factors in real-life, at-scale applications of RA that support or mitigate positive outcomes for students been identified and investigated?
Enhanced collaboration and alignment (T)	▪ Are these improvements focused on exam type writing, or do they translate to improvements in writing beyond the exam context? ▪ What are the specific known impacts of using RA on students (immediate term and long term)?	▪ What are the known impacts of RA implementation and use on teachers (immediate term and long term)? ▪ Are targeted interventions based on RA data designed to improve RA scores, performance on standardized assessments, or scores on writing tasks in non-testing environments?	▪ Is the use of RA a narrowing force within and across schools? ▪ Does adoption of RA drive an expansive or a closed view (constrained by the RA algorithm, training data, and prompt library) of writing?

Note. The parenthetical letters in column 1 refer to respective parts of Figure 1.