Received: 19 June 2017 | Revised: 1 August 2018 | Accepted: 24 August 2018

DOI: 10.1111/jcal.12315

ORIGINAL ARTICLE

WILEY Journal of Computer Assisted Learning

Learning linkages: Integrating data streams of multiple modalities and timescales

Ran Liu[1] | John Stamper[1] | Jodi Davenport[2] | Scott Crossley[3] | Danielle McNamara[4] | Kalonji Nzinga[5] | Bruce Sherin[5]

[1]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania

[2]WestEd, San Francisco, California

[3]Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, Georgia

[4]Department of Psychology, Arizona State University, Tempe, Arizona

[5]Learning Sciences Department, Northwestern University, Evanston, Illinois

Correspondence
Ran Liu, Carnegie Mellon University, Pittsburgh, PA.
Email: ranliu@gmail.com

Funding information
National Science Foundation; Division of Research on Learning in Formal and Informal Settings, Grant/Award Numbers: 1417997, 1418020, 1418072 and 1418181; Institute of Education Sciences, Grant/Award Numbers: R305A100069 and R305A170049

Abstract

Increasingly, student work is being conducted on computers and online, producing vast amounts of learning-related data. The educational analytics fields have produced many insights about learning based solely on tutoring systems' automatically logged data, or "log data." But log data leave out important contextual information about the learning experience. For example, a student working at a computer might be working independently with few outside influences. Alternatively, he or she might be in a lively classroom, with other students around, talking and offering suggestions. Tools that capture these other experiences have potential to augment and complement log data. However, the collection of rich, multimodal data streams and the increased complexity and heterogeneity in the resulting data pose many challenges to researchers. Here, we present two empirical studies that take advantage of multimodal data sources to enrich our understanding of student learning. We leverage and extend quantitative models of student learning to incorporate insights derived jointly from data collected in multiple modalities (log data, video, and high-fidelity audio) and contexts (individual vs. collaborative classroom learning). We discuss the unique benefits of multimodal data and present methods that take advantage of such benefits while easing the burden on researchers' time and effort.

KEYWORDS

collaborative learning, intelligent tutoring systems, log data, multi-modal data analytics, natural language processing, STEM education

1 | INTRODUCTION

Student work is increasingly conducted on computers and online, producing vast amounts of learning-related data. At the same time, advances in computing, data mining, and learning analytics are providing new tools for the collection, analysis, and representation of this data. Together, the available data and analytical tools enable smart and responsive systems with strong potential to personalize learning experiences for individual learners.

At a basic level, we can access data logged by computer systems, which provides indicators of students' behaviours within the system, such as time spent on screens or modules, keystrokes, and responses to questions or problems. We have gained many insights into learning behaviour through analyses of tutoring systems' software-logged data, or "log data." Ideally, such data will be used to create a rich picture of student knowledge and learning processes as they unfold across time (e.g., Graesser, Conley, & Olney, 2012). There is more to the learning experience, however, than can be revealed from log data alone. For example, a student working at a computer might be working independently with few outside influences. Alternatively, he or she might be in a lively classroom, with other students around him or her, talking and even offering suggestions. Data that capture these other experiences have potential to augment and complement log data. In some cases, these additional data may lead to critical insights.

 © 2018 John Wiley & Sons Ltd

In practice, however, the increased complexity of data resulting from adding new, multimodal data streams from different sources can create many challenges. These data are often collected at different grain sizes, which are difficult to integrate. Making sense of data at many levels of analysis, including the most detailed levels, is highly time-consuming. Imagine trying to understand the detailed sequence of events a student exhibits as he or she engages in productive struggle with a difficult concept and the social interactions surrounding this struggle. To fully understand the events that unfold even in this small segment of a student's educational experience, a researcher may need to watch screen video data, listen to the audio dialogue several times, and enter behavioural codes into a separate document. Having to do this for every problem and concept students experience over the course of even one class period of learning technology use would be vastly taxing on human time and effort. Yet this level of detailed analysis provides interesting and temporally rich insights (Worsley, 2014), in contrast to purely quantitative models based solely on coarse-level "correctness" coding. Despite the challenges, progress is being made on ways to integrate multimodal streams of data (Blikstein, 2013; Blikstein & Worsley, 2016).

In this paper, we present two empirical studies that use multimodal data sources to enrich our understanding of student learning. We leverage and extend quantitative models of student learning to incorporate insights derived jointly from streams of data collected in multiple modalities (log data, video, and high-fidelity audio) and contexts (individual vs. collaborative classroom learning). We aim to develop more robust and predictive models of student learning and behaviour. These enhanced, multimodal models provide a more holistic picture about learners and potential success factors for learning.

We also address some critical questions of interest to the fields of educational research, educational data mining, learning analytics, and the learning sciences: What types of learning phenomena can we capture and trace with computer-collected data, and what types do we miss? And are there ways we can enrich computer-collected data by collecting and analysing multimodal data streams, without a massive additional demand on researchers' time and effort?

2 | STUDY 1: CHEMISTRY VIRTUAL LAB TUTOR WITH CAMTASIA VIDEO RECORDINGS

Study 1 data were collected from a classroom study, in which students engaged in a Chemistry "Virtual Lab" tutoring system. ChemVLab+ (chemvlab.org) provides a set of high school chemistry activities designed to build conceptual understanding and inquiry (Davenport, Rafferty, Karabinos, & Yaron, 2015; Davenport, Rafferty, Yaron, Karabinos, & Timms, 2014). Conceptual understanding in chemistry requires students to connect quantitative calculations, chemical processes at the microscopic level (e.g., atoms and molecules), and outcomes at the macroscopic level (e.g., concentrations, colour, and temperature). ChemVLab+ activities are designed to help students connect procedural knowledge of mathematical formalisms with authentic chemistry learning by allowing them to design and carry out experiments. In each activity, students work through a series of tasks to solve an authentic problem and receive immediate, individualized tutoring. As students work, teachers are able to track student progress throughout the activity and assist students that may be lagging behind. Upon completion of the activities, students receive a report of their proficiency on targeted concepts and skills, and teachers can view summary reports that show areas of mastery or difficulty for their students. In the current study, students completed four activities—PowderAde: Using Sports Drinks to Explore Concentration and Dilution, The Factory: Using a City Water System to Explore Dilution, Gravimetric Analysis, and Bioremediation of Oil Spills.

There are a variety of types of interfaces across the four modules, but students spend a significant portion of their time working in open-ended activities such as setting up experiments in a "virtual" laboratory environment (e.g., Figure 1) and making observations.

2.1 | Data collection

Participants were 59 students at a high school in the greater Pittsburgh area enrolled in honors chemistry classes. They participated in four stoichiometry activities of the ChemVLab+ educational tutor. They completed these activities across four 50-min class periods spread over the course of 3 weeks. Before students engaged with ChemVLab+, they completed paper pretest assessments. After the four class periods devoted to using the tutoring system, students completed paper posttest assessments.

Using Camtasia, we collected screen video captures for 47 consenting students during the second and fourth class periods of the study. These multimodal data streams covered the second and fourth ChemVLab+ activities (The Factory: Using a City Water System to Explore Dilution and Bioremediation of Oil Spills, respectively) for the majority of students. Student-facing webcam data were additionally collected from a subset of 25 students who consented to additional recording in addition to the screen video recording. All of the video recordings were initially stored as Camtasia project files but were then exported to MP4 format. There was one video file for each student for each class period, and this video file contained both the screen activity and the student-facing webcam activity (if applicable to that student). All files were labelled with students' anonymized IDs and stored in a secured research hard drive. In total, we collected 90 videos, each averaging about 35 min in length, for a total of approximately 3,150 min of total video time.

2.2 | Analytic approach

The student-centred video data captured rich and detailed student interactions with the ChemVLab+ interface and with peers, the teacher, and/or the experimenter. In general, however, most of the activity captured in the videos was interactions between the student and the interface. There was minimal dialogue, as students were working independently the majority of the time. Because of this, we focused our analyses on coding student behaviours as they interacted with the tutor, within the screen video data.

One of the biggest challenges in multimodal learning analytics is that the large volume of rich, multimodal data collected requires significant human time and effort to make sense of. To gain the most insight from these multimodal data streams beyond what could be obtained through automatic software-logged data, we focused on parts of the activities where students struggled the most. We used the software-

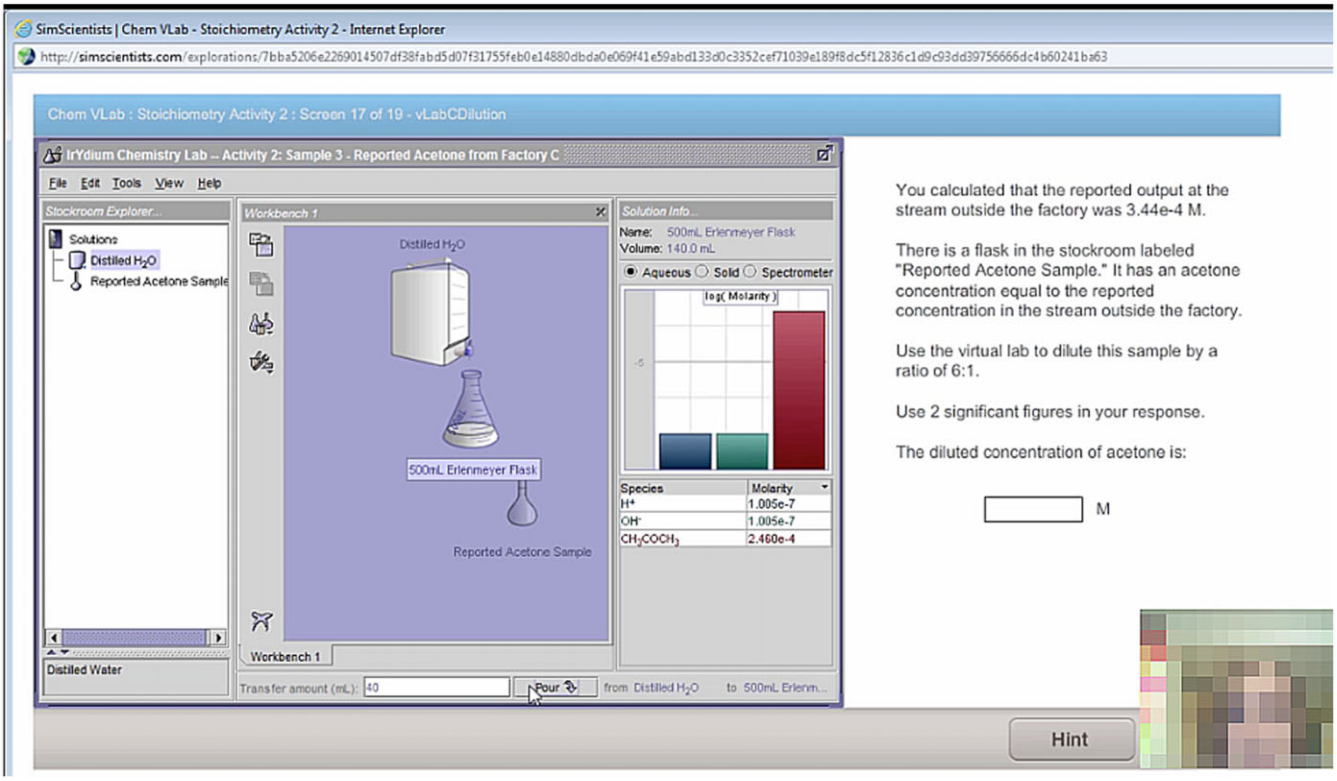

FIGURE 1 Example interface for the experimentation portions of the ChemVLab+ tutor [Colour figure can be viewed at wileyonlinelibrary.com]

logged data to identify these moments and then focused on these moments to do in-depth video analyses.

We then used a set of tools called STREAMS (Structured TRansactional Event Analysis of Multimodal Streams) to facilitate the integration of log data and multimodal data streams (Liu, Davenport, & Stamper, 2016; Liu et al., 2016). These tools allow us to discover insights that uniquely leverage the strengths of both log and multimodal data. STREAMS supports (1) easy temporal alignment of software-logged usage data to any number of additional data streams and (2) log data query-based extraction of video segments.

The first component of STREAMS temporally aligns different multimodal streams of data (video, audio, etc.) with log data and, consequently, to each other. The tool uses the relative times between log data events, combined with the temporal offset between the logged data and the beginning of each media stream, to accomplish the alignment. If the temporal offset is not automatically recorded during data collection, the system provides a command-line-based interface that allows the researcher to provide the time within each media stream at which the first software-logged event occurred. For the videos that we analysed, this part of the process took 30 min of human input. The output of temporal alignment is a data frame that contains the original log data, but with three additional columns per synced media stream: the corresponding media stream's filename, the start time of the event within that stream, and the end time of the event within that stream.

The second component of STREAMS accomplishes log data query-based extraction of video segments. In this component, the user can query any value of any column from the software-logged data (e.g., all problem steps tagged with skill X) or any combination of column values (e.g., all problem steps tagged with skill X on which the student made an incorrect first attempt). STREAMS will then produce a folder of extracted video segments that correspond specifically to the events specified in that query. These video segments can then be swiftly coded by researchers. The code for the tool is available at github.com/ranolabar/STREAMS.

We first mined the software-logged data to identify the single problem screen with the highest error rate across activities. Identifying moments with high error rates can reveal common conceptual difficulties, difficulties with the tutoring interface, and students' metacognitive strategies (Mathan & Koedinger, 2005), as these moments showcase how students handle difficult problems that they do not yet know how to solve.

We then used STREAMS to extract all of the relevant screen video segments pertaining to this problem screen for students who did not get the problem correct on the first attempt. The extracted video data totalled approximately 188 min (just under 6% of the total length of all video data collected). We coded the extracted video data at multiple levels and then used these codes to address the following questions:

- How do the students' behaviours following their initial failure with the problem affect their later performance on problems requiring the same concept?
- How do the students' behaviours reveal different metacognitive strategies, and do these strategies relate to learning outcomes?

In addition, we tested whether the insights derived from observing and coding student behaviours during those video segments could improve upon a baseline quantitative model based solely on the software-logged data.

2.3 | Qualitative analyses

The highest aggregate error rate, 85%, across students (Figure 2) was found on the first task in the Bioremediation of Oil Spills activity (http://chemvlab.org/activities/activity.php?id=4). This problem required students to create balanced chemical equations based on an image of reactions that occur when two solutions are combined. Chemical equations describe the process of how atoms in molecules (the reactants) recombine to form new molecules (the products). In the chemical equation activity, the diagram labels two types of atoms, A and B, and students need to determine what molecules existed before the solutions were combined (the reactants) and after (the products).

The most common incorrect strategy students exhibited was mistaking the chemical equation as describing the state of the system (e.g., how many molecules are present in the reactant and the product containers) as opposed to describing the process of a reaction (e.g., the rules by which molecules combine). The videos revealed a variety of indicators of this misconception. Many students approached the problem by using the diagram very literally. That is, they would count the four A and 10 B molecules in the reactant containers and use those as coefficients for A and B on the reactant side (on the left). They would count the four AB2 molecules in the product container and the two leftover B molecules and use those as the coefficients for AB2 and B on the product side (on the right). This process was evident in the video clips as the cursor was visibly moving over the diagram on the screen, and students could be heard quietly counting. As the students' approach was deliberate and coherent, the data suggest the error was due to a misconception as an "alternative conception" (Chi, 2005), rather than a fragmented understanding (diSessa, 1988). In other words, students appear to have a systematic strategy, but it is not a correct one.

The video analysis also revealed that interface limitations prevented students from making certain errors related to this misconception. If the students conceive as the chemical equation describing the state of the system rather than a process of reaction, they would count each atom in the diagram. In the diagram, there are 10 molecules of B; however, the interface does not allow two-digit coefficients. The inability to respond with a two-digit number indicated to students that the purely visual strategy was not correct. This was evidenced in the videos: Of the 39 videos in which students did not succeed on their first attempts at this problem, in 22 video segments, we observed the student type "1" and repeatedly hitting another key while being frustrated that it would not show up in the text box. In one video, a student is seen initiating dialogue with a nearby peer in which he says "Do you know why it won't take 10? [inaudible peer response] But there's 10 of them!"

2.4 | Video-coded behaviour produces quantitative modelling improvements

On the basis of these qualitative observations, we coded for the presence of certain common behaviours in the video segments. These behaviours are summarized in Table 1. The coding scheme was binary, indicating the presence or absence of each behaviour for each student. Two researchers independently coded the data based on a prior agreed-upon description of evidence of each behaviour (Table 1, right column). Any disagreements in coding for an individual behaviour were then discussed by both researchers until an agreement was reached.

We then investigated whether the insights derived from these coded behaviours would improve upon a baseline quantitative model based solely on the software-logged data. As a baseline model for this comparison, we used the additive factors model on data from the Bioremediation of Oil Spills activity, which contained the problem screen of interest. The additive factors model is a logistic regression model that extends the Rasch model from item response theory (Rasch, 1993) by incorporating a growth/learning term (Cen, Koedinger, & Junker, 2006; Draney, Wilson, & Pirolli, 1996; Spada & McGaw,

FIGURE 2 The problem screen with the highest aggregate error rate across all activities. This problem required the application of the Balance Chemical Reactions concept [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Behaviours coded for in the video analysis and the evidence that constituted coding the presence of each behaviour

Behaviour	Evidence of behaviour
"Literal" misconception	Student attempting 4A + 10B → 4AB + 2B or similar variants (such as 4A + 5B + 5B → 4AB + 2B upon discovering the interface one-digit limitation) Student counting the molecules in the diagram with mouse
Attempting a "10" coefficient for B on the reactants side	Student typing a "1" into the coefficient box followed by audible keystrokes that do not register in the field Any verbal acknowledgement of trying to attempt "10" in the box
Leftover molecule misconception	Student officially types a response with any coefficient for B, including both submitted (software-logged) and nonsubmitted responses
Removed leftover molecules	Student removes a leftover molecule (B coefficient) from a prior attempt prior to reaching a bottom-out hint
Reach unreduced equation	Student arrives at an unreduced-coefficients version of the correct balanced reaction (e.g., 4A + 8B → 4AB2) prior to reaching a bottom-out hint
Bottom-out hint	Student reaches a bottom-out hint on either the reactant or product side
External dialogue	Student arrives at the answer with outside-of-tutor intervention (either by asking the teacher, the experimenter, or a peer)

1985). It models the probability that a given student will get a problem correct on his or her first attempt based on estimates of the student's ability, the difficulty of the concept(s) required on each problem, and the improvement in that concept with each problem the student is required to apply it. We used R (R Team, 2014) and the package *lme4* to construct a linear mixed effects (LMEs) model implementation for the additive factors model and to obtain the Akaike information criterion (AIC) and Bayesian information criterion (BIC) metrics reported here.

The problem screen of interest was the first problem on which students had to apply the Balancing Chemical Reactions concept within the Bioremediation of Oil Spills activity. We sought to discover whether splitting the students into groups based on their video-coded activity would improve the model's prediction of subsequent performance on problems involving the Balancing Chemical Reactions concept. To this end, we test several variations on the baseline model using the video-coded behaviours to group students, to discover which video-coded features produced significant improvements to the model's goodness-of-fit to the data. Goodness-of-fit was measured using AIC and BIC. Both criteria are likelihood-based measures of predictive fit that penalize for model complexity. For both criteria, lower numbers indicate a better relative model fit.

The top row of Table 2 shows the fit of the baseline additive factors model to the data from the Bioremediation of Oil Spills activity. The best improvement to this baseline model was generated by splitting students who incorrectly attempted the video-coded Balancing Chemical Reactions problem (Figure 1) based on whether they received information about the correct answer (either through external dialogue or by seeing a bottom-out hint that provided the answer), or not. This model's fit-to-data is shown in the bottom row of Table 2 and shows a substantial drop in both AIC and BIC, indicating a better model fit.

Because this split was not applicable to students who got the problem right on their first attempt, these students were treated as a separate group. A more stringent baseline model, which just split students based on their first-attempt correctness on the problem screen of interest, was included as an additional control (second row of Table 2). The model with an additional split for video coded data (bottom row of Table 2) fit better yet than this more stringent control model.

Figure 3 shows the aggregate learning trajectories of students classified into each of the three groups: correct on first attempt (green), incorrect on first attempt and required either a bottom-out hint (BOH) or external dialogue to reach the answer (red), or incorrect on first-attempt and did not require either to reach the answer (blue). Though there is some fluctuation between the relative performance of correct on first attempt (green) and incorrect on first attempt and did not require BOH or external dialogue (blue) at problems 1–4, the differences are not significant.

These results show that, among students who incorrectly attempted the first Balancing Chemical Reactions problem, students who received external information about the correct answer, through either external dialogue or a bottom-out hint (Figure 3, red line), before submitting the correct answer had slower learning trajectories with ultimately lower performance across the board on the Balancing Chemical Reactions skill. One possible explanation for this finding is that students who reached the correct answer *without* either of these activities engaged in productive struggle (also referred to as productive failure; Kapur, 2010, 2012). Engaging in productive struggle without resorting to assistance that guides directly to the correct answer has proven particularly effective for learning in some difficult problem-solving contexts (Kapur, 2012). Productive struggle may help students become aware of their knowledge gaps, which eases the process of repairing their misconceived mental models.

TABLE 2 Comparative model fits based on AIC and BIC, likelihood-based measures of predictive fit that penalize for model complexity. For both criteria, lower numbers indicate a better relative mode

Model	AIC	BIC
Baseline	1,085.959	1,115.241
Baseline + Balancing Chem Reactions concept split by success on problem with highest error rate	1,064.283	1,093.565
Baseline + Balancing Chem Reactions concept split by success on problem with highest error rate *and* video-coded activity	1,058.018	1,087.3

Note. AIC: Akaike information criterion; BIC: Bayesian information criterion.

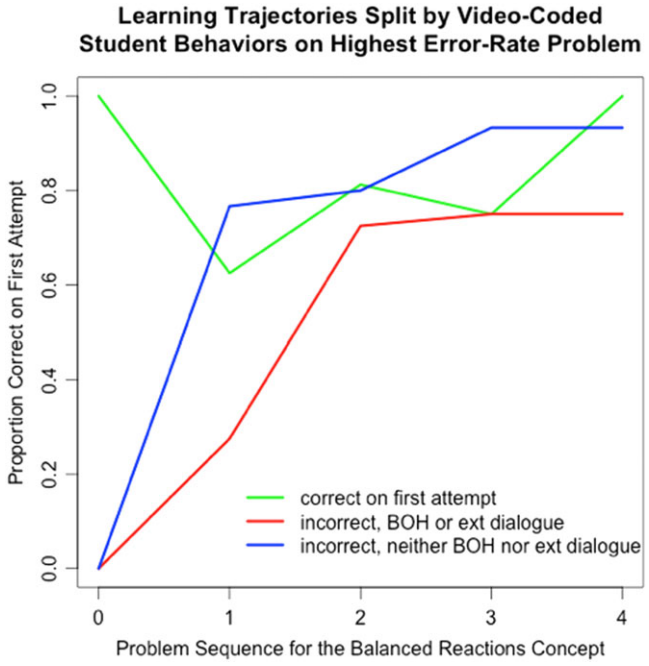

FIGURE 3 Aggregate learning trajectories of students classified into each of the three groups: correct on first attempt, incorrect on first attempt and required either a bottom-out hint (BOH) or external dialogue to reach the answer, or incorrect on first attempt [Colour figure can be viewed at wileyonlinelibrary.com]

An alternative explanation might be that students who reached solutions without many hints had a couple of potential solutions in mind and simply realized (after getting feedback from the system) that their initial approach was incorrect. However, this interpretation is unlikely as the video clips showed a very common progression of errors from the initial attempt to the correct attempt. This progression indicates students did not initially understand that the chemical reaction describes a process, rather than the state of the system. Specifically, students began by counting the number of molecules present in the reactant and product containers ("literal" misconception) and/or attempted to type in a "10" coefficient for B on the reactants side as opposed to describing the process of a reaction (e.g., attending to which new molecules are being formed). Next, students typically progressed to exhibiting evidence of the error of not understanding that leftover molecules are not part of the reaction by typing a response with any coefficient for B on the products side (leftover molecule misconception or removed leftover molecules on Table 1). Finally, students progressed to realizing that they needed to present the equation with coefficient values simplified (i.e., A + 2B = AB2 rather than 4A + 8B = 4AB2). Sixty percent of students who produced an error on their first attempt did not reach a bottom-out hint and did not receive external assistance exhibited this specific progression of errors. This provides further evidence that students were struggling with creating a correct understanding (without being handed the correct answer) along the way.

Although the detailed video coding focused on one specific problem out of the entire set of ChemVLab+ activities, the moments where students solve this problem were impressively revealing of rich details about their learning and help-seeking behaviours. Furthermore, using the video-coded information in conjunction with quantitative modelling allowed us to identify features that uniquely predict subsequent learning trajectories for the concept of interest. The study provides evidence that analysing multimodal data even during limited moments of student activity can lead to unique insights—that log data would not have been able to provide—with implications for quantitative modelling of subsequent learning trajectories.

To further test the power of multimodal models for identifying critical information about learners and isolating potential success factors, we carried out an additional study using a highly collaborative math-based tutor with elementary students. Though the multimodal data available were very different, our research questions remained the same, that is, what types of learning phenomena can we capture and trace with computer-collected data, and what types do we miss? And are there ways we can enrich computer-collected data by collecting and analysing multimodal data streams, without a massive additional demand on researchers' time and effort?

3 | STUDY 2: COLLABORATIVE FRACTION TUTOR WITH HIGH-FIDELITY AUDIO RECORDINGS

Study 2 data were collected from a classroom study of students working on the Collaborative Fraction Tutor (Olsen, Belenky, Aleven, & Rummel, 2014; Olsen, Aleven, & Rummel, 2015), an intelligent tutoring system developed by researchers at Carnegie Mellon University that helps students become better at understanding and working fractions. The tutor was created using Cognitive Tutor Authoring Tools, which facilitate rapid development and easy deployment of intelligent tutors. The tutor supports collaboration between student partners to learn fraction skills such as addition (Figure 4), subtraction, comparing fractions to determine which is larger or smaller, finding the least common denominator, and finding equivalent fractions. The tutor's effectiveness has been demonstrated in prior classroom deployment studies (Rau, Aleven, & Rummel, 2009; Rau, Aleven, Rummel, & Rohrbach, 2012). These studies showed that students' mistakes decrease as they progress through the tutor; students score higher on a fractions test after using the tutoring system compared with before; and scores remain higher than pretutoring a week after they have finished using the tutoring system.

Although each student worked on the tutoring system on his or her own computer screen, each student in a pair could control only part of the screen. The students needed to work together to finish the problem (i.e., one student could not do everything). Students worked together at the same time and, ideally, talked about what they were doing, asked for help from their partner, defended a position or explained why they thought something was the correct answer, and built off of each other's contributions.

All collaborative dyads randomly received a problem set focused on either procedural or conceptual knowledge building. The procedural versus conceptual comparison had been included to investigate whether there were any interactions between collaborative learning and type of knowledge acquired. Figure 4 shows one example of each type of problem, conceptual and procedural. The top and bottom panels show example problems from the conceptual and procedural knowledge conditions,

respectively. The figure depicts correctly completed screens; student-input fields are marked with either green text or borders.

3.1 | Data collection

Participants were 104 fourth and fifth graders from two schools in the greater Pittsburgh area. There were 19 fifth graders from one classroom of one school and 50 fourth graders and 35 fifth graders from the second school. Of these, only a subset of students were present for the full study, had the same partner during the entire study (no absences for either individual), and consented to audio recording of their dialogue. For consistency purposes, we only analysed data from students who fit all of these criteria. Thus, our analyses were conducted on this subset of 36 students (14 fifth graders from the first

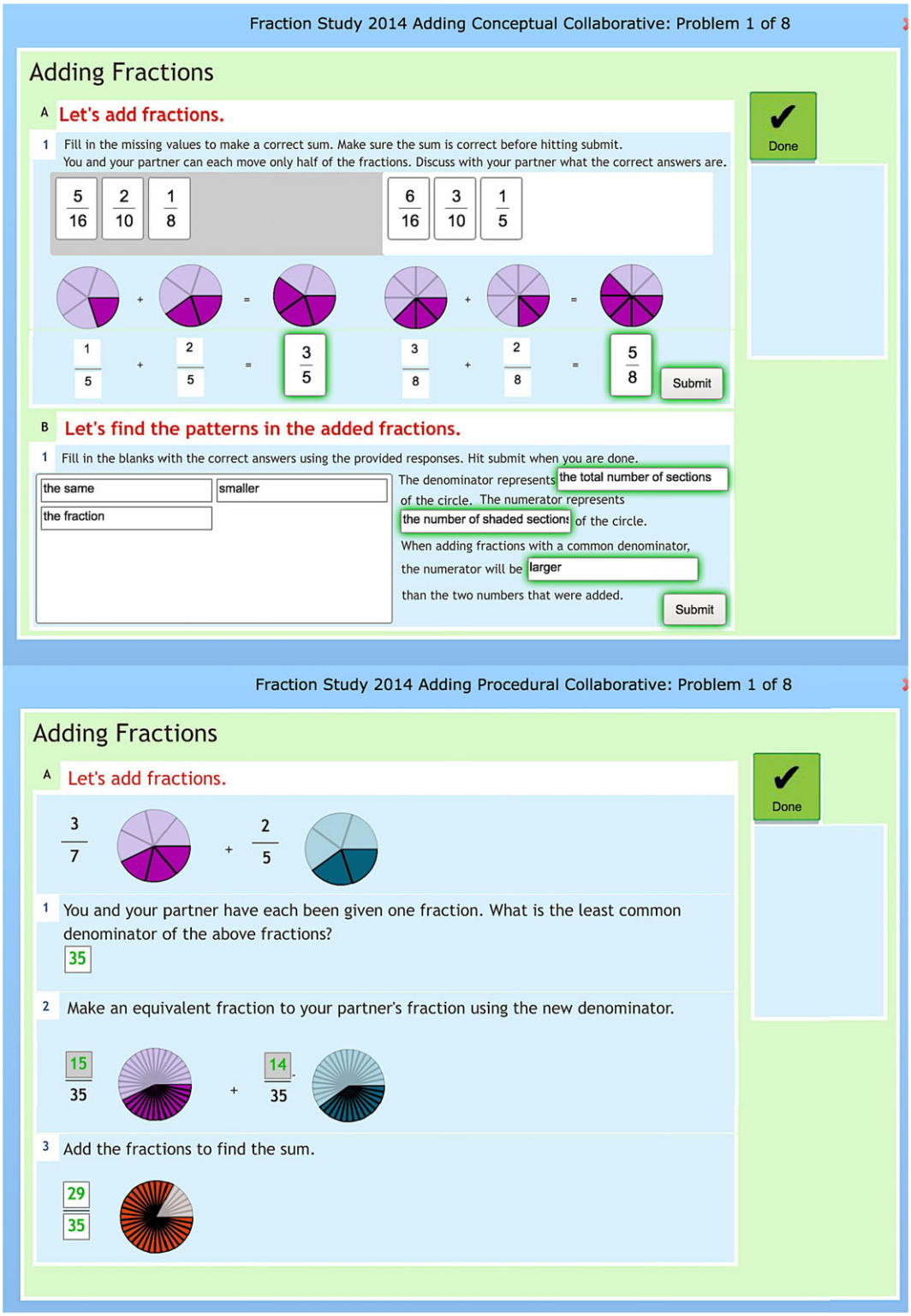

FIGURE 4 Example interface for the "Adding Fractions" skill in the collaborative fraction tutor. The top panel shows the interface for the conceptual condition. The bottom panel shows the interface for the procedural condition [Colour figure can be viewed at wileyonlinelibrary.com]

school and 16 fourth graders and six fifth graders from the other school). There were 15 males and 21 females in the analysis subset. Student pairs were determined by the teachers. Teachers were asked to pair each student with a partner with whom they would get along and who was at a similar knowledge level.

The study took place over five consecutive days. On the first day, students individually took a pretest to establish their baseline fractions knowledge. In the following 3 days, students worked through the tutoring system with a partner. On the last day, students individually took a posttest that also tested fractions knowledge, with content similar to the pretest.

For the consenting students, high-quality audio data were collected for each individual student using a headset outfitted with a microphone. The microphone was linked to a tablet computer to store the recordings. In each class, we additionally collected full-classroom video recorded from one camera located in the corner of the room.

3.2 | Analytic approach

Due to the collaborative nature of the learning experience, we anticipated that dialogue would be plentiful and that much of the interesting and rich learning phenomena would be captured in this dialogue. By transcribing the high-fidelity audio data collected through headset microphones and using automated natural language processing (NLP) tools, we were able to make use of large quantities of dialogue data without direct coding by human researchers. NLP involves the automatic extraction of linguistic features using a computer programming language. NLP has the potential to provide information about language at multiple levels and dimensions (Graesser & McNamara, 2011). Thus, as an initial pass at analysing these data, we took the approach of applying NLP analyses to the transcribed dialogue data.

In particular, we focused on whether linguistic factors related to lexical sophistication, cohesion, and affect present in students' collaborative dialogue predict unique variance in math performance beyond what is accounted for by nonlinguistic factors.

3.3 | Analyses

3.3.1 | Transcription

First, we temporally synced the recordings between the two members of each dyad and merged them so that professional transcribers would work off of one recording for each dyad's conversation. A professional transcriber transcribed each of the speech samples collected from the participants. The transcriptions contained the speaker's words, some metalinguistic data (singing, laughing, and sighing), and filler words (e.g., ummm and ahhhh). Disfluencies that were linguistic in nature (e.g., false starts, word repetition, and repairs) were also retained. If any portion of the audio was not able to be transcribed, the words were annotated with either an underscore or the flag "INAUDIBLE" depending on the transcriptionist. The files were cleaned so that metalinguistic data, filler words, and portions unable to be transcribed were removed prior to analysis.

3.3.2 | Linguistic variables

Transcripts were separated by learner and cleaned to remove all nonlinguistic information including metadata and nonlinguistic vocalizations such as coughs and laughs. Each transcript was run through three NLP tools: the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015), the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016), and the SEntiment ANalysis and Cognition Engine (SEANCE; Crossley, Kyle, & McNamara, 2017). These tools report on language features related to lexical sophistication, text cohesion, and sentiment analysis, respectively. These are discussed briefly below.

TAACO indices

TAACO incorporates over 150 classic and recently developed indices related to text cohesion. For many indices, the tool incorporates a part of speech (POS) tagger and synonym sets. The POS tagger allows TAACO to report on indices of cohesion that are specific to content words such as nouns, verbs, and adjectives. As well, TAACO provides linguistic counts for both sentence and paragraph markers of cohesion. Specifically, TAACO calculates sentence overlap indices for words and lemmas, paragraph overlap indices for words and lemmas, and a variety of connective indices such as sentence linking connectives (e.g., *nonetheless*, *therefore*, and *however*).

TAALES indices

TAALES incorporates about 150 indices related to basic lexical information (e.g., type and token counts for both content and function words), lexical frequency (i.e., how frequent words are), psycholinguistic word information (e.g., concreteness and meaningfulness), academic language for both single word and multiword units (i.e., academic word and formula lists), and word polysemy (i.e., the number of senses a word has) and hypernymy (the specificity of a word).

SEANCE indices

SEANCE is a sentiment analysis tool that relies on a number of preexisting sentiment, social positioning, and cognition dictionaries. SEANCE contains a number of predeveloped word vectors taken from freely available source databases that were developed to measure sentiment, cognition, and social order. For many of these vectors, SEANCE also provides a negation feature (i.e., a contextual valence shifter) that ignores positive terms that are negated (e.g., not happy). SEANCE also includes a POS tagger. A number of these vectors have been combined use statistical analyses to create component scores that relate to large sentiment, social positioning, and cognition construct (e.g., terms related to respect).

3.3.3 | Statistical analyses

We first conducted a paired samples *t* test between students' pretest and posttest scores to assess evidence of learning from the collaborative fraction tutor. We then conducted LME models to determine if linguistic features in the students' language output, along with other fixed effects, successfully predicted students' pretest and posttest math scores. Thus, the LME model modelled the pretest and posttest results in terms of random or within-subjects factors (i.e., repeated

variance explained by the students as they moved through the intervention longitudinally) and fixed or between-subjects factors (e.g., the linguistic features in their transcripts, gender, age, and school). Such an approach allows us to examine math growth over time for individual learners using random factors and to investigate if individual differences related to the learner such as demographic information, age, and linguistic ability predict math development.

We used R (R Team, 2014) for our statistical analysis and the package *lme4* to construct LMEs models. We also used the package *lmerTest* to analyse the LME output and derive *p* values for individual fixed effects. We used stepwise variable selection for the final linguistic model, and interpretation of the model was based on *t* and *p* values for fixed effects and visual inspection of residuals distribution. To obtain a measure of effect sizes, we computed correlations between fitted and predicted residual values, resulting in an R^2 value for both the fixed factors and the fixed factors combined with the random factor (i.e., the repeated participant data from the pretest and the posttest).

We first developed a baseline model that included gender, grade, condition, and school as fixed effects and participants as random effect. We next developed a full model that included gender, grade, condition, and school as fixed effects along with linguistic features and participants as random effects.

3.4 | Results

3.4.1 | Math gains

A paired *t* test examining differences between the pretest and the posttests scores indicated significant differences between the pretest (M = 0.469, SD = 0.170) and the posttest (M = 0.603, SD = .185); $t(35)$ = 5.988, p < 0.001. Overall, students improved in their knowledge of fraction concepts and procedures after engaging with the collaborative tutoring system.

3.4.2 | Baseline model

A baseline model considering all fixed effects aside from linguistic revealed no significant effects on math scores. Table 3 displays the coefficients, standard errors, *t* values, and *p* values for each of the nonlinguistic fixed effects. Inspection of residuals suggested the model was not influenced by homoscedasticity. The nonlinguistic variables explained around 2% of the variance (R^2 = 0.016), whereas the fixed and random variables together explained around 55% of the variance (R^2 = 0.553). Thus, the majority of change found in the pretest and posttest was due to time.

TABLE 3 Baseline model for predicting math scores

Fixed effect	Coefficient	Std. error	*t*	*p*
(Intercept)	0.564	0.059	9.543	<0.001
Gender (male)	−0.039	0.061	−0.650	0.521
Grade (5)	−0.029	0.082	−0.350	0.729
Condition (procedural)	−0.024	0.060	−0.397	0.694
School	0.038	0.086	0.436	0.666

3.4.3 | Full model

A full model was developed including the nested baseline model and linguistic fixed effects. The model included five linguistic features related to cohesion (sentence linking connectives and adjacent overlap of adjectives), affect (respect terms), and lexical proficiency (number of function word types and verb hypernymy). None of the variables showed suppression effects. The model indicated that a greater number of sentence linking connectives (e.g., nonetheless, therefore, and however), function word types (e.g., prepositions, connectives, and articles), and overlap of adjectives predicted higher math scores.

Conversely, more respect terms and greater use of more specific words (i.e., greater hypernymy scores) related to lower math scores. Table 2 displays the coefficients, standard errors, *t* values, and *p* values for each of the fixed effects ordered by strength of *t* value. A log likelihood comparison revealed a significant difference between the baseline and full models, $\chi^2(2)$ = 42.486, p < 0.001, indicating that the inclusion of linguistic features contributed to a better model fit. Together, the fixed factors including the linguistic and nonlinguistic variables explained around 30% of the variance (R^2 = 0.303), whereas the fixed and random variables combined to explain around 82% of the variance (R^2 = 0.823).

The full model LME model demonstrated that a number of linguistic features were significant predictors of math performance (Table 4). Specifically, a greater number of sentence linking connectives and function words were predictive of math performance. These findings indicate that math performance is likely linked with the production of more complex syntactic structures such as those found in coordinated sentences and sentences with more structural components (i.e., function words). Lexically, math performance is associated with the production of more abstract words (i.e., words with greater hypernymy scores). This may be due to the fact that math solutions often require abstract thinking. In addition, a greater overlap of adjectives between sentences is a strong predictor of math performance, likely due to the repetition of math adjectives such as "greater than" and "less than." Lastly, our analysis demonstrated that math performance was related to the use of fewer words related to respect. This finding seems counter-intuitive, but performance within a math tutoring system that requires collaboration and timed completion of problems may favour curt and direct discourse between participants

TABLE 4 Full model for predicting math scores

Fixed effect	Coefficient	Std. error	*t*	*p*
(Intercept)	0.557	0.055	10.106	<0.001
Gender (male is contrast)	0.007	0.057	0.121	0.905
Grade (fifth grade is contrast)	−0.021	0.077	−0.284	0.778
Condition (procedural content is contrast)	−0.036	0.057	−0.639	0.527
School	0.032	0.080	0.401	0.691
Sentence linking connective	0.059	0.018	3.246	<0.001
Number of function word types	0.044	0.0193	2.273	<0.050
Respect words	−0.032	0.013	−2.518	<0.050
Adjacent overlap of adjectives	0.039	0.015	2.549	<0.050
Verb hypernymy	−0.038	0.017	−2.265	<0.050

that may be interpreted as less respectful. In total, the linguistic factors explained about 28% of the variance in the math performance data over and above the 2% explained by the nonlinguistic factors.

These linguistic analyses provide strong evidence that multimodal data streams focused around dialogue can provide unique insights that dramatically improve quantitative predictions of learning outcomes. Importantly, these findings also provided a greater understanding of how language features within student output can explain math performance indicating that language proficiency is likely linked to math proficiency. This link may be related to language skills that specifically help students discuss and analyse mathematical principles. The link may also reflect some general cognitive proficiencies that underlie both math and language skills. This proficiency may be related to analytic ability or an ability toward conceptual knowledge, both of which would assist in learning language and math. In all cases, the findings aid in our understanding of student learning by demonstrating links between different knowledge domains.

Our NLP analyses also provide a more holistic picture of learners that goes beyond what we can infer from log data. The majority of educational analytic research relies on features calculated from the log data recorded in intelligent tutoring systems and massive online open courses, such as video views, forum post reads, and assignment attempts (Baker & Inventado, 2014). Additional approaches to assessing student performance include the use of individual difference measures such as demographics, content knowledge, and literacy skills (DeBoer, Ho, Stump, & Breslow, 2014) although even these are rare in educational analytics. Here, we provide a different approach that goes beyond log data and provide the opportunity to examine individual differences in learners that provide a richer assessment of cognitive performance. By measuring language production features in learners, we can better assess individual differences that tap into cognitive production and likely increase the sensitivity of learner feedback algorithms by providing data approaches that go beyond log data.

4 | GENERAL DISCUSSION

We presented two empirical studies, collected in classroom studies with two distinct learning technology systems in different contexts (individual and collaborative). Our analyses and findings showcase a few different ways, in which multimodal data sources can enrich our understanding of student learning and provide a more holistic picture.

The two studies illustrated different types of multimodal data streams collected alongside automatically software-logged data. In Study 1, we collected student-focused screen and webcam video. This was useful for understanding students' learning processes and approaches based on detailed analyses of their interactions with the tutor interface, mouse movements, and out-of-tutor (in person) help-seeking. In Study 2, we collected high-fidelity audio of students' collaborative dialogue during their use of a fraction tutor designed to support collaboration between pairs of students. Because this classroom activity was dialogue heavy, and we collected audio using microphone headsets outfitted for each individual student, we were able to get high-quality transcriptions of students' dialogue and apply an NLP approach to make use of the large quantity of audio dialogue. The verbal data allowed us to identify linguistic features in students' collaborative dialogue that were highly predictive of math performance on pretest and posttest assessments, above and beyond any nonlinguistic variables.

Reflecting upon our results, we return to our initial questions of interest:

What types of learning phenomena can we capture and trace with computer-collected data, and what types do we miss? The video coding of student behaviours in Study 1 suggested that hint usage was a significant proxy into help-seeking behaviours that predict overall learning gains. Some of this information is available in log data, and one lesson learned is that we may gain insights by putting more effort into automatically quantifying hint usage behaviour. Additionally, we learned that log data misses all out-of-tutor interaction, and dialogue, and we found both to add significantly to the learning picture. Finally, log data also miss "mouse gestures" and other activity that is no longer present in the interface when students submit their answer. In some cases, this activity is very revealing of both common student misconceptions and tutor interface limitations. The linguistic analyses in Study 2 suggest that a lot of rich information is present in students' language during dialogue, and this linguistic information adds significantly to our understanding of students' cognitive and collaborative abilities. The studies also show how behaviours that may help learning in one context can hinder learning in another context. For instance, in Study 1, students seeking outside assistance from peers had negative effects on learning, whereas in Study 2, students collaborating had positive effects. These studies differed in numerous ways (student population, type of the tutor, and subject matter), and the differences reveal that enforcing productive collaboration (as in Study 2) may be more effective than spontaneous collaboration (as in Study 1), in which the assistance most often involved giving the other student the correct answer.

Are there ways we can enrich computer-collected data by collecting and analysing multimodal data streams, without a massive additional demand on researchers' time and effort? Both studies showcased the common multimodal learning analytics challenge of dealing with a large quantity of rich data with limited human time available. We addressed this challenge in different ways depending on what could be done with the available data, as well as what was useful given the learning context (individual vs. collaborative/interactive). For the chemistry study, we focused on moments of special interest and doing in-depth video coding of multimodal data for just these moments. We leveraged the information available in the log data to find these segments of video that seemed worthy of focus. For the collaborative fraction tutor, we applied automated linguistic analysis techniques to help discover linguistic features of student dialogue that predicted a substantial amount of unique variance in math performance.

There is no one-size-fits-all method for making use of the richness present in multimodal data while easing the burden on human time and effort. A careful analysis of the educational domain, the context (classroom vs. other and individual vs. interactive), and what the practical constraints are can help guide researchers to decide which multimodal streams to focus on collecting and how to make the best

use of the additional streams. Here, we have showcased some distinct methodological approaches that we found useful and enriching for learning occurring in different educational domains and in different contexts. Both studies and methodological approaches yielded different *kinds* of insights about student learning as well as quantitative model improvements uniquely beyond what was possible to infer from the log data. As a result, these new models of learning can be used to generate actionable knowledge for systems, students, teachers, and researchers.

ORCID

Ran Liu http://orcid.org/0000-0002-2832-5216
Jodi Davenport http://orcid.org/0000-0001-9091-9616

REFERENCES

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). New York, NY: Springer.

Blikstein, P. (2013). Multimodal learning analytics. In *Proceedings of the 3rd international conference on learning analytics and knowledge (LAK'13)* (pp. 102–106). New York, NY: ACM.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis: A general method for cognitive model evaluation and improvement. In *Proceedings of the 8th international conference on intelligent tutoring systems (ITS 2006)* (pp. 164–175). Jhongli, Taiwan.

Chi, M. T. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, *14*(2), 161–199.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods*, *49*(3), 803–821.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Davenport, J. L., Rafferty, A., Karabinos, M., & Yaron, D. (2015). *Investigating chemistry learning using virtual lab activities in real classrooms*. Paper presented at the 2015 Annual Meeting of the American Educational Research Association, Chicago, IL.

Davenport, J.L., Rafferty, A., Yaron, D, Karabinos, M., & Timms, M. (2014). ChemVLab+: Simulation-based lab activities to support chemistry learning. Paper presented at the 2014 Annual Meeting of the American Educational Research Association, Philadelphia, PA.

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher*, *43*(2), 74–84.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman, & P. Pufal (Eds.), *Constructivism in the computer age* (pp. 49–70). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Draney, K., Wilson, M., & Pirolli, P. (1996). Measuring learning in LISP: An application of the random coefficients multinomial logit model. In G. Engelhard, & M. Wilson (Eds.), *Objective measurement: Theory into Practice* (ed., Vol. 3) (p. 195). Norwood, NJ: Ablex.

Graesser, A. C., Conley, M., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 3. Applications to learning and teaching*. (pp. 451–473). Washington, DC: American Psychological Association.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371–398.

Kapur, M. (2010). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science*, *39*(4), 561–579.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Liu, R., Davenport, J., & Stamper, J. (2016). Beyond log files: Using multimodal data streams towards data-driven KC model improvement. *Proceedings of the 9th International Conference on Educational Data Mining (EDM'16)*. Raleigh, NC.

Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, *40*(4), 257–265.

Olsen, J. K., Aleven, V., & Rummel, N. (2015). Predicting Student Performance In a Collaborative Learning Environment. *Proceedings of the 8th International Conference on Educational Data Mining*.

Olsen, J. K., Belenky, D. M., Aleven, V., & Rummel, N. (2014). Using an intelligent tutoring system to support collaborative as well as individual learning. *Proceedings of the 12th international conference on intelligent tutoring systems* (ITS'14), 134–143.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Rau, M. A., Aleven, V., & Rummel, N. (2009). Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. *Proceedings of the artificial intelligence in education (AIED) conference*, 441–448.

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In *Proceedings of the international conference on intelligent tutoring systems* (pp. 174–184).

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 169–193). New York: Academic Press.

Team, R. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Worsley, M. (2014). Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 1–4). ACM.

How to cite this article: Liu R, Stamper J, Davenport J, et al. Learning linkages: Integrating data streams of multiple modalities and timescales. *J Comput Assist Learn*. 2019;35:99–109. https://doi.org/10.1111/jcal.12315