Check for updates

AERA Open
April-June 2016, Vol. 2, No. 2, pp. 1–19
DOI: 10.1177/2332858416641907
© The Author(s) 2016. http://ero.sagepub.com

Big Data Comes to School: Implications for Learning, Assessment, and Research

Bill Cope
University of Illinois

Mary Kalantzis
University of Illinois

The prospect of "big data" at once evokes optimistic views of an information-rich future and concerns about surveillance that adversely impacts our personal and private lives. This overview article explores the implications of big data in education, focusing by way of example on data generated by student writing. We have chosen writing because it presents particular complexities, highlighting the range of processes for collecting and interpreting evidence of learning in the era of computer-mediated instruction and assessment as well as the challenges. Writing is significant not only because it is central to the core subject area of literacy; it is also an ideal medium for the representation of deep disciplinary knowledge across a number of subject areas. After defining what big data entails in education, we map emerging sources of evidence of learning that separately and together have the potential to generate unprecedented amounts of data: machine assessments, structured data embedded in learning, and unstructured data collected incidental to learning activity. Our case is that these emerging sources of evidence of learning have significant implications for the traditional relationships between assessment and instruction. Moreover, for educational researchers, these data are in some senses quite different from traditional evidentiary sources, and this raises a number of methodological questions. The final part of the article discusses implications for practice in an emerging field of education data science, including publication of data, data standards, and research ethics.

Keywords: *big data, assessment, research methods, research ethics, education*

Big data has become a much-used phrase in public discourse, optimistically as well as controversially. In more optimistic moments, big data heralds "a revolution that will transform how we live, work, and think" (Mayer-Schönberger & Cukier, 2013), changing the way we do business, participate in government, and manage our personal lives. In moments of anxiety, we worry about the effects upon our lives of surveillance by corporations and governments (Podesta, Pritzker, Moniz, Holdern, & Zients, 2014). In education, we have witnessed a similar range of promises and anxieties about the coming era of big data. On the one hand, it is claimed that big data promises teachers and learners a new era of personalized instruction, responsive formative assessment, actively engaged pedagogy, and collaborative learning. On the other hand, critics worry about issues such as student privacy, the effects of profiling learners, the intensification of didactic pedagogies, test-driven teaching, and invasive teacher-accountability regimes. Whether one's orientation is optimistic or anxious, all agree that the changes are substantial and that we educators have yet barely explored the implications.

This article maps the nature and consequences of big data in education. We set out to provide a theoretical overview of new sources of evidence of learning in the era of big data in education, highlighting the continuities and differences between these sources and traditional sources, such as standardized, summative assessments. These sources also suggest new kinds of research methodology that supplement and in some cases displace traditional observational and experimental processes.

We ground this overview in the field of writing because it offers a particularly interesting case of big data in education, and it happens to be the area of our own research (Cope & Kalantzis, 2009; Kalantzis & Cope, 2012, 2015b).[1] Not only is writing an element of "literacy" as a discipline area in schools; it is also a medium of for knowledge representation, offering evidence of learning across a wide range of curriculum areas. This evidence has greater depth than other forms of assessment, such item-based assessments, which elicit learner response in the form of right and wrong answers. Writing, in contrast, captures the complex epistemic performance that

Creative Commons CC-BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://www.creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

underlies disciplinary practices. Examples of such disciplinary practices that are best represented in writing include argument in science, information texts in technical subjects, worked examples in math, and documentation in computer code. Writing is also a medium for the representation of affect in self-reflective, metacognitive, and critical student texts. Most relevantly for this article, writing pushes the boundaries of new measurement technologies and processes, illustrating the range of possibilities in the collection and analysis of evidence of learning in technology-mediated learning environments (Cope, Kalantzis, McCarthey, Vojak, & Kline, 2011; Dawson & Siemens, 2014). Writing offers us a case study of the range and depth of data that can be collected incidental to learning.

The article proceeds in four steps, each step using the example of writing but aiming also to make generalizations that go beyond writing and writing assessment. In a first step, we attempt to classify types of data emerging from technology-mediated learning environments. Second, we explore the ways in which these data can be used in learning and assessment, particularly for the purposes of formative assessment of writing and disciplinary learning that has been represented in writing. Third, we examine the ways in which these data expand our sources of evidence and in which big data and learning-analytic methods might supplement traditional quantitative and qualitative research methods. Finally, we consider some of the implications of these developments for research infrastructure, including data access, data sharing, and research ethics.

To set the stage with a definition, "big data" in education is

1. the *purposeful or incidental recording* of activity and interactions in digitally mediated, network-interconnected learning environments—the volume of which is unprecedented in large part because the data points are smaller and the recording is continuous;
2. the *varied types of data* that are recordable and analyzable;
3. the *accessibility and durability* of these data, with potential to be (a) immediately available for formative assessment or adaptive instructional recalibration and (b) persistent for the purposes of developing learner profiles and longitudinal analyses; and
4. *data analytics*, or syntheses and presentations based on the particular characteristics of these data for learner and teacher feedback, institutional accountability, educational software design, learning resource development, and educational research.

In just a few years, two new subdisciplines of education have emerged to address specific questions raised by the phenomenon of big data, each from a somewhat different perspective—*educational data mining* and *learning analytics*. The field of educational data mining mainly focuses on what we will in the next section of this article characterize as "unstructured data," attempting to analyze and interpret evidence of learning from large and noisy data sets—including, for instance, log files, keystrokes, clickstream data, and discussion threads in natural language (R. Baker & Siemens, 2014; Castro, Vellido, Nebot, & Mugica, 2007; Siemens & Baker, 2013). The field of learning analytics tends to be more concerned with what we characterize as structured data, including data models that are "designed in" (Ho, 2015), as is the case, for instance, of intelligent tutors, games, simulations, and rubric-based peer review (Bienkowski, Feng, & Means, 2012; Knight, Shum, & Littleton, 2013; Mislevy, Behrens, Dicerbo, & Levy, 2012; Pea & Jacks, 2014; Siemens & Baker, 2013; West, 2012). The subfields have their own conferences, journals, and burgeoning communities of designers and researchers. In practice, the fields overlap considerably. So, rather than attempting to disentangle the subdisciplines in this article, we construe the field as a whole as "education data science" (Pea & Jacks, 2014).

Evidence of Learning in Computer-Mediated Learning Environments

In computer-mediated educational environments, evidence of learning can be gleaned from a wide range of sources. We want to classify these sources into three major categories, each of which can provide data about learning to write and learning-in-writing across a range of discipline areas. These are summarized in Table 1.

Within each of these major classes of educational data, there are enormous variations in technology and data type. It has been possible to generate these kinds of data in computer-mediated learning environments for some time. What is new is the amount of data that can be generated, the possibility that they can be generated continuously, their variety, and the possibility of analyzing data sets aggregated and integrated across varied sources.

Machine Assessment

Over the past several decades, traditional assessments have been transformed by computerization in two major areas: computer adaptive testing (CAT) in the case of select-response tests and natural language processing for supply-response tests, from short answers to extended essays.

CAT extends long-standing item response theory, where correct student response to test items varies according to what the student knows or understands (a latent cognitive trait) and the relative difficulty of the item. Computer adaptive tests serve students progressively harder or easier questions depending on whether they answer correctly. Such tests provide more accurately calibrated scores for students across a broader range of capacities, reach an accurate score

TABLE 1
A Typology of Educational Data Sources in Computer-Mediated Learning Environments

Data type	Mode of data collection	Assessment genres: Examples
Machine assessments	Computer adaptive testing	Select response assessments, quizzes (e.g., reading comprehension, grammar, vocabulary)
	Natural language processing	Automated essay scoring, feedback on language features
Structured, embedded data	Procedure-defined processes	Games, intelligent tutors
	Argument-defined processes	Rubric-based peer review of writing
	Machine learning processes	Semantic tagging and annotation, text visualizations, accepted textual change suggestions
Unstructured, incidental data	Incidental "data exhaust"	Keystroke patterns, edit histories, clickstream and navigation paths, social interaction patterns
	Dedicated devices for collecting unstructured data	Video capture, eye trackers, movement detectors

faster, and are harder to game because no two students end up taking quite the same test (Chang, 2015). Computer diagnostic testing (CDT) allows for the coding of topic areas within a test and disaggregation of scores within the subdomains addressed within the test (Chang, 2012). In the domain of literacy, CAT assessments are most frequently used for reading comprehension—so frequently, in fact, that reading comprehension often becomes a proxy for literacy in general, at the expense of writing assessments. CAT and CDT assessments can also be used to test specific features of writing, such as grammar and vocabulary.

These testing processes and technologies are increasingly embedded into pedagogical practice, for instance, in the form of end-of-chapter tests in e-textbooks, comprehension tests in online reading programs, or quizzes delivered through learning management systems (Waters, 2014; Woolf, 2010). These in-course tests can be used as decision points in support of adaptive, self-paced and personalized learning, asking and offering an answer to the question, "Is this student ready to proceed?" A single student may now answer thousands of such questions in a year, adding up to more test data than ever in the past. It is possible by this means to develop a comprehensive view of student progress, identifying specific areas of strength and weakness across a succession of interim and summative assessment instruments.

A new species of responsive items offers students immediate feedback on questions, thus serving a formative assessment function. Machine learning techniques (Chaudhri, Gunning, Lane, & Roschelle, 2013) can also be applied whereby item-based assessments improve through use. For instance, newly designed items—even teacher-developed items—that have not yet been validated can be mixed with well tested ones in order to determine their difficulty, and students could offer feedback based on their underlying thinking (correct thinking/wrong answer or incorrect thinking/correct answer may prompt reframing of the item; Cope, Kalantzis, McCarthey, et al., 2011). These developments support the "crowdsourcing" (Surowiecki, 2004) of item development and evaluation. Moreover, to build a comprehensive view of learner progress, much work is needed to develop ontologies (E. Baker, 2007; Cope, Kalantzis, & Magee, 2011) that specify the semantics of items across multiple assessments (what is the underlying cognitive trait?) and support detailed mapping of standards (precisely what, of the curriculum, has the student covered?).

Natural language processing technologies are today able to grade short-answer and essay-length supply-response assessments with reliability equivalent to human graders (Burstein & Chodorow, 2003; Chung & Baker, 2003; Cotos & Pendar, 2007; Shermis, 2014; Warschauer & Grimes, 2008). Perhaps the greatest obstacle to the assessment of writing has been the cost of human grading, including not only the time humans take to read student test scripts but also rater training and moderation to ensure interrater reliability. This is why item-based reading comprehension has until now so often been taken to be a proxy for literacy. Not only do the U.S. Common Core State Standards (CCSS) require a rebalancing of writing within the literacy curriculum; they also recognize the importance of writing across a range of curriculum areas, including science, social studies, and technical subjects (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). As a consequence, there has been a rebalancing of pedagogical emphasis in literacy from a primarily receptive mode (reading), increasing the relative importance of its productive mode (writing). Such rebalancing aligns with "21st-century skills" of active engagement, participatory citizenship, and innovative creativity (Darling-Hammond & Wood, 2008; Partnership for 21st Century Skills, 2008). So, to develop more efficient and effective writing assessments is a challenge of utmost importance.

Natural language processing offers two types of tools for writing assessment, often used in concert with each other:

statistical corpus comparison and analytical text parsing (Cope, Kalantzis, McCarthey, et al., 2011). In the case of the corpus comparison, the computer is "trained" by being given a corpus of human-graded texts; the machine compares new texts and grades them based on statistical similarity with the human-graded texts. In the case of text parsing, computers are programmed to search for language features, such as markers of textual cohesion, the range and complexity of vocabulary, and latent semantics based on word clustering and frequencies (Crossley, Allen, Snow, & McNamara, 2015; Landauer, McNamara, Dennis, & Kintsch, 2007; McNamara, Graesser, McCarthy, & Cai, 2014).

It has been shown that statistical corpus comparison generates reliable retrospective scores in summative assessments (Burstein, 2009). However, its limitation is that it is unable to provide specific meaningful feedback, and some of its measures allow users to game the system. For instance, length of text, longer sentences, and variety of vocabulary have a significant impact on scores (Vojak, Kline, Cope, McCarthey, & Kalantzis, 2011). Textual parsing is capable of giving meaningful feedback, including specific suggestions about language features, such as voice or simple/technical vocabulary (Bailey, Blackstock-Bernstein, Ryan, & Pitsoulaki, in press; Cope & Kalantzis, 2013). In context of big data, the potential number of noticeable language features is as large as the complexity of language itself. The key to natural language processing in the context of formative assessment is to provide salient suggestions—to reduce big data, much of which may be relevant for summative analysis, to data that are germane to immediate student learning. Emerging areas of development in machine parsing technologies include conceptual topic modeling (Li & Girju, 2015; Paul & Girju, 2010; Riaz & Girju, 2010), mapping argument structures (Ascaniis, 2012) and sentiment analysis (Gibson & Kitto, 2015; Shum et al., 2016; Ullmann, 2015; Wen, Yang, & Rose, 2014).

Structured, Embedded Data

Whereas CAT and natural language processing extend long-standing methods of select-response and supply-response assessment, computer-mediated learning can support the generation of innovative forms of structured data designed into the instructional sequence. Here we highlight three kinds of technology and pedagogical process: procedure-defined machine response; the organization and collation of machine-mediated, argument-defined response by humans; and distributed machine learning.

Procedure-defined processes are well suited for highly structured domains where evidence of learning is to be found in correct and incorrect answers. Typical examples are intelligent tutors, learning games, and simulations, most frequently and successfully used for formally structured domains, such as algebra or chemistry (Koedinger, Brunskill, Baker, & McLaughlin, 2013). In the case of writing, procedure-defined processes may be used to address relatively clear-cut aspects of language learning, such as phonics in beginning literacy, and vocabulary and language conventions across a range of levels of capacities to write. Writing tutors can also include strategically placed formal procedure or game activities at different stages in the writing process, from outline to revision (Roscoe, Brandon, Snow, & McNamara, 2014; Roscoe & McNamara, 2013).

Underlying intelligent tutors, educational games, and simulations are cognitive models that lay out the elements of a target domain, anticipating a range of learning paths (Conrad, Clarke-Midura, & Klopfer, 2014). In these cases, learning analytics rely on knowledge tracing, or tracking learning paths. Illustrating this process, VanLehn characterizes the inner and outer feedback loops that underlie intelligent tutors: An inner loop consists of a step in the execution of a task, which generates a correct or incorrect response, feedback on that response, or in the case of expression of uncertainty or incorrect response, a hint. Then, in an outer loop, student response determines the next suitable task (Chi, Jordan, & VanLehn, 2014; VanLehn, 2006). To the extent that there may be more than one "correct" path at each decision point, decision trees may be complex and navigation paths varied (Y. Xu, Chang, Yuan, & Mostow, 2014). However the alternatives at each step are limited by the procedural nature of the domain—in the case of literacy, straightforward language "facts."

Advanced developments in procedure-defined learning technologies include conversational tutors that use latent semantic analysis to "read" short written responses (Chi et al., 2014; Graesser, VanLehn, Rosé, Jordan, & Harter, 2001) and automatic hint generation based on historical data using machine learning methods (Barnes & Stamper, 2008). In all of these scenarios, the bigness of the data derives from the large number of data points as a student progresses. These data points can be made semantically legible to the student and teacher in the form of immediate feedback. Across time and across many students (a class, all the users of the software, a demographic), the size of the data grows proportionately. Pedagogically significant variation between students also may become visible via knowledge tracing visualizations.

Argument-defined processes involve nonformal reasoning (Walton, 2008) that allows scope for a range of more or less plausible conclusions. This contrasts formal logic, where inarguably correct deductions are possible—the underlying logic of procedure-based digital learning environments. Nonformal reasoning processes are necessary for complex and contextually dependent matters that are potentially disputable, involving human judgment and requiring a person to make his or her reasoning explicit while at the same time demonstrating awareness of other plausible reasoning (Brandom, 1994). Examples of nonformal reasoning include

supporting claims with evidence, ethical reasoning, aesthetic judgment, critical analysis, and self-reflection (Co. Lynch, Ashley, Pinkwart, & Aleven, 2009; Ullmann, 2015). Examples of disciplinary practice at the school level include scientific or historical argument, opinion texts, information or explanations, and narratives—to use the terminology of text types in the U.S. CCSS for writing (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). In higher education, examples of such reasoning include clinical case analyses, legal arguments, business case studies, industrial or architectural designs framed within an articulated program, environmental analyses, metacognitive self-reflections, and documentation of computer code or worked solutions to mathematical or statistical problems. Disputability in such domains does not mean that "anything goes"; rather, there is room for discussion about the nature and cogency of underlying reasoning.

These processes of open-ended reasoning are ideally represented in writing—defined broadly here to capture the richness of multimodal knowledge representation now ubiquitous in the era of digital media across a range of disciplines, including embedded media, such as diagrams, videos, data sets, and formulae (Dawson & Siemens, 2014; Kalantzis & Cope, 2012, 2015a). Rubric-based review is one way to systematize the judgment process. Long-standing scholarly practice has established a level of objectivity when writers and reviewers are anonymous (Cope & Kalantzis, 2014). Research shows that interrater reliability improves when criteria are clearly articulated and rating levels and cut scores are explicitly specified. Under these conditions, mean peer ratings are close to expert ratings (Cope, Kalantzis, Abd-El-Khalick, & Bagley, 2013). Peer review in massive open online courses is a prominent example where the judgment of anonymous course participants against clearly specified review criteria and rating levels is assumed to be equivalent to expert judgment (Piech et al., 2013). Using a single, cloud-located source, it is possible to manage what is otherwise a difficult-to-administer process of anonymization, randomization, and simultaneous review by multiple reviewers interacting simultaneously. Such spaces are also designed for the collection of quantitative and qualitative data, distribution of feedback, and synthesis of results. Reviews can also be moderated and aggregated from multiple perspectives—peer, self, expert/teacher—and different reviewer ratings calibrated. Computer-mediated review processes manage social complexity, including multiple reviews, multiple reviewer roles, multiple review criteria, quantitative rating plus qualitative rating, and tracking progress via version histories. As a consequence of these processes of machine mediation, rigorous multiperspectival review becomes feasible as a routine process (Abrams, 2013; Cope & Kalantzis, 2013; Kline, Letofsky, & Woodard, 2013; Lammers, Magnifico, & Curwood, 2014; McCarthey, Magnifico, Woodard, & Kline, 2014). Every data point can be legible for the purposes of formative and summative assessment. From a formative point of view, a learner may receive multiple comments from multiple reviewers across multiple criteria, every one of which is legible for the purposes of improving the next draft. By *legible*, we mean immediately actionable based on specific suggestions, or legible in a specific sense that an overall grade or score is not. At the same time, from a summative point of view, these data quickly become large. When aggregated, they can show individual learner progress from draft to draft within a project or over a number of projects and can offer comparisons across groups of students of various sizes.

Machine learning processes recruit users to provide structured data. Using machine learning techniques, intelligent writing systems can become more accurately responsive with use. Natural language—in essays, class discussion forums, and the like—is from a computational point of view unstructured data or at best lightly structured data. Language itself is of course highly structured, but its structures are only to a limited degree analyzable by computers, whose processes of calculation are circumscribed by a myriad of difficulties, including ambiguity, context dependency, and metaphorical association.

The question then is, how can users constantly train intelligent writing systems by adding a layer of computable structure as they write, self-assess, and assess each other to the unstructured data of natural language? The answer in part is that it is possible in semantically aware writing environments to tag text with machine-readable structure and semantics (Cope, Kalantzis, & Magee, 2011) in order to help computers make better sense of texts and the knowledge that these texts represent. Here are some examples: A heading-level tag can identify the structure of a written text. A student's own writing can be distinguished from quotations by using a "block quote" command for the purpose of analysis of a student's academic language level or plagiarism. Context-dependent machine annotation can add precision (*I* tagged with the name of the person, *today* tagged with a date). Semantic tagging can specify precise meanings against domain-specific glossaries and ontologies (E. Baker, 2007). The structure of text can also be mapped visually (Rekers & Schürr, 1997; Southavilay, Yacef, Reimann, & Calvo, 2013). Argument maps, for instance, might make explicit the underlying thinking in a text (C. F. Lynch, 2014; C. F. Lynch, Ashley, & Chi, 2014), such as the notions of thesis, claims, evidence, counterclaims, rebuttal, and conclusions articulated in the CCSS and other writing standards (Cope et al., 2013). In one example of visual markup, we have created in our "Scholar" environment a tool whereby students highlight sections of information texts (readings, their own texts, their peers' texts) in different colors in order to identify CCSS information text ideas of concept, definition, fact, example, and opinion. This creates nodes for a diagram beside the text in which they outline the structure of the

information presentation (Olmanson et al., 2015). Additional user structuring directly supports the assessment process. It also supports a "crowdsourced" training model where every piece of student markup can contribute to a collective intelligence that grows with system use. For instance, when a peer reviewer codes as "vernacular" a certain term and suggests that the author substitutes another term that the reviewer codes "technical," and if the author subsequently accepts the reviewer's suggestion, this may become training data for the system used at a later time as a suggestion, confirmed by reuse or discarded when the next user rejects.

Unstructured, Incidental Data

Technology-mediated learning environments, such as learning management systems, games, discussion boards, and peer-reviewed writing spaces, create large amounts of "data exhaust" (DiCerbo & Behrens, 2014). This can be captured and recorded in log files: time stamps, keystrokes, edit histories, and clickstreams that show periods of engagement, forms of activity, navigation paths and social interaction patterns. *Unstructured* here means that the data are not framed in terms of a predetermined data model (such as an allowable move in a learning game or a comment against a review criterion in rubric) and that each data point does not have an immediately obvious meaning. The data points are mostly even smaller than embedded structured data, more numerous, and inscrutable except in a larger context of aggregated data. To be made meaningful, the computer's statistical pattern recognition must trained by human inference—certain patterns of activity correlates with what a human has elsewhere judged to be relative success or lack of success in learning. These training data may be created by experts or collected incidental to learner activity within a learning management system, for instance. In empirical practice, structured and unstructured data are generated simultaneously, and much data might be classified as semistructured.

Incidental data exhaust may be mined for patterns of activity that predict learning outcomes (Atzmueller, 2012). A particular pattern of drafting, peer interaction, and revision may predict, for instance, a high grade in a writing assignment. Going beyond single students, patterns of success may be compared with classes, across demographics, and between teachers. These data can be used to provide advance warning that a student requires attention in a particular area (Cope & Kalantzis, 2015). Social interaction analyses (Speck et al., 2014; Wise, Zhao, & Hausknecht, 2013; X. Xu, Murray, Woolf, & Smith, 2013), or edit histories showing relative contributions in collaborative online writing environments (McNely, Gestwicki, Hill, Parli-Horne, & Johnson, 2012) may offer important predictive data as well as retrospective assessment data. Affective states that impact learning outcomes may also be detectable in patterns of action and interaction, including signs of confusion, frustration, boredom, or flow/engagement (R. Baker, D'Mello, Rodrigo, & Graesser, 2010; Chung, 2013; D'Mello, 2013; Dowell & Graesser, 2014; Fancsali, Ritter, Stamper, & Berman, 2014; Paquette, de Carvalho, & Baker, 2014; Winne & Baker, 2013; Wixon et al., 2014). For such cognitive or affective analyses, machine learning algorithms use training data created by experts (a judgment that a detectable pattern in the data corresponds with a certain affective state). Alternatively, users can train the system on a continuous basis in parallel structured data collection, for instance, in emote-aloud meters (D'Mello, 2013). Natural language processing methods can also be used to parse student written reactions for sentiment (Fancsali et al., 2014).

Dedicated devices for collecting unstructured data may include hardware and software to capture eye movements, gaze, facial expressions, body posture and gesture, in-class speech, and movement around the classroom (D'Mello et al., 2010; Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2014; Be. Schneider & Pea, 2014; Vatrapu, Reimann, Bull, & Johnson, 2013). Technologies include video capture, bracelets, watches, radio-frequency identification chips, quick-response codes, and specialized detectors that capture patterns of bodily movement, gesture, and person-to-person interaction (Lane, 2013; Lindgren & Johnson-Glenberg, 2013). Such devices generate massive, noisy data sets the meaning of which requires human training. Human-interpretable conclusions applied in training data sets are matched with patterns discernable in new data sets, so the human interpretation can presumptively be applied to the new data. In this supervised machine learning, labels are applied in advance to a range of computable patterns. Unsupervised machine learning works around the other way, clustering computable patterns and suggesting to humans that a human-interpretable label may be applicable to commonly occurring patterns. In our Scholar research and development, we have created a tool that traces learner thinking in the form of a sequence of moves as users create a visualization of the underlying logic of their information and argument texts. The question then is, what patterns of thinking predict successful or less successful written texts (Olmanson et al., 2015)?

Any and all of these data sources can provide evidence of learning to write and learning-in-writing across a number of subject areas. The challenge for the moment is that any one learning and assessment environment today offers only a subset of these opportunities. There are significant challenges to build the more comprehensive view that big data in education promises in theory.

Putting Big Data Evidence to Work in Learning and Assessment

Learning is a complex, multifaceted activity. To return to the example we have been using in this article, any and every

technology and process of educational evidence gathering can be applied to learning to write and to represent disciplinary knowledge in writing. A number of implications for the future of assessment can drawn from the diverse and rich sources of evidence of learning to write and to learn through writing in technology-mediated environments. These are summarized in Table 2.

The "big data" we have been classifying and describing are bigger than older data sources only because the data points have become smaller and as a consequence of the incidental recording of learning activity. The activities of learning, meanwhile, are no larger and no more complex than they ever were. The data are bigger only as a consequence of the finely grained mechanisms to collect evidence of learning that can be embedded within technology-mediated learning environments (Thille et al., 2014). This is particularly the case for complex disciplinary performances, such as writing and knowledge represented in writing. We can now create writing environments that capture trace patterns in learning events (Winne, 2014) and that collate learning process data (Knight, Shum, & Littleton, 2014). Some of these data points may be structured data that are semantically legible to teachers and students—the answer to a grammar or vocabulary question in the case of procedure-defined evidence generated in tutors or games or, in the case of argument-defined evidence, a comment or rating in the case of a response to a criterion in a rubric by a peer. Other of the data points may not be so immediately legible, requiring aggregation and interpretation of unstructured data, for instance, applying machine learning processes to predictors of success at writing in the edit history of a wiki or blog.

The intrinsic embeddedness of these sources of evidence points to exciting possibilities for the more comprehensive realization of long-held aspirations for formative assessment (Airasian, Bloom, & Carroll, 1971; Black & Wiliam, 1998; Bloom, 1968; Shute, 2009; Wiliam, 2011), until recently a comparatively neglected mode of assessment compared to summative assessments for the purposes of institutional accountability (Armour-Thomas & Gordon, 2013; Gorin, 2013; Kaestle, 2013; Ryan & Shepard, 2008; Shepard, 2010). Computer-mediated environments can support immediate machine feedback by means of natural language processing, CAT, and procedure-based games, for instance. They can also offer extensive peer and teacher feedback by streamlining the complex social processes of machine-mediated, argument-defined human feedback.

Where such feedback mechanisms are built in, the potential arises to end the historical separation of instruction and assessment. A myriad of small moments of learning may also be a moment of formative assessment. The potential arises for feedback that is always available on the fly. The feedback can be recursive in the sense that it prompts a response that prompts further feedback. Feedback on feedback ("That was helpful/not helpful") can also produce a quick response. Such feedback is immediately actionable in specific ways. It can determine appropriate pedagogical progression for more personalized learning. These have been precisely the objectives we have set ourselves in research and development for our Scholar platform (Cope & Kalantzis, 2013). In this context, instruction and assessment are integrated (Armour-Thomas & Gordon, 2013; Cope & Kalantzis, 2015). In this way, pedagogical design also becomes "evidence-centered design" (Mislevy et al., 2012; Rupp, Nugent, & Nelson, 2012).

These transformations point to the emergence of records of evidence of learning that are more comprehensive, and the analytics more thorough, than legacy summative assessments. Indeed, we may be able to use assessment data that were in the first instance formative, for summative purposes. In our traditional practices of learning and assessment, we have conceived formative and summative assessments as different kinds of processes, creating different kinds of data, in different ways, at different times and for different purposes. However, today's data-rich learning environments may blur the formative/summative distinction, where every data point in a summative perspective may already have served a formative purpose. The difference then is one of perspective rather than a fundamental distinction of assessment type.

Moving from finely grained perspectives of individual learner progress to comparative analyses of cohorts and demographics, it is now possible to "zoom out" from specifics to wider views and to "zoom in" from the larger views in order to identify the dynamics of specific learning sequences at an individual or group level (Worsley & Blikstein, 2014). At the most granular level, it is possible in our Scholar environment and others to see any and every semantically legible data point, where each such data point has been a symptomatic waypoint in a student's progress map (DiCerbo & Behrens, 2014). Different learning paths now become visible.

In these conditions, we also witness a shift in emphasis from making inferences about cognition to a focus on the artifacts created by learners in the process of their knowledge construction. The classical "assessment argument" is a three-cornered triangle: observation (for instance, responses to a bank of test items), interpretation (which and how many of the responses are right or wrong), and cognition (in the form of an inference about student understanding of the domain under assessment) (Pellegrino, Chudowsky, & Glaser, 2001). However, now we can assess the artifacts of knowledge making, and as process as well as product. To take writing, we can keep a progress record across versions, including, among the many other sources of evidence that we have discussed, clickstream-records sources that have been read, annotations taken and notes made, and contributions of peer reviews to the development of a text. DiCerbo and Behrens (2014) call this an activity paradigm in contrast to an item paradigm. In this scenario, learners are conceived

TABLE 2
Traditional Compared to Emerging Models of Assessment

Traditional assessment model	Emerging assessment model
Assessment is *external* to learning processes; the challenge of "validity" or alignment of the test with what has been taught	Assessment is *embedded* in learning; "validity" no longer a challenge
Limited opportunities for assessment, *restricted data sets* (select and supply response assessments)	Data are big because there can be *many small data points* during the learning process (structured and unstructured data)
Conventional focus on *summative assessment*	Renewed focus on *formative assessment*
Summative assessment is an outcomes or *end view* of learning	Summative assessment is a progress view, using data that were at first formative to trace learning progressions; feedback is *recursive*
Expert or teacher assessors	*Crowdsourced*, moderated assessments from multiple perspectives, including peers and self
Focus *on individual memory* and deductions leading to correct or incorrect answers	Focus on knowledge representations and *artifacts* that acknowledge textual provenance and trace peer *collaborations*
Assessment of *fact and correct application*	Assessment of *complex epistemic performance*, disciplinary practice
Assessment experts as *report grades*	Learners and teachers as data analysts, with the support of *analytics dashboards and visualizations*

as agents who have a range of choices as they construct knowledge (Winne, 2006). At the completion of a writing project, we can assess a constructed knowledge artifact (Berland, Baker, & Blikstein, 2014) that is a product of complex epistemic performance, a tangible outcome of disciplinary practice. The integration of assessment into pedagogy in these ways also addresses long-standing challenges of validity (Messick, 1989) in legacy summative assessment systems. The distinction between what is being taught and what is being assessed is reduced or eliminated when assessment is designed into the learning.

Moreover, there is a shift from a focus on individual thinking and the recall of facts or operations. This is the traditional cognitive focus of assessment (Dixon-Román & Gergen, 2013). Rather than memory and isolated mental skill, learning analytics can now trace the learner's synthesis of readily available knowledge sources and tools, and refinement based on peer and teacher feedback. For instance, an assessment need not focus on what a learner can remember about climate change for a select-response science test. Rather, the focus is how well he or she is able to write a scientific argument about climate change, having accessed available sources, analyzed different hypotheses, and evaluated the evidence provided to support claims related to alternative scientific arguments. If a student's peers have given him or her feedback on a draft, the social provenance of the student's thinking is traceable, as are the sources he or she has accessed via clickstream records and has recognized in citations. Some of these social sources of knowledge may in an earlier era have been construed to be cheating. Now they become an integral part of the collaborative learning ecologies. These parallel the knowledge ecologies that constitute the practice of "real" science (Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014; Lansiquot, 2013). Instead of assessing individual cognition in the form of memory and correct application of theorems, we now assess evidence of cognition in the documented social provenance of information, the peer activities that went into the collaborative construction of knowledge artifacts, and the quality of reasoning behind a conclusion. In other words, we are interested in something much deeper than whether an answer happens to be considered right or wrong (Cope & Kalantzis, 2015).

This is the shape of an emerging infrastructure that captures data providing evidence of learning and to represent a wide range of complex disciplinary understandings in complex forms of knowledge representation, such as writing. All the pieces of such an infrastructure are already available in a fragmentary way, and we have brought a number of them together in our Scholar web writing and assessment environment (Cope & Kalantzis, 2013).

To teach and learn in such environments requires new professional and pedagogical sensibilities. Everyone becomes to some extent a data analyst—learners using analytics to become increasingly self-aware of their own learning and teachers as they acquire a level of data literacy required to interpret a student's progress and calibrate their instruction (Twidale, Blake, & Gant, 2013). Much learning will be required on the part of both students and teachers as they become familiar with these environments. Importantly also, as we will see in the next section of this article, these data also become rich sources for analysis by researchers, instructional designers, and educational software developers in a new "education data science" (Pea & Jacks, 2014).

Toward Education Data Science

At the dawn of the computer age, Warren Weaver, one of the founders of information theory, spoke of a coming third phase of science. The first phase had until about 1900 dealt with two-variable problems, such as the laws of Newtonian physics, which resulted in the great mechanical inventions of modern times. In a second phase, probability theory was used to deal with the disorganized complexities of biological and social realities. In a third phase, he predicted a science in which computers could support analysis of organized complexity (Weaver, 1948). We contend that this is now becoming possible in education with big data, whose evidential complexity lies in their fine granularity and wide variety. One such example of complexity, we have been arguing, is the process of writing and organizing disciplinary knowledge in written form. How then do the big data, collected incidental to the learning process, allow us to see into this complexity? What are the implications for educational researchers? And what opportunities and challenges arise for the social sciences in general (Burrows & Savage, 2014; Kitchin, 2014; Savage & Burrows, 2007) as well as the learning sciences? Our responses to these questions follow, and are summarized in Table 3.

An Embedded Role for the Researcher and Distributed Data Collection

The traditional role of the educational researcher, particularly within the experimental model, is that of an independent observer. As a consequence, researchers design and implement instruments of measurement that are mostly separate from the objects being measured—surveys, tests, interviews, observation protocols, and the like.

However, in the case of data collected incidental to learning, the instruments of measurement are embedded in the learning. Some of the measurement may be done by teachers, peers, and the learner in his or her own self-evaluation. The data are collected via mechanisms that are also integral to the learning. The subjects are, in effect, recruited as data collectors—as is the case in peer essay assessments, think-aloud annotations, or crowdsourced training in machine learning. In our research on science writing in the middle school, we have demonstrated that when rating-level descriptors are clear, mean scores of several non-expert raters are close to those of expert raters (Cope et al., 2013). According to a logic now termed the "wisdom of crowds" in online and big data contexts (Ranade & Varshney, 2012; Surowiecki, 2004), the expert human judgment of teachers or researchers can be meaningfully supplemented by non-expert judgments, such as those of students themselves (Strijbos & Sluijsmans, 2010). Web 2.0 technologies have demonstrated the effectiveness of non-expert reputational and recommendation systems (Farmer & Glass, 2010; O'Reilly, 2005). In these ways, the role separation between data collector and research subject is blurred.

In this scenario, researchers need to reposition themselves as data collaborators—working alongside the instructional software designer, teacher, and learner. To the extent that the division of instruction and assessment is blurred in the era of big data, so also is the division blurred between the data used for pedagogy and the data used by researchers in the educational data sciences. And to the extent that formative and summative assessment becomes perspectives on the same data, research is also grounded in these data.

Sample Sizes, Where N = All and N = 1

In the standard educational research model, the ideal sample size is "*N* = just enough." There are costs of time and effort in instrument development, implementation, collection, and analysis. For this reason, *N* has to be enough to minimize sample error or bias while still supporting generalizability. This *N* may be small in the case of thick qualitative data or larger for quantitative analyses. However, in an era when data are collected incidental to learning and embedded assessment, there is no marginal cost in analyzing data sets of any size. The possibility, in fact, arises to study *N* = all, where "all" may be every user of a piece of cloud software or every student in a data collection catchment area. We already have whole-population or census data available in administrative data sets. Now we can also seek a more granular view of learner activity in the data emerging from computer-mediated learning environments. At the other end of the scale, with enormous amounts of collectable at the level of an individual student or a single case, *N* = 1 can yield reliable data, too. This means also that there need not be the bifurcation of sample sizes and methods that has traditionally been the mark of the qualitative/quantitative divide. Big data can simultaneously support *N* = 1 and *N* = all.

Multiscalar Data Perspectives

Schools have always offered feedback at different scales, from immediate feedback in the form of classical classroom discourse—where teacher initiates, student responds, and teacher evaluates (Cazden, 2001)—to summative assessments (Mislevy, 2013). But these are different kinds of feedback processes, created for different feedback orientations and generating different kinds of data that could not practicably be brought into relation with each other in a comprehensive view. In the case of big data, scaling up or down, zooming in or out, offers a range of viable perspectives on a shared data source—a micromoment of feedback in the writing process, for instance, to larger patterns of revision, to overall progress of a student or a class or cohort measured in terms of writing standards over a longer time frame.

TABLE 3
Traditional Compared to Emerging Models of Research

Traditional research model	Emerging research model
Researcher as *independent observer*	Researchers recruit subjects as data collectors, *co-researchers*
Optimal *sample N* to produce reliable results	There is no marginal cost for $N = all$, and data are rich enough to support $N = 1$
Practical limits to research perspective determined by the *scale of data collection*	*Multiscalar perspectives*, from $N = 1$ to $N =$ all
Fixed time frames, long enough to demonstrate overall effect; longitudinal analyses expensive and thus infrequent	*Short time frames*, feeding small incremental changes back into the learning environment; *longitudinal time frames* as a consequence of data persistence
Standardization effects (fidelity, average effect)	Tracing *heterogeneity* in data, e.g., different paths in adaptive learning environments, salient activities of outliers
Causal effects: *overall*, for whole populations or population subsets	*Microgenetic* casual analysis, e.g., learning progressions for different students, differential effects traceable in varied learning paths
Relatively *separate quantitative and qualitative* research practices; *low significance of theory* in empirical analyses	*Integration of quantitative and qualitative* analyses; *increasing importance of theory* in data analyses

Variable Time Frames

In the standard experimental model of educational research, the intervention has to last long enough to demonstrate statistically significant, overall, differential effect when an intervention group is measured against a comparison group. This takes time and costs resources, which also means that for practical purposes, the research process is linear: intervention → data collection → data analysis → conclusion regarding effect. More fluid models have been advocated for some time in the form of design experiments (Laurillard, 2012; Schoenfeld, 2006) and microgenetic classroom research (Chinn, 2006). Such approaches did not until now bear the weight of statistical proof of effect afforded in the standard model. However, new possibilities emerge in the era of big educational data. For instance, in the case of $N =$ all, experimental conditions can be engineered into everyday practice in the form of A/B studies (Tomkin & Charlevoix, 2014), where "all" is divided into Group A (users working in a beta version that includes the proposed software revisions) and Group B (current software version) in order to compare the effects of the changes created in the A software instantiation before full implementation. This is, in fact, how user testing occurs in "agile" software development methodologies (Martin, 2009). Such an approach contrasts with an earlier generation of linear "waterfall" software development: specification → coding → testing → delivery. If the standard educational research model was best suited to evaluate "waterfall" development, new research methods required "agile" software development. In the development of our web-based Scholar writing and assessment environment, we go through a 2-week cycle of development → A/B trials → implementation. Virtual learning environments today are typically in constant development (Dede, 2015; Wolf, 2010), because these development methodologies emphasize rapid, frequent, and incremental cycles of design, testing, and release (Martin, 2009; Stober & Hansmann, 2009). In these circumstances, researchers can assume a role deeply embedded in the design, implementation, and testing process, for instance, in micro intervention-result-redesign cycles. Not only are research time frames dramatically compressed; research becomes an integral part of an incremental and recursive development process. On the other hand, in this context, research time frames can also be dramatically extended. For instance, when learning data are persistent, longitudinal data can be explored on longer time frames than practicable in the traditional model, possibly extending even to lifelong learner models (Kay, 2008).

Dynamic and Heterogeneous Data

In the standard model of experimental educational research, fidelity of implementation is required. To demonstrate an overall effect, every learner in an intervention needs to have the same experience in the intervention, and the comparison group needs to be held constant in order for the difference to be consistent. Measurement of effect needs to be standardized (Mitros, 2015). However, adaptive or personalized learning (Conati & Kardan, 2013; Graesser & McNamara, 2012; Koedinger et al., 2013; McNamara & Graesser, 2012; Wolf, 2010) has continuous recalibration built into it. It is nonstandardized by design. A certain kind of infidelity is built in. The same is true for software that gives teachers the scope to teach content they have developed themselves and in their own way—as opposed to faithfully following the designer's script. This last scenario is particularly pertinent in the case with writing and writing assessment software, where learner outcomes are strongly

related to the teacher's writing prompts, rubrics, disciplinary expectations, and pedagogical framing. In the nature of this wave of software and instructional design, these data cannot be standardized and homogenized because they are by nature dynamic and heterogeneous. Rather than speak to average overall effects, we can track differential effects through big data techniques, such as network mapping, systems analysis, model tracing, diagramming, and visualization (Maroulis et al., 2010). Now that we can see the detail as well as the norm, we may gain valuable insights from outliers and edge cases. We may also see different processes producing different effects for different demographic groupings.

Tracing Causal Effect

In the standard model of experimental educational research, causal effect can be inferred when the

> difference between what would have happened to the participant in the treatment condition and what would have happened to the same participant if he or she had instead been exposed to the control condition.… Because the statistical solution to the fundamental problem of causal inference estimates an average effect for a population of participants or units, it tells us nothing about the causal effect for specific participants or subgroups of participants. (Ba. Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007, pp. 9, 19)

Today, it is possible to supplement these analyses to some extent with more detailed causal explanations using multilevel statistical models. These can be cross-validated with embedded "big data" analytics. For instance, in the case of data collection embedded within periods of learning, it is possible to drill down to constituent data points to explore learning processes. In these ways, big data analyses can complement or supplement analyses of overall effects and multilevel analyses with microgenetic causal analysis, allowing researchers to investigate the details of process and even to find things that may not have been anticipated within a statistically normed "causal black box" of overall effect (Stern et al., 2012, p. 7). By means of emerging tools, such as machine learning, neural nets, and support vector machines, educational data science can cross-validate multilevel experimental and measurement methods, including "cross-level" interaction, that is, interactions between groups and individuals and micro- as well as macrolevel processes. This has the potential to offer new insights into the differences between individuals and similarities within subgroups, multiple causality, contributing factors, contingencies, nonlinear pathways, causes and effects that are mutually influential, and emergent patterns.

Convergence of Qualitative and Quantitative Methods, and Empirical and Theoretical Work

At times, discussions of big data appear to presage a future dominated by quantitative research and perhaps even an "end of theory" (Anderson, 2008) where algorithmic processes, such as "data smashing," will produce results that emerge directly from the data (Chattopadhyay & Lipson, 2014). The reality of big data, however, is also likely to be one where theory is as important as ever and qualitative methods are needed beside quantitative. In physics, the Higgs Boson became visible in the mass of noisy data generated by the Large Hadron Collider only because its possibility had already been hypothesized in theoretical conjecture (Hey, Tansley, & Tolle, 2009). Similarly, we may find patterns in big educational data only on the basis of conjectural logic models. Statistical patterns in machine learning data are to a significant extent creatures of patterns already built into supervised training models. In the case of unsupervised machine learning, the statistical patterns make sense only when they are given explanatory labels. For these reasons indeed, theory is needed more than ever to frame data models, to create ontologies that structure fields for data collection, and for model tracing. Patterns, moreover, may become meaningful only after drilling down to semantically legible data points—asking questions, such as "What was this outlier in fact doing?" Quantifiable judgments by self, peers, or teachers (an evaluative judgment in a point selected in a Likert scale) may be supported by qualitative justifications (a comment supporting that judgment). In these ways, the qualitative and the quantitative are productively interleaved. Furthermore, natural language processing technologies use statistical methods to parse data that have traditionally been regarded as qualitative par excellence. It is through these variously combined qualitative and quantitative methods that complex knowledge representations, such as those made in writing, become data that provide evidence of learning.

The methodological developments we have described here do not overturn the established practices of quantitative and qualitative educational research. In some instances they incorporate them, reducing human effort and making them less expensive. At other times, they supplement and complement established research practices. Perhaps the greatest value, however, is the possibility in any particular case to analyze a variety of data types using a variety of methods, cross-validating these against each other in a more powerfully holistic, evidence-based repertoire of research practices.

Implications for Research and Data Infrastructure

So far in this article, we have explored the emergent possibilities of big data in education, and educational data science, illustrating these with the example of writing and complex knowledge representations in writing. However, the practicalities of big data present considerable challenges for our research infrastructure and research dissemination practices. Here we mention three: data access, data models, and data privacy. These are summarized in Table 4.

TABLE 4
Traditional Compared to Emerging Research and Data Infrastructures

Traditional infrastructure	Emerging infrastructure
Journal articles and monographs *summarize results*	Publication of *full data sets*
Meta-analyses are based on *results as reported*; few replication studies	Meta-analyses can *mine multiple data sets*; closely aligned, easily implemented replication studies
Divergent data models mean that it is difficult to align datasets	Data *standards* support interoperability of data
Research ethics protocols based on *consent prior to research*; distinct *research activities*	The ethics of *mining historical data*; creating data effects by experimental intervention where the *data collection and instruction are integrally related*

Data Access and Publishing

In legacy research models, data were collected and kept by academicians. In the era of big data, they are found objects, and their owners are by and large outside of academe (Savage & Burrows, 2007). In the case of educational data, they are often stored in the data warehouses of commercial software hosts or school administrations. This makes it harder to access data for research processes. In the case of web-based writing, environments such as Google Classroom/Docs, blogs, and wikis now capture enormous amounts of student writing but thus far offer limited supports to analyze this writing for the purposes of pedagogy and assessment. Our Scholar project is one attempt to provide a more comprehensive range of writing assessment data.

At the knowledge dissemination end of the process, new opportunities as well as challenges emerge. The historical medium for knowledge declaration is the peer-reviewed journal article or monograph. These have typically declared summary results but not the data sets in which these results are grounded. Today, there is a move to publish data online alongside or linked to the article or monograph, often in repositories (Cl. Lynch, 2008; Shreeves, 2013). This opens the possibility of replication and comparison studies based on the same or parallel data sets. It also allows for deeper meta-analyses, which until now have been able to do little more than report on aggregated effects (Glass, 2006; Hattie, 2009). In other words, meta-analyses could in future be grounded in underlying data and not just reported results. Of course, issues of data availability, replicability, and commensurable data models emerge, although these are not unique to big data. The difference is that the order of complexity in addressing these issues increases with data size and variability, as do the potential benefits to the extent that we as a profession manage to address these technical challenges.

Attempts are under way to create accessible, specialized data repositories, including the University of Illinois' National Data Service or, in the case of educational data sciences, the Pittsburgh Science of Learning Center's DataShop initiative (Koedinger et al., 2010). In the case of writing, these data sets may include full text, revision histories, and ancillary assessment data. However, how are underlying data peer reviewed? How is they cited? How is the provenance of data sets acknowledged, as well as representations of data? These are infrastructural challenges that we are only now beginning to address.

Data Models and Interoperability

It is one thing to make data more accessible, however quite another for data sets to take commensurable forms that will allow for replication studies or meta-analysis. How does one set of questions in a computer diagnostic test compare in difficulty and content with another set of questions in a different test? Which language or writing standards or objectives is one set of assessable writing tasks and texts designed to address, compared to another set of writing tasks and texts in a different piece of software, or school, or subject area? How does the learning model in one language learning game map to the learning model for a different game? How do data generated in learning and assessment environments align with institutional and administrative data sets (Wagner & Yaskin, 2015)? Without addressing these questions, the data remain isolated in self-referencing islands.

These questions of data commensurability can be addressed by emerging methodologies for interoperability across data models. One approach involves creating data standards. In the United States, the Common Education Data Standards (Office of Educational Technology [OET], 2016, p. 60) and the Schools Interoperability Framework set standards for system-level data. Instructional Management Systems (IMS) and its Learning Tools Interoperability standard create a common framework for learning management systems and educational software. The IMS "Caliper" development offers a more finely grained view of learning activity. In the spirit of further development in this direction, development of an educational data dictionary has been recommended (Woolf, 2010, p. 65), as has a notion of "open learning analytics" that ties together data emerging from a number of platforms (Siemens et al., 2011). Much work needs to be done to create interoperability in the area of federated data, using methods broadly known as the "semantic web" (Cope, Kalantzis, & Magee, 2011).

Data Privacy and Research Ethics

Critical issues also arise in the areas of data privacy and research ethics (Data Science Association, n.d.; Hammer, 2015; OET, 2016, p. 74; Piety, 2013; Zwitter, 2014). One of the most celebrated early big data initiatives in the field of education was the $100 million inBloom educational data warehouse, funded by the Gates and Carnegie Foundations. Within a year of its launch, it had collapsed as states, districts, teachers, and parents took fright at the prospect of "big data" seeing into children's lives to determine their educational destinies and seeing into teachers' lives to determine their professional destinies (McCambridge, 2014). Just as predictive analytics can be used to raise one's insurance premium or increase one's chance of arrest, so they might be used to predetermine a child's place in a learning track or a teacher's employment prospects (Heath, 2014; Mayer-Schönberger & Cukier, 2013, pp. 151, 160; Podesta et al., 2014).

When it comes to educational research, we may attempt to anonymize data in order to evaluate curriculum and schools—however, even when names are stripped out, every person's data is unique and his or her profile remains visible in institutional data sets and on the web. With the self-same big data methods, identities can readily be inferred (Daries et al., 2014). Big data itself makes guaranteed anonymization hard to achieve.

Then there is the interventionary, and in some senses inherently manipulative, nature of embedded research. Perhaps the most notorious instance of this was the study in which 700,000 Facebook users were split into A and B groups who were then fed different mixes of positive and negative posts. "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks" was the alarming title of the subsequent paper in which the researchers report on the results of this experiment (Kramera, Guillory, & Hancock, 2014). Institutional review board approval for this project from Cornell University relied on consent via the research subjects' Facebook user agreement whereby the company owns personal data and can use them for a wide range of purposes. The same is the case with much learning management and other education software, where users effectively consent for their learning experiences to be used as data, including manipulation of those experiences for research and development purposes. The Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences concedes that traditional consent protocols are impractical in the context of big data and has recommended a new category of "excused" research (National Research Council, 2014). If big data is not to become Big Brother, users need to be recruited as co-collectors, co-analyzers, co-researchers—equal parties in the data-driven decisions that may today be made over their own lives (Nichols, Twidale, & Cunningham, 2012; Renear & Palmer, 2009; Wickett, Sacchi, Dubin, & Renear, 2012).

Conclusions

Discussion of "big data" in education is recent, at times making it sound like yet another passing educational fad. In this article, we have attempted to provide an overview of the developing scene, the continuities with traditional data sources and research methodologies, and a map of emerging potentials in the form of novel data sources and modes of analysis. We have focused on the example of writing in order to illustrate the range and complexity of data sources offering evidence of learning, not only in the subject area of literacy (writing as form) but as writing as a medium for knowledge representations and complex disciplinary performance across a range of discipline areas.

As is to be seen in unfolding developments in the field of technology-mediated writing and writing assessment, big data and education data sciences may in time offer learners, teachers, and researchers new windows into the dynamics and outcomes of learning, finely grained in their detail, varied in their sources and forms, and massive in their scope. However, much work still needs to be done in the nascent field of education data sciences before the affordances of computer-mediated learning can be fully realized in educational practice. For this reason, the case we have presented here is by necessity part description of an emergent reality and at the same time part agenda for future research and development. This is a journey that we have barely begun.

Note

1. U.S. Department of Education, Institute of Education Sciences, "The Assess-as-You-Go Writing Assistant: A Student Work Environment That Brings Together Formative and Summative Assessment" (R305A090394); "Assessing Complex Performance: A Postdoctoral Training Program Researching Students' Writing and Assessment in Digital Workspaces" (R305B110008); "u-Learn.net: An Anywhere/Anytime Formative Assessment and Learning Feedback Environment" (ED-IES-10-C-0018); "The Learning Element: A Lesson Planning and Curriculum Documentation Tool for Teachers" (ED-IES-lO-C-0021); and "InfoWriter: A Student Feedback and Formative Assessment Environment for Writing Information and Explanatory Texts" (ED-IES-13-C-0039). Bill and Melinda Gates Foundation, "Scholar Literacy Courseware." Scholar is located at http://CGScholar.com.

References

Abrams, S. S. (2013). Peer review and nuanced power structures: Writing and learning within the age of connectivism. *e-Learning and Digital Media*, *10*(4), 396–406.

Airasian, P. W., Bloom, B. S., & Carroll, J. B. (1971). *Mastery learning: Theory and practice* (J. H. Block, ed.). New York, NY: Holt Rinehart & Winston.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved from http://www.wired.com/2008/06/pb-theory/

Armour-Thomas, E., & Gordon, E. W. (2013). *Toward an understanding of assessment as a dynamic component of pedagogy*. Princeton NJ: Gordon Commission.

Ascaniis, S.de. (2012). Criteria for designing and evaluating argument diagramming tools from the point of view of argumentation theory. In N. Pinkwart & B. M. McLaren (Eds.), *Educational technologies for teaching argumentation skills* (pp. 3–27). Sharjah, UAE: Bentham Science.

Atzmueller, M. (2012). Data mining. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 75–94). Hershey, PA: IGI Global.

Bailey, A. L., Blackstock-Bernstein, A., Ryan, E., & Pitsoulaki, D. (in press). Data mining with natural language processing and corpus linguistics: Unlocking access to school-children's language in diverse contexts to improve instructional and assessment practices. In S. E. Atia, O. Zaiane, & D. Ipperciel (Eds.), *Data mining and learning analytics in educational research*. Malden, MA: Wiley-Blackwell.

Baker, E. L. (2007). *Moving to the next generation system design: Integrating cognition, assessment, and learning*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.

Baker, R. S. J.d., D'Mello, S. K., Rodrigo, M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human–Computer Studies*, *68*(4), 223–241.

Baker, R. S. J.d., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (2nd ed., pp. 253–274). New York, NY: Cambridge University Press.

Barnes, T., & Stamper, J. (2008, June). *Toward automatic hint generation for logic proof tutoring using historical student data*. Paper presented at the Intelligent Tutoring Systems 9th International Conference, Montreal, Canada.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, *19*(1/2), 205–220. doi:10.1007/s10758-014-9223-7

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Washington, DC: Office of Educational Technology, U.S. Department of Education.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2), 1–2.

Brandom, R. (1994). *Making it explicit: Reasoning, representing and discursive commitment*. Cambridge, MA: Harvard University Press.

Burrows, R., & Savage, M. (2014). After the crisis? Big data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), 1–6. doi:10.1177/2053951714540280

Burstein, J. (2009). Opportunities for natural language processing in education. In A. Gebulkh (Ed.), *Springer lecture notes in computer science* (Vol. 5449, pp. 6–27). New York, NY: Springer.

Burstein, J., & Chodorow, M. (2003). Directions in automated essay analysis. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 487–497). New York, NY: Oxford University Press.

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Studies in computational intelligence* (Vol. 62, pp. 183–221). Berlin, Germany: Springer-Verlag.

Cazden, C. B. (2001). *Classroom discourse: The language of teaching and learning* (2nd ed.). Portsmouth, NH: Heinemann.

Chang, H.-H. (2012). Making computerized adaptive testing diagnostic tools for schools. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 195–226). Charlotte, NC: Information Age.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20. doi:10.1007/s11336-014-9401-5

Chattopadhyay, I., & Lipson, H. (2014). Data smashing: Uncovering lurking order in data. *Journal of the Royal Society Interface*, *11*(101), 20140826–20140826. doi:10.1098/rsif.2014.0826

Chaudhri, V. K., Gunning, D., Lane, H. C., & Roschelle, J. (2013). Intelligent learning technologies: Applications of artificial intelligence to contemporary and emerging educational challenges. *AI Magazine*, *34*(3/4), 10–12.

Chi, M., Jordan, P., & VanLehn, K. (2014, June). *When is tutorial dialogue more effective than step-based tutoring?* Paper presented at the Intelligent Tutoring Systems 12th International Conference, Honolulu, HI.

Chinn, C. A. (2006). The microgenetic method: Current work and extensions to classroom research. In J. L. Green, G. Camilli, P. B. Elmore, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 439–456). Mahwah, NJ: Lawrence Erlbaum.

Chung, G. K. W.K. (2013). *Toward the relational management of educational measurement data*. Princeton, NJ: Gordon Commission.

Chung, G. K. W.K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary assessment* (pp. 23–40). Mahwah, NJ: Lawrence Erlbaum.

Conati, C., & Kardan, S. (2013). Student modeling: Supporting personalized instruction, from problem solving to exploratory open-ended activities. *AI Magazine*, *34*(3/4), 13–26.

Conrad, S., Clarke-Midura, J., & Klopfer, E. (2014). A framework for structuring learning assessment in a massively multiplayer online educational game: Experiment centered design. *International Journal of Game Based Learning*, *4*(1), 37–59.

Cope, B., & Kalantzis, M. (2009). "Multiliteracies": New literacies, new learning. *Pedagogies: An International Journal*, *4*, 164–195.

Cope, B., & Kalantzis, M. (2013). Towards a new learning: The "Scholar" social knowledge workspace, in theory and practice. *e-Learning and Digital Media*, *10*(4), 334–358.

Cope, B., & Kalantzis, M. (2014). Changing knowledge ecologies and the transformation of the scholarly journal. In B. Cope & A. Phillips (Eds.), *The future of the academic journal* (2nd ed., pp. 9–84). Oxford, UK: Elsevier.

Cope, B., & Kalantzis, M. (2015). Assessment and pedagogy in the era of machine-mediated learning. In T. Dragonas, K. J. Gergen, & S. McNamee (Eds.), *Education as social construction: Contributions to theory, research, and practice* (pp. 350–374). Chagrin Falls, OH: Worldshare Books.

Cope, B., Kalantzis, M., Abd-El-Khalick, F., & Bagley, E. (2013). Science in writing: Learning scientific argument in principle and practice. *e-Learning and Digital Media*, *10*(4), 420–441.

Cope, B., Kalantzis, M., & Magee, L. (2011). *Towards a semantic web: Connecting knowledge in academic research*. Cambridge UK: Woodhead.

Cope, B., Kalantzis, M., McCarthey, S., Vojak, C., & Kline, S. (2011). Technology-mediated writing assessments: Paradigms and principles. *Computers and Composition*, *28*(2), 79–96.

Cotos, E., & Pendar, N. (2007, September). *Automated diagnostic writing tests: Why? How?* Paper presented at the Technology for Second Language Learning Conference, Ames, IA.

Crossley, S., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015, March). *Pssst. . . Textual features. . . There is more to automatic essay scoring than just you!* Paper presented at the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082–1099. doi:10.1037/a0032674

D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., . . . Graesser, A. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems: 10th international conference, ITS 2010*, Pittsburgh, PA, USA, June 14-18, 2010, proceedings, part I (pp. 245–254). Berlin, Germany: Springer.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Seaton, D. T., . . . Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Queue*, *12*(7), 1–12.

Darling-Hammond, L., & Wood, G. H. (2008). *Assessment for the 21st century: Using performance assessments to measure student learning more effectively*. Stewart, OH: Forum for Education and Democracy.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational data mining: Applications and trends* (pp. 345–378). Heidelberg, Germany: Springer.

Data Science Association. (n.d.). *Data science code of professional conduct*. from http://www.datascienceassn.org/code-of-conduct.html

Dawson, S., & Siemens, G. (2014). Analytics to literacies: The development of a learning analytics framework for multiliteracies assessment. *International Review of Research in Open and Distributed Learning*, *25*(4).

Dede, C. (Ed.). (2015). *Data-intensive research in education: Current work and next steps*. Washington, DC: Computing Research Association.

DiCerbo, K. E., & Behrens, J. T. (2014). *Impacts of the digital ocean on education*. London, UK: Pearson.

Dixon-Román, E. J., & Gergen, K. J. (2013). *Epistemology in measurement: Paradigms and practices*. Princeton, NJ: Gordon Commission.

Dowell, N. M. M., & Graesser, A. C. (2014). Modelling learners' cognitive, affective, and social processes through language and discourse. *Journal of Learning Analytics*, *1*(3), 183–186.

Fancsali, S. E., Ritter, S., Stamper, J. C., & Berman, S. (2014, June). *Personalization, non-cognitive factors, and grain-size for measurement and analysis in intelligent tutoring systems*. Paper presented at the 2nd Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium, Pittsburgh, PA.

Farmer, F. R., & Glass, B. (2010). *Web reputation systems*. Sebastapol, CA: O'Reilly.

Gibson, A., & Kitto, K. (2015). Analysing reflective text for learning analytics: An approach using anomaly recontextualisation. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 275–279). New York, NY: ACM.

Glass, G. V. (2006). Meta-analysis: The quantitative synthesis of research findings. In J. L. Green, G. Camilli, P. B. Elmore, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 427–438). Mahwah, NJ: Lawrence Erlbaum.

Gorin, J. S. (2013). *Assessment as evidential reasoning*. Princeton, NJ: Gordon Commission.

Graesser, A. C., & McNamara, D. S. (2012). Reading instruction: Technology based supports for classroom instruction. In C. Dede & J. Richards (Eds.), *Digital teaching platforms: Customizing classroom learning for each student* (pp. 71–87). New York, NY: Teachers College Press.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, *22*(4), 39–51.

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014, July). *Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue*. Paper presented at the 7th International Conference on Educational Data Mining (EDM 2014), London, UK.

Hammer, P. (2015). Implications of and approaches to privacy in educational research. In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 92–94). Washington, DC: Computing Research Association.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.

Heath, J. (2014). Contemporary privacy theory contributions to learning analytics. *Journal of Learning Analytics*, *1*(1), 140–149.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

Ho, A. (2015). Before "data collection" comes "data creation." In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 34–36). Washington, DC: Computing Research Association.

Kaestle, C. (2013). *Testing policy in the United States: A historical perspective*. Princeton, NJ: Gordon Commission.

Kalantzis, M., & Cope, B. (2012). *Literacies*. Cambridge, UK: Cambridge University Press.

Kalantzis, M., & Cope, B. (2015a). Learning and new media. In D. Scott & E. Hargreaves (Eds.), *The Sage handbook of learning* (pp. 373–387). Thousand Oaks, CA: Sage.

Kalantzis, M., & Cope, B. (2015b). Regimes of literacy. In M. Hamilton, R. Hayden, K. Hibbert, & R. Stoke (Eds.), *Negotiating spaces for literacy learning: Multimodality and governmentality* (pp. 15–24). London, UK: Bloomsbury.

Kay, J. (2008). Lifelong learner modeling for lifelong personalized pervasive learning. *IEEE Transactions on Learning Technologies*, *1*(4), 215–228.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data and Society*, *1*(1), 1–12. doi:10.1177/2053951714528481

Kline, S., Letofsky, K., & Woodard, B. (2013). Democratizing classroom discourse: The challenge for online writing environments. *e-Learning and Digital Media*, *10*(4), 379–395.

Knight, S., Shum, S. B., & Littleton, K. (2013, April). *Epistemology, pedagogy, assessment and learning analytics*. Paper presented at the Third Conference on Learning Analytics and Knowledge (LAK 2013), Leuven, Belgium.

Knight, S., Shum, S. B., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics*, *1*(2), 23–47.

Koedinger, K. R., Baker, R. S. J.d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining* (pp. 43–56). Boca Raton FL: CRC Press.

Koedinger, K. R., Brunskill, E., Baker, R. S. J.d., & McLaughlin, E. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, *34*(3/4), 27–41.

Kramera, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(29), 8788–8790. doi:10.1073/pnas.1412469111 doi:10.1073/pnas.1412583111

Lammers, J. C., Magnifico, A. M., & Curwood, J. S. (2014). Exploring tools, places, and ways of being: Audience matters for developing writers. In K. E. Pytash & R. E. Ferdig (Eds.), *Exploring technology for writing and writing instruction* (pp. 186–201). Hershey, PA: IGI Global.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. New York, NY: Routledge.

Lane, H. C. (2013). Enhancing informal learning experiences with affect-aware technologies. In R. A. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 435–446). New York, NY: Oxford University Press.

Lansiquot, R. D. (2013). Real classrooms in virtual worlds: Scaffolding interdisciplinary collaborative writing. In A. Peña-Ayala (Ed.), *Intelligent and adaptive educational-learning systems: Achievements and trends* (pp. 269–292). Heidelberg, Germany: Springer.

Laurillard, D. (2012). *Teaching as a design science: Building pedagogical patterns for learning and technology*. London, UK: Routledge.

Li, C., & Girju, R. (2015, June). *Detecting causally embedded structures using an evolutionary algorithm*. Paper presented at the 3rd Workshop on Events: Definition, Detection, Coreference, and Representation, Denver, CO.

Lindgren, R., & Johnson-Glenberg, M. (2013). Emboldened by embodiment: Six precepts for research on embodied learning and mixed reality. *Educational Researcher*, *42*(8), 445–452. doi:10.3102/0013189x13511661

Lynch, Cl. (2008). How do your data grow? *Nature*, *455*(7209), 28–29.

Lynch, Co., Ashley, K. D., Pinkwart, N., & Aleven, V. (2009). Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, *19*, 253–266.

Lynch, C. F. (2014, July). *AGG: Augmented graph grammars for complex heterogeneous data*. Paper presented at the 7th International Conference on Educational Data Mining, London, UK.

Lynch, C. F., Ashley, K. D., & Chi, M. (2014, June). *Can diagrams predict essay grades?* Paper presented at the Intelligent Tutoring Systems 12th International Conference, Honolulu, HI.

Maroulis, S., Guimerà, R., Petry, H., Stringer, M. J., Gomez, L. M., Amaral, L. A. N., & Wilensky, U. (2010). Complex systems view of educational policy research. *Science*, *330*, 38–39.

Martin, R. C. (2009). *Agile software development, principles, patterns, and practices*. Upper Saddle River, NJ: Prentice Hall.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A Revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

McCambridge, R. (2014). Legacy of a failed foundation initiative: inBloom, Gates and Carnegie. *Nonprofit Quarterly*. Retrieved from https://nonprofitquarterly.org/2014/07/02/legacy-of-a-failed-foundation-initiative-inbloom-gates-and-carnegie/

McCarthey, S. J., Magnifico, A., Woodard, R., & Kline, S. (2014). Situating technology-facilitated feedback and revision: The case of Tom. In K. E. Pytash & R. E. Ferdig (Eds.), *Exploring technology for writing and writing instruction* (pp. 152–170). Hershey, PA: IGI Global.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 188–205). Hershey, PA: IGI Global.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

McNely, B. J., Gestwicki, P., Hill, J. H., Parli-Horne, P., & Johnson, E. (2012, April/May). *Learning analytics for collaborative writing: A prototype and case study*. Paper presented at the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Mislevy, R. J. (2013). *Four metaphors we need to understand assessment*. Princeton, NJ: Gordon Commission.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-Centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.

Mitros, P. (2015). Challenges in assessing complex skills in MOOCs. In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 36–39). Washington, DC: Computing Research Association.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Author.

National Research Council. (2014). *Proposed revisions to the common rule for the protection of human subjects in the behavioral and social sciences*. Washington, DC: National Academies Press.

Nichols, D. M., Twidale, M. B., & Cunningham, S.Jo. (2012, February). *Metadatapedia: A proposal for aggregating metadata on data archiving*. Paper presented at the iConference 2012, Toronto, ON.

Office of Educational Technology. (2016). *Future ready learning: Reimagining the role of technology in education*. Washington, DC: U.S. Department of Education.

Olmanson, J., Kennett, K., McCarthey, S., Searsmith, D., Cope, B., & Kalantzis, M. (2015). Visualizing revision: Leveraging student-generated between-draft diagramming data in support of academic writing development. *Technology, Knowledge and Learning*. Advance online publication.

O'Reilly, T. (2005). *What is web 2.0? Design patterns and business models for the next generation of software*. Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014, July). *Towards understanding expert coding of student disengagement in online learning*. Paper presented at the 36th Annual Cognitive Science Conference, Quebec City, Canada.

Partnership for 21st Century Skills. (2008). *21st century skills, education and competitiveness*. Washington, DC: Author.

Paul, M., & Girju, R. (2010, July). *A two-dimensional topic-aspect model for discovering multi-faceted topics*. Paper presented at the 24th AAAI Conference on Artificial Intelligence, Atlanta, GA.

Pea, R., & Jacks, D. (2014). *The Learning Analytics Workgroup: A report on building the field of learning analytics for personalized learning at scale*. Stanford, CA: Stanford University.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013, July). *Tuned models of peer assessment in MOOCs*. Paper presented at Educational Datamining 2013, Memphis, TN.

Piety, P. J. (2013). *Assessing the big data movement*. New York, NY: Teachers College Press.

Podesta, J., Pritzker, P., Moniz, E., Holdern, J., & Zients, J. (2014). *Big data: Seizing opportunities, preserving values*. Washington, DC: Executive Office of the President.

Ranade, G. V., & Varshney, L. R. (2012). To crowdsource or not to crowdsource? In *HCOMP2012* (pp. 150–156). Palo Alto, CA: AAAI Press.

Rekers, J., & Schürr, A. (1997). Defining and parsing visual languages with layered graph grammars. *Journal of Visual Languages & Computing*, *8*(1), 27–55.

Renear, A. H., & Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science*, *325*(5942), 828–832. doi:10.1126/science.1157784

Riaz, M., & Girju, R. (2010, September). *Another look at causality: Discovering scenario-specific contingency relationships with no supervision*. Paper presented at the Fourth IEEE International Conference on Semantic Computing, Pittsburgh, PA.

Roscoe, R. D., Brandon, R. D, Snow, E. L., & McNamara, D. S. (2014). Game-based writing strategy practice with the Writing Pal. In K. E. Pytash & R. E. Ferdig (Eds.), *Exploring technology for writing and writing instruction* (pp. 1–20). Hershey, PA: IGI Global.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*(4), 1010–1025. doi:10.1037/a0032340

Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining*, *4*(1), 1–10.

Ryan, K. E., & Shepard, L. A. (Eds.). (2008). *The future of test-based accountability*. New York, NY: Routledge.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885–899. doi:10.1177/0038038507080443

Schneider, Ba., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs*. Washington, DC: American Educational Research Association.

Schneider, Be., & Pea, R. (2014, July). *The effect of mutual gaze perception on students' verbal coordination*. Paper presented at the 7th International Conference on Educational Data Mining (EDM 2014), London, UK.

Schoenfeld, A. H. (2006). Design experiments. In J. L. Green, G. Camilli, P. B. Elmore, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 193–205). Mahwah, NJ: Lawrence Erlbaum.

Shepard, L. A. (2010). Formative assessment: Caveat emptor. *Educational Measurement: Issues and Practice*, *28*(3), 32–37.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. doi:10.1016/j.asw.2013.04.001

Shreeves, S. L. (2013). The role of repositories in the future of the journal. In B. Cope & A. Phillips (Eds.), *The future of the academic journal* (2nd ed., pp. 299–316). Cambridge, UK: Woodhead.

Shum, S. B., Sándor, Á., Goldsmith, R., Wang, X., Bass, R., & McWilliams, M. (2016, April). *Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool*. Paper presented at the Sixth International Learning Analytics and Knowledge Conference, Edinburgh, UK.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning and Media*, *1*(2), 1–11.

Siemens, G., & Baker, R. S. J.d. (2013, April/May). *Learning analytics and educational data mining: Towards communication and collaboration*. Paper presented at the Second Conference on Learning Analytics and Knowledge (LAK 2012), Vancouver, BC.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., . . . Baker, R. S. J.d. (2011). *Open learning*

analytics: An integrated and modularized platform. Society for Learning Analytics Research.

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013, April). *Analysis of collaborative writing processes using revision maps and probabilistic topic models*. Paper presented at the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Speck, J., Gualtieri, E., Naik, G., Nguyen, T., Cheung, K., Alexander, L., & Fenske, D. (2014, March). *ForumDash: Analyzing online discussion forums*. Paper presented at the First ACM Conference on Learning @ Scale, Atlanta, GA.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations*. London, UK: Department for International Development.

Stober, T., & Hansmann, U. (2009). *Agile software development: Best practices for large software development projects*. Berlin, Germany: Springer.

Strijbos, J.-W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional and conceptual developments. *Learning and Instruction*, *20*(4), 265–269.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York, NY: Doubleday.

Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data-enriched assessment. *Research and Practice in Assessment*, *9*(Winter), 5–16.

Tomkin, J. H., & Charlevoix, D. (2014, March). *Do professors matter? Using an a/b test to evaluate the impact of instructor involvement on MOOC student outcomes*. Paper presented at the First ACM Conference on Learning @ Scale, Atlanta, GA. Retrieved from http://dl.acm.org/citation.cfm?id=2556325

Twidale, M. B., Blake, C., & Gant, J. (2013, February). *Towards a data literate citizenry*. Paper presented at the iConference 2013, Fort Worth, TX.

Ullmann, T. D. (2015, September). *Keywords of written reflection: A comparison between reflective and descriptive datasets*. Paper presented at the 5th Workshop on Awareness and Reflection in Technology Enhanced Learning, Toledo, Spain.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

Vatrapu, R., Reimann, P., Bull, S., & Johnson, M. (2013, April). *An eye-tracking study of notational, informational, and emotional aspects of learning analytics representations*. Paper presented at the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, *28*(2), 97–111.

Wagner, E., & Yaskin, D. (2015). Predictive models based on behavioral patterns in higher education. In C. Dede (Ed.), *Data-intensive research in education: Current work and next steps* (pp. 23–31). Washington, DC: Computing Research Association.

Walton, D. (2008). *Informal logic: A pragmatic approach*. Cambridge, UK: Cambridge University Press.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–36.

Waters, J. K. (2014). Adaptive learning: Are we there yet? *Technological Horizons in Education*, *41*(4).

Weaver, W. (1948). Science and complexity. *American Scientist*, *36*, 536–544.

Wen, M., Yang, D., & Rose, C. (2014, July). *Sentiment analysis in MOOC discussion forums: What does it tell us?* Paper presented at the 7th International Conference on Educational Data Mining (EDM 2014), London, UK.

West, D. M. (2012). *Big data for education: Data mining, data analytics, and web dashboards*. Washington, DC: Brookings Institution.

Wickett, K. M., Sacchi, S., Dubin, D., & Renear, A. H. (2012, October). *Identifying content and levels of representation in scientific data*. Paper presented at the Proceedings of the American Society for Information Science and Technology, Baltimore, MD.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington IN: Solution Tree Press.

Winne, P. H. (2006). How software technologies can improve research on learning and bolster school reform. *Educational Psychologist*, *41*(1), 5–17. doi:10.1207/s15326985ep4101_3

Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning*, *9*(2), 229–237. doi:10.1007/s11409-014-9113-3

Winne, P. H., & Baker, R. S. J.d. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining*, *5*(1), 1–8.

Wise, A. F., Zhao, Y., & Hausknecht, S. N. (2013, April). *Learning analytics for online discussions: A pedagogical model for intervention with embedded and extracted analytics*. Paper presented at the Third Conference on Learning Analytics and Knowledge (LAK 2013), Leuven, Belgium.

Wixon, M., Arroyo, I., Muldner, K., Burleson, W., Rai, D., & Woolf, B. (2014, July). *The opportunities and limitations of scaling up sensor-free affect detection*. Paper presented at the 7th International Conference on Educational Data Mining (EDM 2014), London, UK.

Wolf, M. A. (2010). *Innovate to educate: System [re]design for personalized learning, a report from the 2010 symposium*. Washington, DC: Software and Information Industry Association.

Woolf, B. P. (2010). *A roadmap for education technology*. Global Resources for Online Education.

Worsley, M., & Blikstein, P. (2014, November). *Deciphering the practices and affordances of different reasoning strategies through multimodal learning analytics*. Paper presented at the Third Multimodal Learning Analytics Workshop, Istanbul, Turkey.

Xu, X., Murray, T., Woolf, B. P., & Smith, D. (2013, July). *Mining social deliberation in online communication: If you were me and I were you*. Paper presented at the 6th International Conference on Educational Data Mining (EDM 2013), Memphis, TN.

Xu, Y., Chang, K.-M., Yuan, Y., & Mostow, J. (2014, July). *Using EEG in knowledge tracing*. Paper presented at the 7th International Conference on Educational Data Mining, London UK.

Zwitter, A. (2014). Big data ethics. *Big Data and Society*, *1*(2), 1–6. doi:10.1177/2053951714559253

Authors

BILL COPE is a professor in the Department of Education Policy, Organization & Leadership, University of Illinois, Urbana-Champaign, USA and an adjunct professor in the Globalism Institute at RMIT University, Melbourne. He is also a director of Common Ground Publishing, developing and applying new publishing technologies. He is a former first assistant secretary in the Department of the Prime Minister and cabinet and director of the Office of Multicultural Affairs. His research interests include theories and practices of pedagogy, cultural and linguistic diversity, and new technologies of representation and communication. His recent research has focused on the development of new digital writing and assessment technologies, with the support of a number of major grants from the US Department of Education and the Bill and Melinda Gates foundations. The result has been the Scholar multimodal writing and assessment environment. Among his recent publications are an edited volume on *The Future of the Academic Journal*, Edn 2, Elsevier 2014, and with Kalantzis and Magee, *Towards a Semantic Web: Connecting Knowledge in Academic Research*, Elsevier, 2009.

MARY KALANTZIS is dean of the College of Education at the University of Illinois, Urbana-Champaign, USA. Before this, she was dean of the Faculty of Education, Language and Community Services at RMIT University, Melbourne, Australia, and president of the Australian Council of Deans of Education. She has been a board member of Teaching Australia: The National Institute for Quality Teaching and School Leadership, a Commissioner of the Australian Human Rights and Equal Opportunity Commission, chair of the Queensland Ethnic Affairs Ministerial Advisory Committee, vice president of the National Languages and Literacy Institute of Australia and a member of the Australia Council's Community Cultural Development Board. With Bill Cope, she is co-author or editor of a number of books, including: *The Powers of Literacy: Genre Approaches to Teaching Literacy*, Falmer Press, London, 1993, *Multiliteracies: Literacy Learning and the Design of Social Futures*, Routledge, London, 2000; *New Learning: Elements of a Science of Education*, Cambridge University Press, 2008 (2nd edition, 2012); Ubiquitous Learning, University of Illinois Press, 2009; *Literacies*, Cambridge University Press 2012 (2nd edition, 2016) and *A Pedagogy of Multiliteracies*, Palgrave, 2016.