2016-06-01

Using Transaction-Level Data in Online Assessment

Robert Scott Nyland
Brigham Young University

Using Transaction-Level Data in Online Assessment

Robert Scott Nyland

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Randall S. Davies, Chair
Peter J. Rich
Ross Larsen
Charles R. Graham
Gove Allen

Department of Instructional Psychology & Technology

Brigham Young University

June 2016

ABSTRACT

Using Transaction-Level Data in Online Assessment

Robert Scott Nyland
Department of Instructional Psychology & Technology, BYU
Doctor of Philosophy

This article format dissertation explores the benefits of using detailed forms of assessment to enable feedback in educational contexts, and includes three separate, yet related articles. In the first article, I reviewed the current state of educational research in using online learning tools that collect detailed data regarding student learning. The article examined the type of data being collected, the way that these data are processed, and how the results are presented to instructors and students as feedback. In the second article, I describe a special case of these detailed forms of assessment in an Introduction to Microsoft Excel class, and look at the potential benefits of using transaction-level data to give feedback to instructors and students. This article provides empirical evidence for the difference between transaction-level data and final answer data in identifying student knowledge gaps and misconceptions. In the final article, I analyzed knowledge gaps and misconceptions identified in the Introduction to Microsoft Excel class by using additional student activity data (video watching and reading) to predict these knowledge gaps. This article serves as a case study for using data from integrated learning environments to provide feedback regarding student performance.

Keywords: data, feedback, performance based assessment, educational technology

ACKNOWLEDGMENTS

I would like to thank, first and foremost my wife, Stephanie, for supporting me during my studies. We will always look back on this time of our lives with fondness. I would like to thank Dr. Randy Davies for bringing me into his research world, while still allowing me to explore my own interests. Similarly, I would like to thank my other research group members, John Chapman and Gove Allen for their mentoring in many aspects of my project. I would also like to thank my other committee members for their feedback and direction, especially Dr. Ross Larsen, for shepherding me through the world of multivariate statistics. I will always remember to check my assumptions. In addition, I would like to thank Dr. Conan Albrecht for his many hours of support on Article 3 and Bob Bodily for his help double coding my literature review.

TABLE OF CONTENTS

LIST OF TABLES

Article 2

Article 3

LIST OF FIGURES

Article 1

Article 2

Article 3

DESCRIPTION OF RESEARCH AGENDA AND STRUCTURE OF THE DISSERTATION

This dissertation is primarily concerned with the issue of feedback, which Hattie (1999) has argued is one of the most important components in education. While feedback is relatively straightforward in face-to-face environments—the instructor observes students doing some kind of task and then the instructor gives them feedback regarding their performance—it is unclear how scalable this process is. How do we best give feedback in online environments, where it is often difficult to observe students engaged in task performance?

The answer may be in learning analytics. This field, which is described "as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Siemens, 2011, para. 2), attempts to deal with the increasing amount of data that exists as a byproduct of the educational process. These data come primarily from Student Information Systems (SISs) and Learning Management Systems (LMSs). However, while the data that comes from these systems may be used for the purpose of prediction, it often lacks the specificity that is needed for giving diagnostic feedback to students regarding how well they completed a task. As described by Thille et al. (2014), data from such systems has breadth (yielding large amounts of data about many learners) but not depth (large amounts of data about individual learners). They further argue that "data-enriched assessment in appropriately instrumented online learning environments can, for a large number of learners, provide insights into each individual learner's problem-solving processes, strategic learning choices, misconceptions, and other idiosyncratic aspects of performance" (p. 6). The goal of this dissertation is to explore the potential of data from online learning environments that are both broad and deep, in helping

students receive feedback regarding task performance. This dissertation will follow an article format. Three articles described below form the body of this dissertation.

Article 1: A Review of Data-Enabled Formative Assessment

The first article in the dissertation is a literature review of systems that have been built to collect data for the purpose of formative assessment. In this paper, we explored the types of data being collected from these learning environments, how the data is being processed, and how feedback on performance is presented back to the instructor or teacher. I originally submitted this paper to the AECT Young Scholar award where I was a finalist. The article is currently under review after a resubmission process to *Educational Technology Research & Development.*

Article 2: Transaction-Level Learning Analytics in Online Authentic Assessments

The second article presents result from the first phase of research with data from an online spreadsheet course being taught at several universities. In this article, we wanted to demonstrate the benefits of using transaction-level data (deep data) in helping identify student misconceptions and areas of struggle. In this paper, we took a single knowledge component (the use of absolute references) from an online Introduction to Microsoft Excel class and tracked student performance regarding that knowledge component over four occasions. Overall, we found that transaction-level data gives us a better sense of student misunderstanding than final answer assessment data alone. This paper is currently in the resubmission process in *The Journal of Computing in Higher Education.*

Article 3: Linking LMS Activity Data with Transaction-level Assessment Data

My final article is a case study in using data from an integrated learning environment. The online Introduction to Microsoft Excel course collects transaction-level data regarding student performance on authentic assessments (which was examined in Article 2) along with

specific usage data from the LMS (i.e., video watching and text reading). The goal of this article was to determine if there was a predictive relationship between video and text usage and a student's struggle regarding specific knowledge components in the course. We are still deciding where this final article will be submitted for publication.

Following all of the articles is dissertation conclusion along with references for materials cited in the introduction and conclusion. This references section does not include articles referenced inside of the three articles.

ARTICLE 1:

A Review of Data-enabled Formative Assessment

A Review of Data-enabled Formative Assessment

Rob Nyland (robnyland@gmail.com)

Brigham Young University

Corresponding Author

Rob Nyland

150 MCKB, IPT Department

Brigham Young University

Provo, UT 84602

Abstract

Feedback and formative assessment are critical processes that are often difficult to enact in online learning environments. Because of this, we need to identify options that collect student data from these environments to facilitate the process of formative assessment. The purpose of this literature review is to understand the current state of research on these types of tools. Namely, we were interested in identifying the types of data being collected by these tools, how these data were processed, and how the processed data were presented to the instructor or student for the purpose of formative assessment. We identified two categories of data: machine graded and activity stream data. The data were processed using three methods: activity streams, descriptive data analysis, and data mining. Processed data were presented to students through reports and real-time feedback, and to instructors through reports and visual dashboards. In conclusion, we make design recommendations for future systems looking to collect student data for the purpose of formative assessment and feedback.

Keywords: data, formative assessment, feedback, dashboards, data mining

A Review of Data-enabled Formative Assessment

Feedback is a critical component of successful teaching and learning. This is powerfully demonstrated in Hattie's (1999) meta-analysis of 196 educational studies in which he found that feedback had an average effect size of .79, nearly double the average effect size of all other interventions combined (.40). He concluded, "The simplest prescription for improving education must be 'dollops of feedback' —providing information [on] how and why the child understands and misunderstands, and what directions the student must take to improve" (Hattie, 1999, p. 11). In the classroom, feedback can be produced through formative assessment which "provide[s] feedback on performance to improve and accelerate learning" (Sadler, 1998, p. 77).

Good formative assessment relies on observation of the student learning process. This observation can be difficult as student work is increasingly done in online environments. While instructors might be able to read the body language or facial expressions of a student in a face-to-face environment to assess understanding, such signs of misunderstanding are not readily present from the data that comes from these environments. We need to find ways that we can harness the information that is collected by these online learning environments to enable formative assessment from an instructor.

This need for smarter formative assessment tools is echoed by the U.S. Department of Education (2011) who stated that online learning systems have the potential to "be used formatively to diagnose and modify the conditions of learning and instructional practices while at the same time determining what students have learned for grading and accountability purposes" (pg. xi).

While there has been much research in the last several years regarding the potential for finding patterns in educational data (see Papamitsiou & Economides, 2014), there has not been a

systematic review of research regarding data-enabled learning tools designed for formative assessment. In this review, we wanted to understand the current state of the literature regarding the use of technology-provided, data-enabled formative assessment. In particular, we wanted to understand the type of data being collected by these systems, how the data are processed, and how the processed data are presented to either the instructor or the students. By understanding the methods that are currently being employed for the purposes of technology-enabled formative assessment, we hope to identify the best way to enable feedback to students and instructors through these systems.

The literature review is guided by three questions:

1. What types of data are technology-enhanced formative assessment systems capturing?
2. What methods are being used to process the data?
3. Once the data are processed, how are the findings from the data being used for the purposes of feedback to either instructor or student?

Methods

The aim of this review was to identify how technological systems have been used for the purpose of formative assessment. Because many research fields—including learning analytics, educational data mining, artificial intelligence, and online learning—address topics in technology-enhanced assessment, we began with a broad set of search terms. We searched the terms *learning analytics, data mining, data analysis, assessment, formative evaluation, visualizations, dashboards, intelligent tutoring systems, computer-mediated communication*, and *data analysis* in the following electronic databases: ERIC, Education Full Text (H.W. Wilson), PsycInfo, Computers & Applied Sciences Complete, the ACM Digital Library, and Google

Scholar. Further searching was accomplished through backwards referencing of collected studies. The search was not bounded within any specific time period, but due to the types of technologies that we were looking for, articles tended to be from the last 20 years. Only peer-reviewed articles and conference presentations were included in the search.

Inclusion Criteria

After scanning through the initial set of articles, articles were removed according to the following inclusion criteria, matching our review questions:

- The system implemented in the research needed to have an explicit purpose of providing feedback to students or instructors based on data collected in the course.
- The system needed to capture student data digitally (e.g. through computers, mobile devices, or other sensor data) rather than through paper and pencil or Scantrons.
- The articles needed to report on implemented formative assessment systems, rather than theoretical designs.

As a result of our inclusion criteria, several revolutionary built for data analysis purposes, such as HIMATT, AKOVIA, ALA-Reader, and SNAPP were not included. While these tools are helpful in understanding student conceptual knowledge and community structure, they did not directly give formative feedback to students or instructors regarding content knowledge.

After excluding articles that did not meet the criteria, 24 total articles remained for review. Each article was reviewed according to each of the three review questions. Articles were coded for the type of data collected, the method used to process the data, and how the processed data were presented to the instructor or student for the purpose of formative assessment. After categories were developed, 20 percent of the articles were independently

examined by another reviewer using the developed codes. Initial agreement between the reviewers was at 80 percent, after which the coding definitions were further clarified and the two reviewers came to complete agreement on the articles in the sample. It was then deemed appropriate that the developed codes could be used on the remaining articles.

Findings

After reviewing the articles, several categories for each of the research questions emerged. In response to the first question, there were two categories of data: machine scored data, and activity stream data. For the second question, there were three data processing methods: activity streams, descriptive data analysis, and data mining. The third question had two sets of categories: methods for presenting data to students and methods for presenting data to instructors. In the student category, there were two presentation methods: reports and dashboards, and real-time feedback. In the instructor category, there were also two presentation methods: reports and visual dashboards. These were broken up into separate categories for the purpose of analysis. The reviewed articles along with their mapping in each of the categories is presented in Figure 1. Note that some articles addressed all aspects of the research questions (e.g., Chen & Chen, 2009), while other articles only addressed one aspect (e.g., Bajzek, Brown, Lovett, & Rule, 2007). In the remainder of this section, we will discuss those aspects of the reviewed studies derived from answering each of the research questions.

Figure 1. Reviewed articles categorized by data type, data processing and feedback presentation.

Data Types

While the types of data found in the review actually more closely resemble a continuum, we grouped the results into two categories (machine scored and activity stream data) for the sake of analysis.

Machine scored data. This data type was most readily represented by objectively scored assessment items. These items allowed assessments to be easily scored because they only allow a finite amount of possible answers to a given assessment item. An early published example of this comes from Buchanan (1998) whose PsyCal tool used objectively scored items as a formative assessment tool for students. Another example of this type of data can be seen in Cassady, Budenz-Anders, Pavlechko, and Mock (2001), who used Quiz Editor JS, an online assessment tool that easily creates objectively scored assessments that provide formative feedback. Additional examples of machine graded data in formative assessment can be seen in Henly (2003), Lin and Lai (2014), Wang (2008), and Alemán, Palmer-Brown, and Jayne (2011).

Activity stream data. As previously mentioned, the shift from machine scored to activity stream data is an artificial categorization placed on a continuum. Therefore, some studies using activity stream data are similar to those in the machine scored section. In these studies, objectively scored assessment data were collected along with student course activity data. For example, Heift (2005) developed a web-based tool for teaching German that collected objectively scored assessment data along with information regarding the context of how those answers were submitted (e.g. the student's ID, the time stamp, the task id, the system feedback, and student navigation patterns).

Similar data were collected from studies investigating the use of *intelligent tutoring systems* (ITSs). Feng, Heffernan, & Koedinger (2009) reported on the use of their ASSISTment

tutoring tool, which was designed to teach and assess mathematics to secondary education students. In their system, student data included the percentage of items that were answered correctly (machine scored data), number of items completed, the total time spent on the items, and the number of hint requests (in ITSs, students can typically ask for a hint when moving through a problem step). Similarly, Chen and Chen (2009) collected student interactions within a closed web-based learning environment. Logged data included correct response rate, reading rate of instructional materials, reading time, effort level of studying course materials, final test grade, attendance rate, accumulated score of question and answer and concentration degree.

This trend of collecting log data from student activities in a computer-based online environment continues in many other studies. May, George, and Prevot's (2011) Track Analysis and Visualization tool (TrAVis) logged time spent, connection frequency, message activity, and discussion threads started. Along with objectively scored assessment data, Ali, Hatala, Gašević, and Jovanović's (2012) LOCO-Analyst collected data regarding student visits on certain lessons and the estimated difficulty of those lessons. In addition, the tool collected social data, including the number of sent/received messages in the course forums and chat rooms. Similar student activity streams were collected in several of the other studies reviewed (Kosba, Dimitrova, & Boyle, 2007; Mazza & Dimitrova, 2007; Merceron & Yacef, 2005; Nedungadi & Raman, 2012; Scheuer & Zinn, 2007).

McNely, Gestwicki, Hill, Parli-Horne, and Johnson (2012) took an unconventional approach to capturing activity stream data outside of a learning environment. They created Uatu, a small program that collects collaboration information from a Google Docs document. Once Uatu is added to a created document, it logs all of the revisions that are made to that document. Their hope was that logging the revision activity would act as a metacognitive tool for its users.

While several of the previously mentioned studies collected data regarding social activity in the class, the data collected was high level (frequency counts, who the message was sent to). Hsu and Ho (2012) took a further step into gathering activity stream data by collecting text content from students' discussion board posts. Their goal was to find ways of combing through text data to automate the feedback process. Leeman-Munk, Wiebe, and Lester (2014) also sought to automate the process of giving feedback to text inputs, but within an assessment environment. They collected short text answers from Leonardo; a virtual environment used to teach science to upper elementary students, and used their tool Write Eval to analyze the text responses.

Real-time activity streams. The most unstructured data encountered in our review could be classified as real-time activity streams. These were moment-by-moment descriptions of student activities, often presented in real-time. The most basic example of this is seen in Shirley and Irving's (2015) investigation into the use of Connected Classroom Technology (CCTs) or audience response systems. In their qualitative investigation of the use of these tools, the authors described four middle and high school science classrooms that use TI-Navigator™, a system that collects and displays the real-time activity stream of students. This system displays exactly what a student has typed into their own calculator, allowing the students to compare their responses with fellow students.

Other studies took on more complex data from student activity streams. Koh, Basawapatna, Nickerson, and Repenning (2014) collected the activity streams of middle school students who were engaged in the process of designing electronic games using the REACT system (Real Time Evaluation and Assessment of Cognitive Thinking). While not specific about what data were collected from the student process, the authors mention that the system "breaks down all collectable student project information and records it in the REACT database" (p. 51).

Therefore, it appears that REACT is collecting a real-time activity stream of a student's game development process.

Finally, the most complex activity stream data were collected by another pair of studies (Kennedy, Ioannou, Zhou, Bailey, & O'Leary, 2013; Sewell et al., 2008). Here the researchers collected activity streams from students working with a 3D dental simulation tool. In the study, subjects manipulated two pen-like haptic tools (that simulate a drill and an irrigator) in a 3D virtual space. While the students are working, the tool simultaneously recorded 48 metrics generated in real-time. Notable metrics included timestamp, tool position and orientation, the size and shape of the tool, and information about the anatomical structure they are working with. This detailed data could then be played back in real-time to give an account of the student's procedure.

Thus far, we have looked at the spectrum of data that is being collected by these systems for the purpose of formative assessment. Data in its most structured form was machine scored assessment data. As the data becomes less structured, we saw systems that collected activity streams of students, sometimes in real-time.

Now that we have looked at the types of data being collected by these tools, we will report how this collected data were processed for the purposes of formative assessment.

Data Processing Methods

In our next question, we wanted to understand the methods that were used to process the data in order to give feedback. Here processing refers to means by which data is retrieved from their database. Three main data processing methods categories emerged from our review: First, we had data that was unprocessed (activity stream data). Second, we had data that was produced

from descriptive data analysis. Lastly, we have data that was processed through the use of data mining techniques.

Activity streams. Data that is presented as activity streams are completely unprocessed. In our review, we found only one example of this—Shirley and Irving's (2015) research into the use of Connected Classroom Tools. In their study, the TI-Navigator™ system would directly display the problem solving processes of individual students on the main screen for everyone in the class to see.

Descriptive data analysis. The next set of results from types of data processing methods consists of descriptive student data compiled from simple database queries—in the form of counts, sums, or averages. This is demonstrated in its most basic form by Cassady et al., (2001). In their study, students took short formative quizzes that were authored in Quiz Editor JS. Data from these quizzes were processed through simple descriptive statistics—counts of the number of questions that were right or wrong. Similarly, Heift's (2005) German language tutor presented objectively scored assessment as well as tracking data with student users through queries. Performance data was in the form of percentages correct, as well as counts of different types of errors (spelling, verb inflection, and word order). The use of these types of simple queries to process and present data is common in many instructional systems that keep a log of student assessments and other activities (Buchanan, 1998; Feng et al., 2009; Henly, 2003; May et al., 2011; McNely et al., 2012; Wang, 2008).

Data mining. When data collected is as an activity stream, many researchers turn to data mining for processing. Data mining is used when answers to questions cannot be found simply by descriptive data analysis alone, but rather hidden patterns in the data need to be discovered

(Dunham, 2003). Data mining has risen within the past few years as a topic of increasing interest to the educational community (Baker, 2010).

In the research we reviewed, we found several different methods of data mining to process activity stream data collected in a technology-enhanced learning environment. While the methods are diverse, we will group them into three categories based on the goal of the method, including: Building a model of student performance, creating an expert model for automated feedback, and making recommendations for remediation.

Building a model of student performance. The goal of data processing methods in this category is to help the instructor or student understand the student's current level of knowledge in a particular domain. While a specific method wasn't stated, Koh et al. (2014) used data mining techniques to parse through the activity streams of students participating in game programming. Their goal was to understand the level of performance on a game design task.

Merceron and Yacef's (2005) TADA-Ed tool was developed with the explicit purpose of helping instructors model their own students using data mining. With it, instructors could pre-process data and then use several data mining techniques—k-means, hierarchic clustering, and decision trees. K-means and hierarchic clustering are both ways of grouping students or student responses together using some criteria, while decision trees are a way to classify an object based on other similar objects.

In another tool for teachers, Kosba et al. (2007) used a fuzzy student modeling approach to generate a student's certainty factor for every concept that they covered in a learning environment. Based on these models, a concept is assigned as completely learned, learned, or not learned for a given student or student group. In addition, the system builds a model of groups within the class based on inputted criteria, e.g. nationality, background, or course

preferences. These features allow instructors to track the progress of individual students in the class as well as selected groups.

Alemán et al. (2011) processed objectively scored data using Snap-Drift Neural Networks, a high-speed data categorization algorithm, to group similar responses together. The resulting groups consist of students making the same mistakes on the same problems. The authors felt that these groups represented students who had similar misconceptions that an instructor might be able to correct through remediation. Similarly, Lin and Lai (2014) used an a priori algorithm to identify questions on an objectively scored assessment that were commonly missed together.

To model student knowledge to the students themselves, J. L. Hsu, Chou, and Chang (2011) used text mining methods to provide students with feedback regarding the quality of their answers in an online discussion board. Using Latent Semantic Analysis and multiclass singular value decomposition, their tool automatically classified student responses to the discussion board according to its depth of understanding as described by Bloom's taxonomy.

Creating an expert model for automated feedback. The next set of studies used data mining to build models of expert performance, which could then be used to generate automatic feedback to students. To create automated feedback to constructed text response data in an online science learning environment, Leeman-Munk et al. (2014) used machine learning techniques. Machine learning is concerned with "design[ing] systems that can learn from data" (Bell, 2014, p. 2). They trained their system on sample human-graded correct and partially correct answers and then used soft cardinality and Latent Semantic Analysis to automatically grade submitted answers that were not completely identical to the sample answers Kennedy et al. (2013) used several algorithms to give students expert feedback in a 3D dental simulator. First,

they identified an association rule that could be calculated in real-time that would demonstrate surgical expertise. Next, they took this association rule and used it to build a Hidden Markov model (taking an observable metric and then inferring a hidden gesture not easily observable through those metrics alone). In this case, the Hidden Markov model would infer whether a participant was engaging in a stabbing motion, a sweeping motion, or a stabbing-sweeping motion with their dental instrument. The researchers then trained these models on the performance of expert and novice surgeons to determine when the tool should give feedback. Many of the same data analysis techniques were used in a similar study by Sewell et al., (2008).

Making recommendations for remediation. In the last set of data mining methods, the goal was to provide recommendations for the future learning for students. Scheuer and Zinn's (2007) Student Inspector used machine learning "to predict future learning outcomes" (p. 6). Instructors could use *the Analyser* (a module within Student Inspector) to suggest appropriate lessons to students based on their performance history. Poorly performing students could have easier tasks recommended to them, while higher performing students were recommended more challenging tasks.

Chen and Chen (2009) used several sequential data mining techniques to make learning recommendations in an online learning environment. First, they performed factor dependence analysis using a fuzzy clustering method to determine student measures that were most essential to the analysis. The final measures included reading rate, correct response rate of test items, accumulated score on the discussion board, and effort level of studying course materials. Second, fuzzy association rule mining was used to create rules that would predict student performance. One example of a rule is "CR_L => GRADE_L", interpreted as *if a student has a*

low correct response rate on a test, they are likely to have a low cumulative grade. The system could then use these rules to provide formative feedback to students in the course.

The studies above demonstrate that there are a variety of data analysis techniques currently being employed to process both machine gradable and activity stream data from online learning environments. In the next section, we will discuss how this information, once retrieved, is being presented to the instructor or student for the purposes of formative assessment.

Data Presentation for Formative Assessment

Once processed, data from learning systems can be presented to both students and instructors for the purpose of providing feedback on the learning process. In this section, we will first talk about how data were presented to students for the purpose of formative assessment, and then we will discuss how data were presented to instructors.

Students. After reviewing the literature, we discovered two main methods of delivering feedback to students based on data collected in the learning process. The first is through the use of student reports and dashboards and the second is through real-time feedback delivered to the student while engaged in the learning activity.

Student reports and dashboards. The first category of feedback delivery to students is in the form of reports and dashboards. These are both visual displays of student activities that can be viewed directly by the student—the report is in a tabular format, while the dashboard includes graphic representations of the data. In our review, the first report was shown in Buchanan (1998), whose PsyCal systems was designed to be a formative assessment tool for an introduction to psychology course. Once a student submitted their answers to an assessment, they were presented with a report that told them how many questions they answered correctly,

along with a list of questions that were incorrectly answered. Correct answers were not given, but rather the student was referred to sections of the text where correct answers could be found.

After students completed learning activities in Heift's (2005) German language tutoring system, they viewed their progress via The Report Manager. It also kept detailed error reports, identifying which items on activities or assessments were missed. For those exercises in which the student wished to achieve a better score, the Report Manager allowed them to redo exercises. In an evaluation of the effectiveness of the Report Manager, Heift (2005) found that of users who viewed a learning report, 70% of them repeated whole exercise sets after viewing the results – suggesting that viewing student performance levels could be a trigger for seeking better performance in a class. A similarly styled learning report was also used by Nedungadi and Raman (2012) to give feedback to students after engaging in a mobile learning session, and Lin and Lai (2014) after answering formative questions in their annotation-sharing and intelligent formative assessment (ASIFA) system. In examining this set of reports, it would seem that the most useful are those that encourage some kind of action on the part of the student. This may include revisited material not understood, or working on a problem set that was not previously mastered.

McNely et al. (2012) were particularly interested in the metacognitive value of a report for students when they developed Uatu—a tool which tracks contributions to a Google document. Their hope was that by seeing a visual display of how group members had contributed to the document, they would be more likely to increase their own participation. However, the authors did not perform any evaluation on the tool, and so it is difficult to know whether the tool had the intended impact.

Hsu et al. (2011) wanted to provide students with a more visual display of their performance of a learning environment. After using text-mining algorithms to classify student discussion responses by their level of understanding on Bloom's cognitive taxonomy, the system produced two graphs—one for individual cognition level and one for the cognition level of the class. The individual cognition displayed a spider (or radar) chart with the student's scores for each of the six cognition levels on a scale from 0 to 1. If a student had a lower score for a certain cognitive level, then they had failed to use words that correctly corresponded with that level. In the class level graph, the student could see common words used by other students in their class and how those words were assessed in terms of cognitive understanding. This graph gave feedback on the types of words that the student should use to increase their cognitive understanding of the underlying concept. While Hsu et al.'s (2011) system is conceptually interesting, it doesn't seem to directly address some large assumptions, namely the validity of using words and phrases to assess underlying cognitive understanding. We feel that using phrases that match a pre-determined level of understanding is not valid evidence that a student is actually thinking at that level.

In a final study where feedback was directed specifically to students, Chen and Chen (2009) used two reports to help learners see their progress in an online learning module. In the first report, students were presented with several metrics gathered from the learning system and the instructor. These included attendance rate, concentration rate, instructor comments, and correct response rate. In addition, the student's final score was predicted based on the associated fuzzy rules created in the earlier data mining process. In the second report, the student was able to see the previously generated fuzzy learning rules for the class. An example of this is *High RT (Reading Time) => High Score*—meaning high reading time leads to a higher score. While such

rules might be interesting for an instructor to look at, they seem inappropriate a group of 9 to 11 year old students (the sample in the study) to understand. In addition, the authors do not address whether there might be possible detrimental effects to student motivation after they are shown a predicted score.

Real-time feedback to students. In the next group of studies, feedback was delivered to students in real-time while engaged in the learning activity. This approach is seen in its most basic form by Cassady et al. (2001) in their work with embedded formative quizzes. While the authors did not delve deeply into the type of feedback given to students, they did mention that the feedback was immediate. Because the data were objectively scored, it is likely that the feedback was pre-programmed by the assessment designer. While these systems can give feedback quickly, the type of feedback is not customized. The system knows whether a student got a question wrong, but it does not know the reason. The use of quick feedback following objectively scored assessment was also utilized by Henly (2003) and Wang (2008).

Another relatively simple mechanism for real-time student feedback was examined by Shirley and Irving (2015). The TI-Navigator™ tool in their study provided a real-time activity stream of student responses viewable to members of the class. Notwithstanding, in this case, the students had to derive feedback themselves from the activity stream by comparing their answers to other students in the class as the system did not process the student data.

Kennedy et al., (2013) demonstrated the most sophisticated real-time student feedback system. In their research with a 3D dental simulator, feedback was presented to the student in the form of real-time suggestions for better performance in a text box on their display. In the example provided in their article, the system detected when a student was not performing the procedure in the optimal manner. At this point a small text box appeared on the student view

saying, “You are too tentative at this stage of the procedure. Apply more force” (p. 179). While their research is promising, the authors note that their real-time feedback process is in its initial stages. It is clear that more research needs to be done to understand how often such feedback could be presented before the student before they feel like they are being pestered.

Instructors. Our review essentially found one way of presenting feedback to an instructor: the report. However, these reports vary in their level of detail and their reliance on visual representations. In this section, we will briefly discuss more traditional reports and then discuss the use of visual reports (commonly referred to as dashboards).

Instructor Reports. Instructor reports are tables that present data to an instructor about student performance in an online environment. In our review, reports appeared in their most simple form in a study by Feng et al. (2009), whose ASSISTment system allows instructor to run several reports, including a grade book. The grade book delivers several data points from the ASSISTment system back to the instructor in tabular form. The instructor could then look at possible patterns and intervene if necessary. A similar report is used by Chen and Chen (2009). In their study, processed student and instructor data were delivered via a report to the instructor’s mobile device. This report includes learner test scores, the variances in learner scores, and the learning rules that were derived from the earlier data mining process. While the information that is reported from this process may be useful, it would seem difficult for a typical instructor to derive meaning and take an action from such a table of numbers. This was also a problem in Aleman, Palmer-Brown, and Jayne's (2011) research which processed a series of formative, multiple-choice questions using data mining. The result of this data processing was then presented to the instructor as groups of responses to a series of questions. The authors explained:

> For example, "b/d c *" represents a group characterized by all the students answering b or d to question 1, c to question 2, and mixed answers to question 3. Hence the educator can easily see the common mistakes in the groups of the student answers highlighted by the tool (p. 503).

Once again, while the information derived from the data processing may have been useful, it still required the instructor to sufficiently understand the content of each of the questions to derive meaning from the data.

The Teach Advisor (TAdv) tool described by Kosba et al. (2007) attempted to simplify the process of turning instructor reports into actionable feedback. In the *Generate advice* and *View advice* sections of the TAdv tool, the system created advice for the instructor based on the algorithm used in the earlier data processing step. This feedback could be addressed to individual students, groups of students within the class, or the class as a whole. The instructor could review the advice, and use the report to send automatically created feedback to students. Such a move seems like step in the right direction, as the system is helping direct the instructor's attention to potential learning problems in the course.

Visual dashboards. While the instructor reports in the previous section typically relied on data delivered to the instructor via a table, reports in the instructor dashboard category rely on visual representations of data. Bajzek et al. (2007) provide a good definition of a dashboard when talking about their own Digital Dashboard for Learning (DDL) for Carnegie Mellon University's Open Learning Initiative (OLI). According to them, a dashboard "provides visibility into key indicators of student learning through simple graphics such as gauges, charts, and tables within a web browser" (p. 2). The authors then gave several key components to an effective educational dashboard:

- Provide a wide variety of different metrics in a single consolidated view
- Roll-up details into higher level summaries
- Provide intuitive visualizations that are instantly understandable – for example, red bars mean a problem
- Provide linkages to the data that they represent. (p. 3)

In one of the earliest studies using visual displays for the purpose of formative assessment, Merceron and Yacef (2005) built a tool that allowed for the cleaning, processing and visualization of educational data (TADA-Ed). While several visualizations were available for the instructor to look at, the visualizations could only be accessed once an advanced statistical analysis was run (such as a cluster analysis). While there is definitely power in such an approach, it is difficult to determine whether such complex visualizations would be helpful to the typical instructor monitoring of the learning of their class.

Mazza and Dimitrova (2007) used several visual representations in their CourseVis tool to help instructors understand how university students were progressing in an online course. In one of the visualizations, discussion forum activity was visualized with a *discussion plot*. In this plot, the discussion posts originator, time of posting, and post follow-up activities are plotted on a two dimensional axis, with point size representing the number of follow-up posts. This *discussion plot* can help the instructor identify students that are making active contributions to the discussion forums. Another visualization discussed is what the authors refer to as a *cognitive matrix*; in this visualization, student performance for each of the course topics in the class was visualized in a grid, with colors representing performance. This matrix enables the instructor to quickly ascertain the class' knowledge on a given topic, or a single individual's performance

across a range of topics. The last described visualization is the *student access plots* that visualize student activity in the course over time. In this set of graphics, overall logins to the course along with activity in certain content pages were graphed. The use of these *student access* plots may help an instructor gauge a student's engagement with the course.

In Scheuer and Zinn's (2007) Student Inspector, course activity data were visualized through a combination of windows that allow instructors to sort, filter, and visualize data. In the *performance measurement* tab of the program, the instructor could identify the performance of individual students compared with their peers. In the *misconceptions* tab, instructors were able to look at those topics for which students struggled the most with, the distribution of which was visualized as a pie chart. The *misconceptions* tab allowed the instructor to quickly see where they may need to remediate for individual students, or adjust instruction for the entire class. In the last visualization, the *topic coverage* tab, the instructor was able to see weak and strong performance areas for each student. Such a view also may be helpful to an instructor in identifying areas for student remediation. Overall, the Student Inspector tool seems to be useful as a tool for data enabled formative assessment. It focuses on providing the instructor with actionable information in the form of areas for student remediation. In fact, after Scheuer and Zinn's (2007) description of their product, they also provide some evaluation data. Overall, instructor reaction was positive and they felt that the most useful aspect of the program was the identification of student misconceptions.

A windowed interface similar to Student Inspector was used by Ali et al. (2012) in their LOCO-Analyst tool. In the latest version of the tool, student information is presented to the instructor in four tabs: forums, chats, learning, and annotations. The forums and chats tabs provided visualization to the instructor regarding the students' online interactions. The number

of postings for each topic are presented via tabular and bar chart format. In the learning tab, the instructor is presented with bar charts regarding student activity time on topics in the course. The researchers also performed an evaluation on the tool to get feedback from potential users. Overall, response to the tool was positive; however, most of the instructors (70%) felt that the tool did "not provide enough information on how to improve student's online interactions" (p. 482). This suggests that there is something about the type of data being presented to the instructors that is not actionable.

May et al.'s (2011) Tracking Data Analysis and Visualization (TrAVis) tool used two visualization tools to give information to instructors. The first, the *Time Machine,* allows instructors to see a timeline view of completed student activities. The second tool, a radar graph, "provid[es] simultaneous observation and analysis of different aspects of user activity" (p. 61). In one example, the radar graph visualizes connection frequency, discussion threads started, messages posted, messages replied, and messages quoted. These radar tools are also applied to group discussion, where an instructor can quickly see online group activity at a glance. While these visualization tools may be good for instructor awareness of course activity, they also offer little by way of actionable information to an instructor.

A radar graph was the primary visualization in Koh et al.'s (2014) Real Time Evaluation and Assessment of Computational Thinking (REACT) tool. In their Computational Thinking Pattern Analysis Graph, student activities were mapped onto nine different *computational thinking patterns* (cursor control, generations, absorptions, collision, transportation, push, pull, diffusion, and hill climbing). Individual student graphs are then aggregated on the Assessment Dashboard and given color codes to indicate how well a student is progressing through their programming process. Green graphs indicate that the student is working correctly, while orange

and red graphs suggest to the instructor that a student may be struggling. In testing the tool with a sample of instructors, the response was positive, and the instructors felt that the visualizations helped them monitor the work that their students were doing.

Discussion

The aim of this research was to review current practices in data-enabled formative assessment. In doing so, we wanted to understand the types of data being collected, the way that this data were being processed, and how the processed data were presented before students or instructors for the purpose of formative assessment.

While a few systems were collecting only objectively scored data (i.e. multiple choice and true-false questions), a majority of the systems were collecting additional data about the users of these systems. While objectively scored data were convenient for giving pre-programmed feedback to the student (as seen in Cassady et al., 2001), it cannot give targeted feedback to the student or make inferences about why the student responded incorrectly to the question. To give more nuanced feedback to students, additional information needs to be collected. This was demonstrated by the many studies that collected data from student activities. By collecting information about assessments, social activity, and course usage, the systems had more information to make recommendations or suggestions for remediation to the instructor or the student. Additionally, by keeping a log of a student's progress throughout a course, the systems were able to see larger trends for an individual student, and identify to instructors that students may be demonstrating that they are struggling with material.

When it comes to processing data, many of the systems that we reviewed used descriptive data analysis to collect information and present it to the instructor. These systems query student logs in the database to create reports for instructors and students. While such queries are useful,

they may not be uncovering information that would normally be hidden to an instructor or student. To uncover these more nuanced patterns, other systems used data mining techniques. Overall, there was not one clear data mining method being used, it largely depended upon the nature of the data and the end goal. A few of the products that we reviewed (especially Merceron & Yacef, 2005) relied on the instructor to pick the appropriate data mining method. While data mining has proven to be an effective way of finding patterns in student data, we cannot expect instructors to be experts in its method. Because of this, most of the actual processes should be hidden to the instructor (with possible advanced option for instructors who want additional control). As data mining processes become more common in educational research, more effort needs to be given to explain these methods to non-experts so that appropriate research questions can be formulated.

Once this educational data were processed, we saw several ways that it was presented to both students and instructors. For the students, several studies found value in using reports and dashboards for meta-cognitive purposes (Heift, 2005; McNely et al., 2012; Nedungadi & Raman, 2012). However, these approaches gave students high-level feedback and not specific areas for remediation. A review of Chen and Chen (2009) also pointed to the need to make recommendations appropriate to the age level and ability of the students—presenting the model directly to the student does not seem like the best approach. Also unknown is the motivational effect of presenting such models directly to students. More research needs to be focused on understanding how best to give feedback to students in these environments. It will also be important for research to look more into appropriateness and practicality of real-time feedback to students as is demonstrated by Kennedy et al., (2013). Such feedback may be more appropriate for simulations (as in surgical training), but it would be interesting to see it applied to other

contexts. For many of these issues, we may want to look to intelligent tutoring literature to see what it has uncovered about feedback in learning environments.

We also discovered many ways that data from these learning environments is being presented back to instructors for the purpose of formative assessment. The first category in these types of feedback was instructor reports. While such reports are likely useful to instructors, they may make it difficult for instructors for identify important patterns in student learning that need to be corrected. Instructor dashboards were more helpful in accomplishing this goal. Much like data mining techniques, the designs of these dashboards varied from project to project, depending upon the type of data that is being collected and the specific needs of the instructor. We should place more effort on developing customizable instructor dashboards that present a variety of choices for displaying student data.

It should be noted that the most salient feature of the dashboard identified in Scheuer and Zinn (2007) was the ability to identify student misconceptions. This points to the question of actionability of the data in these visual dashboards for instructors. More research needs to be focused on identifying the utility of dashboard features for instructors. What data are important for an instructor in a dashboard? What presented information are they already aware of? What is going to help them make pedagogical decisions?

Implications for Future Research

Through this review, we have identified several areas where further research could help us develop smarter tools to enable formative assessment. First, as more and more educators are looking to use data processing and mining techniques to find patterns in student data, we need more accessible materials to help them understand these processes. We hope that in the future, researchers will provide better rationales for their use of data analysis techniques depending on

the type of educational data at hand. Currently, much of the research in the fields of Learning Analytics and Educational Data Mining is being created by computer scientists. Making these methods more accessible to education researchers and practitioners could allow them to contribute more to the direction of the field.

Our review also suggests that we need to look more at how students and instructors are using reports and dashboards. Our critique of Chen and Chen's (2009) research points out the sometimes inappropriate nature of dashboard feedback to students. Other tools that we have identified in our study (Merceron & Yacef, 2005) seemed too complex for a normal instructor to interpret. More empirical research and evaluation on the use of these tools could help designers create visualizations that are appropriate to the development level of the user and also helpful in producing specific recommendations.

A final area of research arises from a finding of Scheuer and Zinn (2007). In the evaluation of their tool, the ability to identify learner misconceptions was seen as the most helpful feature to instructors. This finding does not surprise us, as we feel that instructors want to use dashboards as tools to understand areas of their class that they can improve—the identification of misconceptions gives a concrete starting point for a remediation in their class. We feel that more research is needed regarding how we can use these data-enabled tools to identify misconceptions among students, especially misconceptions that might not be present in larger grained data.

Conclusion

The goal of this review was to review tools that have been used to facilitate formative assessment using computer-collected data. From the amount of research that we have found, it appears that many in the fields are making great strides in harnessing the power of these tools to

help give instructors information about student performance. The most effective of these systems collect as much student performance data as possible, parse through the data using advanced analysis techniques, and then present patterns and trends back to the instructor or teacher using visual techniques.

We hope that this review will act as a starting point for future designs of online learning and assessment tools. Data can be a powerful tool, which when applied appropriately can give instructors great insight into the learning process. These insights will allow them to do what they are best at—teaching.

References

Alemán, J. L. F., Palmer-Brown, D., & Jayne, C. (2011). Effects of response-driven feedback in computer science learning. *IEEE Transactions on Education*, *54*(3), 501–508. doi:10.1109/TE.2010.2087761

Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A qualitative evaluation of evolution of a learning analytics tool. *Computers & Education*, *58*(1), 470–489. doi:10.1016/j.compedu.2011.08.030

Bajzek, D., Brown, W., Lovett, M., & Rule, G. (2007). Inventing the digital dashboard for learning. Proceedings from *World Conference on Educational Multimedia, Hypermedia and Telecommunications 2007*, 1084–1092. Retrieved from http://www.editlib.org/p/25512

Baker, R. S. J. D. (2010). Data mining for education. In *International Encyclopedia of Education* (Vol. 7, pp. 112–118). doi:10.4018/978-1-59140-557-3

Bell, J. (2014). *Machine learning: Hands-on for developers and technical professionals*. Indianapolis, IN: John Wiley & Sons.

Buchanan, T. (1998). Using the World Wide Web for formative assessment. *Journal of Educational Technology Systems*, *27*(1), 71–79. doi:10.1016/s0360-1315(00)00049-x

Cassady, J. C., Budenz-Anders, J., Pavlechko, G., & Mock, W. (2001). The effects of internet-based formative and summative assessment on test anxiety, perceptions of threat, and achievement. In *Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001)* (p. 13).

Chen, C. M., & Chen, M. C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education*, *52*(1), 256–273. Retrieved from https://www.lib.byu.edu/cgi-

bin/remoteauth.pl?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2008-17099-026&site=ehost-live&scope=site

Dunham, M. H. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, New Jersey: Prentice Hall.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243–266. Retrieved from https://www.lib.byu.edu/cgi-bin/remoteauth.pl?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2009-10369-003&site=ehost-live&scope=site

Hattie, J. (1999). Influences on student learning [pdf]. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.8465&rep=rep1&type=pdf

Heift, T. (2005). Inspectable learner reports for web-based language learning. *ReCALL*, *17*(1), 32–46. doi:10.1017/S0958344005000418

Henly, D. C. (2003). Use of Web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education, 7*(3), 116–122. doi:10.1034/j.1600-0579.2003.00310.x

Hsu, C. C., & Ho, C. C. (2012). The design and implementation of a competency-based intelligent mobile learning system. *Expert Systems with Applications*, *39*(9), 8030–8043. doi:10.1016/j.eswa.2012.01.130

Hsu, J. L., Chou, H. W., & Chang, H. H. (2011). EduMiner: Using text mining for automatic formative assessment. *Expert Systems with Applications*, *38*(4), 3431–3439. doi:10.1016/j.eswa.2010.08.129

Kennedy, G., Ioannou, I., Zhou, Y., Bailey, J., & O'Leary, S. (2013). Mining interactions in

immersive learning environments for real-time student feedback. *Australasian Journal of Educational Technology*, *29*(2), 172–183.

Koh, K. H., Basawapatna, A., Nickerson, H., & Repenning, A. (2014). Real time assessment of computational thinking. Proceedings of *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 49–52). IEEE. Retrieved from http://sgd.cs.colorado.edu/wiki/images/9/91/Paper_24.pdf. doi: 10.1109/VLHCC.2014.6883021

Kosba, E., Dimitrova, V., & Boyle, R. (2007). Adaptive feedback generation to support teachers in web-based distance education. *User Modelling and User-Adapted Interaction*, *17*, 379–413. doi:10.1007/s11257-007-9031-z

Leeman-Munk, S. P., Wiebe, E. N., & Lester, J. C. (2014). Assessing elementary students' science competency with text analytics. In Proceeding of the *Fourth International Conference on Learning Analytics and Knowledge* (pp. 143–147). ACM. doi:10.1145/2567574.2567620

Lin, J. W., & Lai, Y. C. (2014). Using collaborative annotating and data mining on formative assessments to enhance learning efficiency. *Computer Applications in Engineering Education*, *22*(2), 364–374. doi:10.1002/cae.20561

May, M., George, S., & Prévôt, P. (2011). TrAVis to enhance online tutoring and learning activities: Real-time visualization of students tracking data. *Interactive Technology and Smart Education*, *8*, 52–69. doi:10.1108/17415651111125513

Mazza, R., & Dimitrova, V. (2007). CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human Computer Studies*, *65*, 125–139. doi:10.1016/j.ijhcs.2006.08.008

McNely, B. J., Gestwicki, P., Hill, J. H., Parli-Horne, P., & Johnson, E. (2012). Learning analytics for collaborative writing: A prototype and case study. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 222–225. doi:10.1145/2330601.2330654

Merceron, A., & Yacef, K. (2005). TADA--Ed for Educational Data Mining. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *7*(1), 267–287. Retrieved from http://imej.wfu.edu/articles/2005/1/03/printver.asp

Nedungadi, P., & Raman, R. (2012). A new approach to personalization: Integrating e-learning and m-learning. *Educational Technology Research and Development*, *60*, 659–678. doi:10.1007/s11423-012-9250-9

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology and Society*, *17*(4), 49–64.

Sadler, D. R. (1998). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 77–84. doi:10.1080/0969595980050104

Scheuer, O., & Zinn, C. (2007). How did the e-learning session go? The Student Inspector. Proceedings of the *Conference on Artificial Intelligence in Education (AIED'07)*, 487–494. Retrieved from http://dl.acm.org/citation.cfm?id=1563601.1563678

Sewell, C., Morris, D., Blevins, N. H., Dutta, S., Agrawal, S., Barbagli, F., & Salisbury, K. (2008). Providing metrics and performance feedback in a surgical simulator. *Computer Aided Surgery*, *13*(2), 63–81. doi:10.1207/s15327752jpa8502

Shirley, M. L., & Irving, K. E. (2015). Connected classroom technology facilitates multiple components of formative assessment practice. *Journal of Science Education and*

Technology, (24), 56–68. doi:10.1007/s10956-014-9520-x

US Department of Education. (2011). *Transforming American Education: Learning powered by technology* (Vol. 8). doi:10.2304/elea.2011.8.2.102

Wang, T.-H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education*, *51*(3), 1247–1263. doi:10.1016/j.compedu.2007.11.011

ARTICLE 2:

Transaction-Level Learning Analytics in Online Authentic Assessments

Transaction-Level Learning Analytics in Online Authentic Assessments

Rob Nyland

Randy Davies, PhD.

John Chapman

Gove Allen, PhD

Brigham Young University

Corresponding Author

Rob Nyland

150 MCKB, IPT Department

Brigham Young University

Provo, UT 84602

robnyland@gmail.com

Abstract

This paper presents a case for the use of transaction-level data when analyzing automated online assessment results to identify knowledge gaps and misconceptions for individual students. Transaction-level data, which records all of the steps a student uses to complete an assessment item, are preferred over traditional assessment formats that submit only the final answer, as the system can detect persistent misconceptions. In this study, we collected transaction-level data from 996 students enrolled in an online introductory spreadsheet class. Each student's final answer and step-by-step attempts were coded for misconceptions or knowledge gaps regarding the use of absolute references over four assessment occasions. Overall, the level of error revealed was significantly higher in the step-by-step processes compared to the final submitted answers. Further analysis suggests that students most often have misconceptions regarding non-critical errors. Data analysis also suggests that misconceptions identified at the transaction level persist over time.

Keywords: Educational data mining, Assessment, Data logs, learning analytics

Transaction-Level Learning Analytics in Online Authentic Assessments

One of the purposes of assessment in education is to evaluate students' comprehension and ability (Harlen, 2007). Accurate assessment becomes especially important as students engage in complex learning tasks that require them to build on prerequisite knowledge. A student who does not properly master prerequisite concepts may have difficulty mastering more intricate concepts and techniques later in the learning process (Khan, 2012). Thus, assessment methods need to be valid and reliable in determining whether a student has mastered these preliminary concepts.

Unfortunately, all forms of assessment come with some degree of measurement error (Miller, Linn, & Gronlund, 2013). This is particularly true of test items that do not have one correct answer and where scoring is automated. A student may answer a question correctly despite holding a misconception or knowledge gap about the content; and selecting a correct answer does not always indicate that the student really knows the material (Pelligrino, Chudowsky, & Glaser, 2001). Ideally, knowledge gaps and misconceptions would be identified through use of more detailed forms of assessment. Similar to a requirement to "show your work" on a math assessment, a detailed observation of the student's step-by-step progressions in completing a task may provide a more accurate understanding of what the student knows. With this additional information, teachers can better identify gaps in the student's understanding and provide remediation to students. Although observing the steps a student takes when completing a task may be feasible in small classes or one-on-one tutoring sessions, it does not typically occur in classrooms with large numbers of students and rarely if ever is undertaken in online learning environments. Instead, most online learning environments tend to rely on objectively scored assessments for the sake of efficiency (Davies & West, 2013). Recently, however, some

online learning platforms have been equipped with the ability to collect detailed student activities and assessment responses in the form of data logs (Siemens, 2012). These logs include step-by-step descriptions of the process a student went through in order to complete a task. By capturing these step-by-step (or transaction-level) data, these systems have the potential to do a more detailed analysis of student learning. However, in many ways online educators are only beginning to utilize data to analyze student learning (Baker & Yacef, 2009).

Here, we offer a case study of the benefits of using detailed assessment data at the transaction level. We examined data from an instructional system designed to teach spreadsheet concepts in an online learning environment. The grading engine for this system captures not just the final solution a student submits but also the steps students take in arriving at their final solution (i.e., transaction-level data). This allows us to compare errors students make in their final answer with those students make as they arrive at the final solution that they submit for grading. This research presents the first phase of a larger project dealing with this type of data. In this initial work we used transaction-level data to uncover and identify misconceptions persisting among students for one knowledge component—the use of absolute references (holding constant the row or column in a cell reference when copying the content of that cell), in an online course. Because we believe that knowledge gaps and misconceptions often go undetected when a learning system relies solely on final submitted answers, we anticipated that this study would be foundational in exploring the use of transaction-level data to facilitate better feedback and remediation in online instruction.

Background Information

Interest in the use of educational data mining and learning analytics has increased dramatically in recent years (Baker & Yacef, 2009; Ferguson, 2012). Researchers have used a

variety of techniques to gather, process, and visualize educational data, with no uniform way to classify the types of data used. In this paper, we have employed Chung's (2014) classification system that describes three levels of educational data: system-level, individual-level, and transaction-level.

System-level data is typically aggregated at the school level, frequently including data regarding students' completed courses, entrance exam scores, grades, and demographic information. Research at this data level has focused on predicting success and developing early-alert systems for students at risk for dropping out (Arnold & Pistilli, 2012; Bowers, 2010; Campbell, 2007; Morris, Wu, & Finnegan, 2005). Since the focus of these systems is to inform institutions regarding general policy decisions, they do not provide specific feedback to students regarding content knowledge and performance that might be useful in the classroom. The second level of data, assessment-level data, is composed of individual students' assessment scores for a given class. This level of data, typified in a course grade book, is generally the most readily available to instructors. Much of the research in learning analytics has focused at this level, including research in adaptive systems using intelligent tutors (Abdous, He, & Yen, 2012). However, these data do not often provide diagnostic insights that might be used for remediation.

Transaction-level data, composed of the individual steps a student went through in completing an assessment (Chung, 2014), offers the most detail, but it is often difficult to obtain and can be challenging to analyze. Despite this difficulty, several researchers in the areas of educational data mining and learning analytics have begun to work with data at the transaction level. Several studies have used transaction-level data to monitor students while learning to program (Berland, Martin, Benton, Petrick Smith, & Davis, 2013; Blikstein, 2011), developing games (Koh, Basawapatna, Nickerson, & Repenning, 2014), engaging in collaborative activities

(Perera, Kay, Koprinska, Yacef, & Zaiane, 2009), or using study tools (Nesbit, Zhou, Xu, & Winne, 2007). These studies, however, were descriptive in nature, trying to understand general patterns of student activity based on the data trace in the transaction-level data.

Our goal is to demonstrate how transaction-level learning analytics can support a model of cognitive apprenticeship (see Collins, Brown, & Newman, 1987). Cognitive apprenticeship attempts to make the thinking of experts visible to students. In our case, expert thinking comes in the form of identifying knowledge gaps and misconceptions a student might have when attempting to solve a problem. As such, the primary goal of our study was to determine the degree to which the use of transaction-level data might better identify misconceptions and knowledge gaps that may not be identified through an analysis of the final answer (i.e., assessment-level) data alone.

Problem Solving and Knowledge Components

Identifying and remediating student knowledge gaps has been the goal of researchers working with intelligent tutoring systems (ITS). Using a cognitive apprenticeship framework, these systems' attempt to communicate an expert's knowledge to students via a computer (Wenger, 1987), they do this by engaging in a process of knowledge tracing to determine the current state of the students' knowledge at each step of the instruction process (Corbett & Anderson, 1995). When a student's knowledge does not match that of the expert model, hints or opportunities for remediation are presented. The ability to accurately identify and report specific knowledge gaps and misconceptions is essential in this process.

ITS literature has classified knowledge as either (a) goal-independent declarative knowledge or (b) goal-oriented procedural rules (Corbett & Anderson, 1995). Both categories can contain collections of knowledge components, which VanLehn (2006) defined as "a

principle, a concept, a rule, a procedure, a fact, an association or any other fragment of task-specific information" (p. 3). A student uses existing knowledge components and learns new knowledge components when participating in a learning event (Koedinger, Corbett, & Perfetti, 2010).

The learning events in a course of study are directed by its learning objectives, which are usually made up of complex tasks that consist of a combination of knowledge components. Therefore, an appropriate combination of knowledge components may be required for a student to accomplish a learning objective. Koedinger et al. (2010) argued that knowledge components could be derived through student behavior on assessment events. Further, they claimed that knowledge components "are not pre-determined by instructional designers, but can be empirically derived from sets of tasks that instructional designers can specify" (p. 11).

However, in the process of learning and applying knowledge components, students are bound to have errors in their knowledge. Brown and VanLehn (1980) worked to identify the systematic errors or "bugs" that exist in student knowledge, which they defined as "complex, intentional actions reflecting mistaken beliefs about the skill" (p. 380). In our study, we refer to these systematic student errors or bugs as *knowledge gaps*. We feel that these knowledge gaps are traditionally overlooked when assessment-level data are utilized alone, and more detailed, transaction-level data are necessary to accurately identify such errors. While several studies have focused on defining errors in spreadsheets (Panko, 2013; Panko & Aurigemma, 2010), no study was identified which attempted to locate errors or misconceptions in student knowledge regarding spreadsheets using transaction-level data.

Although the authentic assessment system in this study is different from an ITS in that (with the exception of the feedback it provides) it is not adaptive in nature, we used a similar

process from the literature to identify knowledge components from the set of tasks students are expected to complete. By looking at the transaction-level data associated with one of these knowledge components, we hoped to identify knowledge gaps that might otherwise go undetected.

Research Purpose and Questions

The purpose of this study was to identify knowledge gaps using transaction-level student log data obtained from an online spreadsheet course. Our intent was to demonstrate the ability and value of transaction-level data to identify otherwise undetected knowledge gaps and to lay the groundwork for future use of transaction-level data to automate feedback, remediation, and possibly potential to adapt and differentiate the instruction provided. While many knowledge components are taught in the targeted spreadsheet class, because of the scale of the data, we chose to look at one knowledge component: the use of absolute references. While absolute references are not the most difficult concept to master in the course, they are important and represent a knowledge component that many students struggle to master. We identified three primary research questions for this study:

- What is the difference between knowledge gaps present in transaction-level data compared to final answer assessment data?
- Which errors (i.e., knowledge gaps and misconceptions) are revealed most often by transaction-level data and by final solution data?
- To what degree do knowledge gaps persist across assessment occasions in the course?

Methods

Data for this research were gathered from student assignments in an Introduction to Excel class hosted on the MyEducator platform. In this course, students were given the assignment to

complete worksheets, each of which incorporated the Hidden Event Log for Individual Observation System or HELIOS, which created a detailed log of the steps each student used to arrive at a solution for the assigned task. Transaction-level data from this log were used along with the final answers for this study. The log collected student ID, assignment and worksheet ID, cell ID, formula entered, and resulting value displayed.

Existing log data were collected from students enrolled in the class at two universities in the western United States in the winter 2014 semester (January–April). Courses from each of the universities used identical learning resources and assessments. Individual student logs were collected and aggregated into a single data file for analysis, which identified four separate assignments on which students had to use absolute references, the specified knowledge component for the study. Only those students who completed all four of the assignments were included in the final analysis; therefore the sample for this study included 996 students out of 1128 who were enrolled in the course.

Problem Description

In a preliminary analysis of the data, we identified a list of basic knowledge components that student would need to complete each of the assigned tasks in this instructional system—sample knowledge components included skills such as (a) the ability to correctly select and use specific functions including their associated arguments (e.g., COUNT, RATE, IF), (b) the ability to reference cells and create basic mathematical formulas, and (c) the ability to copy and paste a formula down a column or across a row. A combination of these knowledge components would be needed to correctly solve a specific problem. For our present analysis, we focused on students' understanding of a single knowledge component, absolute references. This action is designated by placing a $ before the column letter or row number to be held constant when

copying the cell content to another cell, down a column of cells or across a number of rows. We decided on absolute references because, based on a preliminary analysis, the concept of absolute references is difficult for students to master and because there were multiple occasions throughout the course where students were required to use absolute references to correctly complete a task. Focusing on one knowledge component allowed us to create a case study with the purpose understanding the possibilities of using transaction-level data to identify knowledge gaps.

After completing an analysis of answers students provided for each problems, we identified four ways that a student might use an absolute reference incorrectly: (a) using an absolute reference when or where it was not needed, (b) failing to use the absolute reference when it was needed, (c) using the absolute reference inappropriately, and (d) typing in a value rather than using a cell reference to make the solution work. While each error represents a potential knowledge gap or misconception, some would be considered more problematic than others. Therefore, each error was assigned a weight based on the severity of the error. We considered using an absolute reference when it was not needed to be a minor error because doing so does not cause a problem when copying the cell down the column, but the misstep exposed a potential knowledge gap. All of the other mistakes were considered major errors, some more severe than other, as they clearly exposed a knowledge gap that if left uncorrected would likely cause problems when the student attempted to copy the solution or later on endeavored to manipulate values. A student might make more than one error when applying absolute references to their solution.

On each occasion we studied, the task required students to correctly use cell references and a formula or function to obtain a solution. They were then required to copy the solution

down the column to complete the spreadsheet table. On the first occasion they were required to create a formula and place it in cell D11 (see Figure 1). The task, explained in a text box required that the formula utilize an absolute reference to the cell C8 and a relative reference to the cell C11 in order to copy the cell contents down the column correctly. Examples of each of the errors for the first occasion can be found in Table 1.

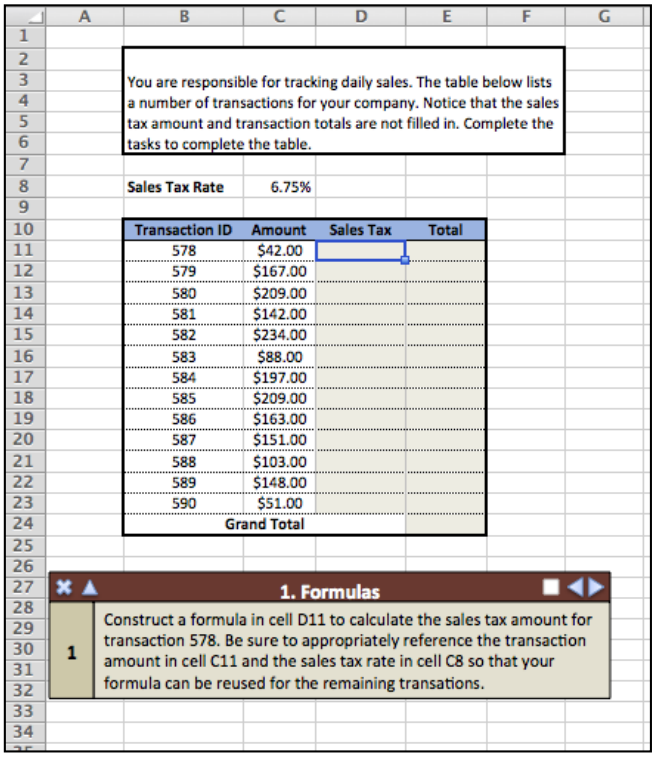

Figure 1. Sample task. Students are required to use a formula or function to solve the problem and place the solution in cell D11. The student must then apply absolute references so the solution it can be copied down the column.

Table 1

Examples of Solutions That Reveal Knowledge Gaps in Absolute Reference

Sample solution	Knowledge gap evidenced, severity of error	Error weighting
=C11*C$8	None	0.0
=$C11*$C$8	Used but not needed, minor issue	0.05
=C11*C8	Needed but not used, major issue	0.3
=C$11*C$8	Used inappropriately, major issue	0.5
=C11*0.0675	Use avoided by typing in value, major issue	0.75

There were four occasion where absolute references were required in the course. While each occasion posed a slightly different problem, on all four occasions students were required to solve a problem using a formula or a function that referenced a value in at least two other cells. One of the cells represented a constant value that required an absolute reference if the cell content was to be copied correctly. In all cases, the problem referenced at least one additional cell in the row (a relative reference) then required the student to copy the solution down the column to complete the spreadsheet table. Thus in each problem a relative and an absolute reference was needed to complete the problem. On the first occasion, the topic of the lesson was absolute references, and the instruction explained how to use them. On each of the subsequent instructional occasions, the instruction was directed toward a different aspect of the spreadsheet program. The midterm exam had no instructional component, only assessment. After the first occasion, students were expected to remember how to use absolute references correctly, but they were allowed to review instructional materials from any lesson. The problem presented on the midterm (the last occasion we observed) was quite similar to the problem described in Figure 1 (the problem present in Lesson 2). Although student may have made errors in the formula or

function on each occasion, only errors associated with absolute references were analyzed in this study.

Data Cleaning and Coding

Data cleaning and coding were completed manually in Microsoft Excel. Student data logs included both the transaction-level data (i.e., time sequenced step-by-step attempts the student made to complete the problem) as well as the final solution submitted by the student. After filtering for the assessment in question data from the student log files were cleaned to remove blank entries and non-numeric symbols. This created a list of unique answers for each of the four tasks that required absolute references. The lists were then sorted first by student and subsequently by chronological step. With these data, we created a master error list in which we identified errors associated with each of the unique solutions submitted by students. The first task was coded manually. The process of error coding was then automated using an Excel function that searched for evidence that one or more of the four error types had occurred. For example, in Lesson 2 the function found evidence of the student using an absolute reference where it was not needed by searching for the presence of $C or $11. To test whether this method of coding was comparable to coding the answers by hand, we computed an intercoder reliability calculation between the methods. The reliability level was .92 (or 92% agreement); so although the automated identification was not perfect, we felt that we could proceed with using a function to code all of the answers to expedite the process.

Once all of the unique answers for each of the problems that required absolute references were coded for errors, a total error score was calculated. Each of the four types of error was weighted based on the severity of the error type. The total error for a specific solution represented a consistent ordinal weighting based on severity for each error evidenced in the data.

Total error for each solution ranged from 0, indicating no error evident, to 1.3, indicating that one or more of the 4 errors were identified in that step (see Table 1). However, the total score did not score more harshly for overlapping issues. For example, if the student typed in the value to avoid using an absolute reference, the error rating was determined to be 0.75; thus a student's answer would not include both an error score for avoiding the use of an absolute reference by typing in a value and another error score for failing to use an absolute reference when needed. To calculate the mean error for a student's transaction-level, the error from each step was averaged. The error value of the final submitted step was based on scoring of the single final submission.

Data Analysis

To determine whether the amounts of error displayed in the transaction-level data were different from the amounts in the final answer data, we used paired sample T-tests. Cohen's d was used to calculate an effect size for each comparison. This statistical test was appropriate because we wanted to determine differences on a student-by-student basis. To answer our other research questions, we also used descriptive statistics and graphic representations to identify persistent patterns in the data.

Results

Our first research question asked whether there was a difference in the level of error that could be identified through data obtained from the transaction-level data compared to data from the final answer alone. Averages of error across the transaction-level solutions were compared with the error found in the final submitted answers. The results of the paired sample T-tests between the solution process and final answer by occasion are displayed in Table 2. As we anticipated, there was a significant difference between the average errors found in the process

data and in the final submitted answer data for each of the occasions. The effect size in each case was large, over two standard deviation units difference in each case, suggesting that there is a greater amount of student knowledge gaps detected in the transaction-level data compared to the final submitted answer. The largest difference between transaction-level and final answer data can be seen in Lesson 4. Here the mean error on the final step for all of the students was .070 (a relatively small amount indicating only minor mistakes were made which did not affect the result) while the mean error on the transaction level data was .177 (indicating that students struggled somewhat to obtain a suitable solution).

Table 2

Results of Comparisons Between Process and Final Answer by Occasion

Occasion	Mean process error	Mean final error	*t*	*p*	*Effect size*
Lesson 2 (D11)	.087 (.130)	.036 (.103)	17.86	<.001	2.53
Lesson 3 (E18)	.236 (.230)	.127(.235)	20.70	<.001	2.71
Lesson 4 (K17)	.177 (.097)	.070 (.100)	38.74	<.001	6.27
Midterm (D20)	.182 (.126)	.078 (.127)	25.57	<.001	4.75

Note. Standard deviations appear in parentheses adjacent to means. Error values range from 0 to 1.3 in severity.

Frequency of Error Comparison

The frequency with which specific errors occurred in the transaction-level data was also studied. The frequency of each type of error is shown in Table 3 disaggregated by error type. Overall, the most common type was Error 1, using an absolute reference when it was not needed. This error was committed by the majority of students (over 70% from the second occasion on). These results tend to indicate students did not yet know when or where to use an absolute reference. And possibly, because it would not affect the outcome, they left it in the final solution

rather than removing the unnecessary addition. Students also seemed to struggle with Error 2, not using an absolute reference when it was needed. However, unlike Error 1, students tended to correct this mistake in the final solution. One possible explanation is that they completed the process in two steps: first creating the formula and then applying the absolute references. However, we could not verify this possibility for all cases.

Table 3

Frequency of Absolute Reference Errors by Occasion

	Error 1		Error 2		Error 3		Error 4	
	Process	Final	Process	Final	Process	Final	Process	Final
Lesson 2 (D11)	268 (26.9%)	246 (24.7%)	341 (34%)	36 (3.6%)	17 (1.7%)	1 (.1%)	44 (4.4%)	13 (1.3%)
Lesson 3 (E18)	716 (71.8%)	701 (70.3%)	488 (48.9%)	5 (.5%)	62 (6.2%)	1 (.1%)	295 (29.6%)	120 (12.0%)
Lesson 4 (K17)	767 (76.9%)	763 (76.5%)	796 (79.8%)	76 (7.6%)	60 (6.0%)	12 (1.2%)	9 (.9%)	5 (.5%)
Midterm (D20)	886 (88.9%)	853 (85.6%)	671 (67.3%)	8 (.8%)	361 (36.2%)	57 (5.7%)	12 (1.2%)	5 (.5%)

Error 1: absolute reference used when not needed, minor error
Error 2: absolute reference not used when needed, major error.
Error 3: absolute reference used inappropriately (on the wrong cell), major error
Error 4: use of absolute reference avoided by typing in cell value, major error

Both Error 3, using an absolute reference incorrectly for the situation, and Error 4, typing in a value to avoid using an absolute reference, were far less likely to be an issue for students. With the exception of the midterm exam, students rarely used the absolute reference on the wrong part of the cell reference. And as with Error 2, students tended to fix this error in their final submitted solution, as the solution would be incorrect if they had not made the

modification. Error 4, tying in a value instead of using a cell reference, was a serious mistake but (with the exception of Occasion 2) students rarely committed it, and many corrected it before submitting the final answer. This indicates that a majority of the students understood that they needed to use an absolute reference to complete the problem; however, they may have had misconceptions regarding how to apply the references correctly.

Patterns of Error over Time

To answer our final research question, we looked for patterns in the data that might inform our understanding of how knowledge gaps regarding absolute references persist throughout the course. Ideally, a student's repeated attempts to practice a skill decrease the rate of error—a phenomenon often referred to as a learning curve (Corbett & Anderson, 1995). A typical learning curve would present itself with high error counts in initial occasions followed by few errors in subsequent occasion. Figure 2 graphically represents the mean error by occasion for both the transaction-level and final answers submitted by students. The graph reveals a learning curve different from that theorized by Corbett and Anderson (1995). Aggregated errors tended to spike on the second occasion and then remained somewhat constant in Lesson 4 and the Midterm. The initial low level might be explained by the fact that the topic of instruction for the first occasion was the use of absolute references. The more heavily scaffolded instruction provided by the learning system may have resulted in fewer errors being observed on this occasion in both the process and the final solutions submitted. On the second and subsequent occasions, students were expected to have already learned how to use an absolute reference and would have had to either remember what to do or relearn the concepts.

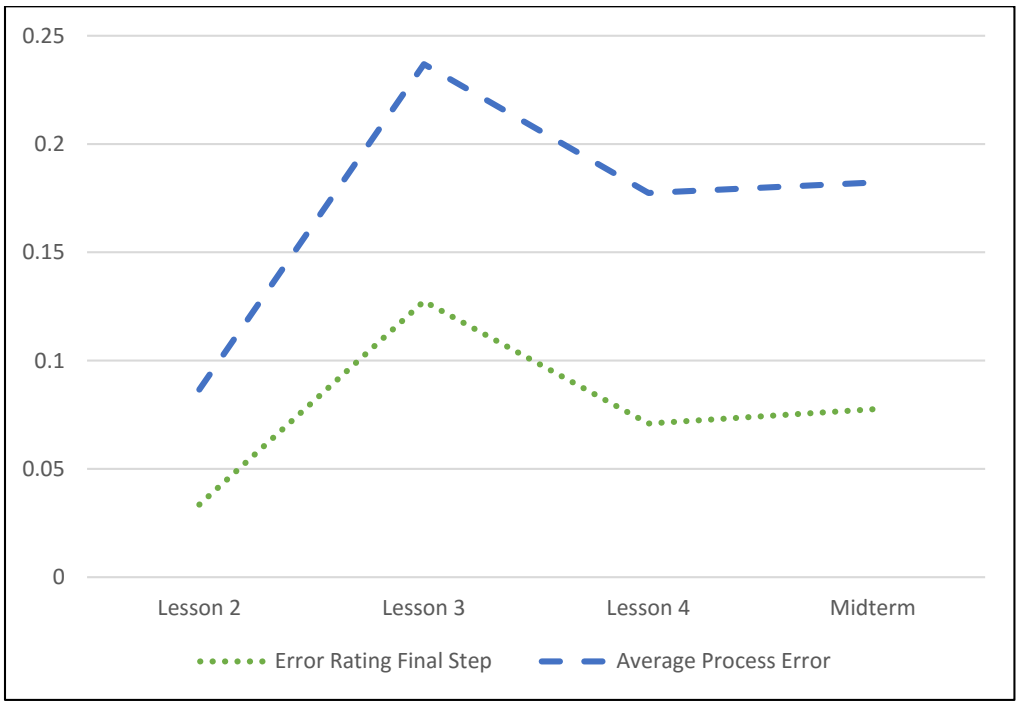

Figure 2. Error level for the final step and average errors for the process, by occasion.

While Figure 2 was helpful in examining overall the progress of students across time, the heat map in Figure 3 is intended to help illustrate how individual students' errors changed over time. This idea is similar to Bowers (2010), who used a heat map to track the performance of students across their entire K-12 education. This visualization illustrates the error level of individual students regarding absolute references across occasions rather than the mean error of all students. Each individual is represented by one row in the graph. A darker line indicates the mean error level for an individual student in the solution process on that occasion. While the heat map is not an empirical measure of student struggle, it provides an at-a-glance picture of struggle for individual students. Ideally, we would hope that error rates for individual students would improve across occasions (i.e., fewer dark lines across time).

In looking at Figure 3, on the first occasion, few errors were found in students' step-by-step process. However, on the second occasion many of those students who previously

demonstrated little error previously demonstrated high error levels. By the third and fourth occasions, the errors seemed to be randomly distributed. Students who had few or no errors on prior occasions made mistakes at this later time, suggesting a knowledge gap or learning gap had developed and was presenting in the third and fourth occasions. Overall, this result illustrates a great deal of fluctuation in student understanding. Several students showed no indication of any learning gap or misconceptions until the midterm exam. Still others consistently struggled with the concept across occasions. Fewer students demonstrated full understanding across each of the four occasions. These results suggest that there is no one single pattern for learning this concept that can be derived from the data, and that any instructional remediation may need to be considered on an individual basis.

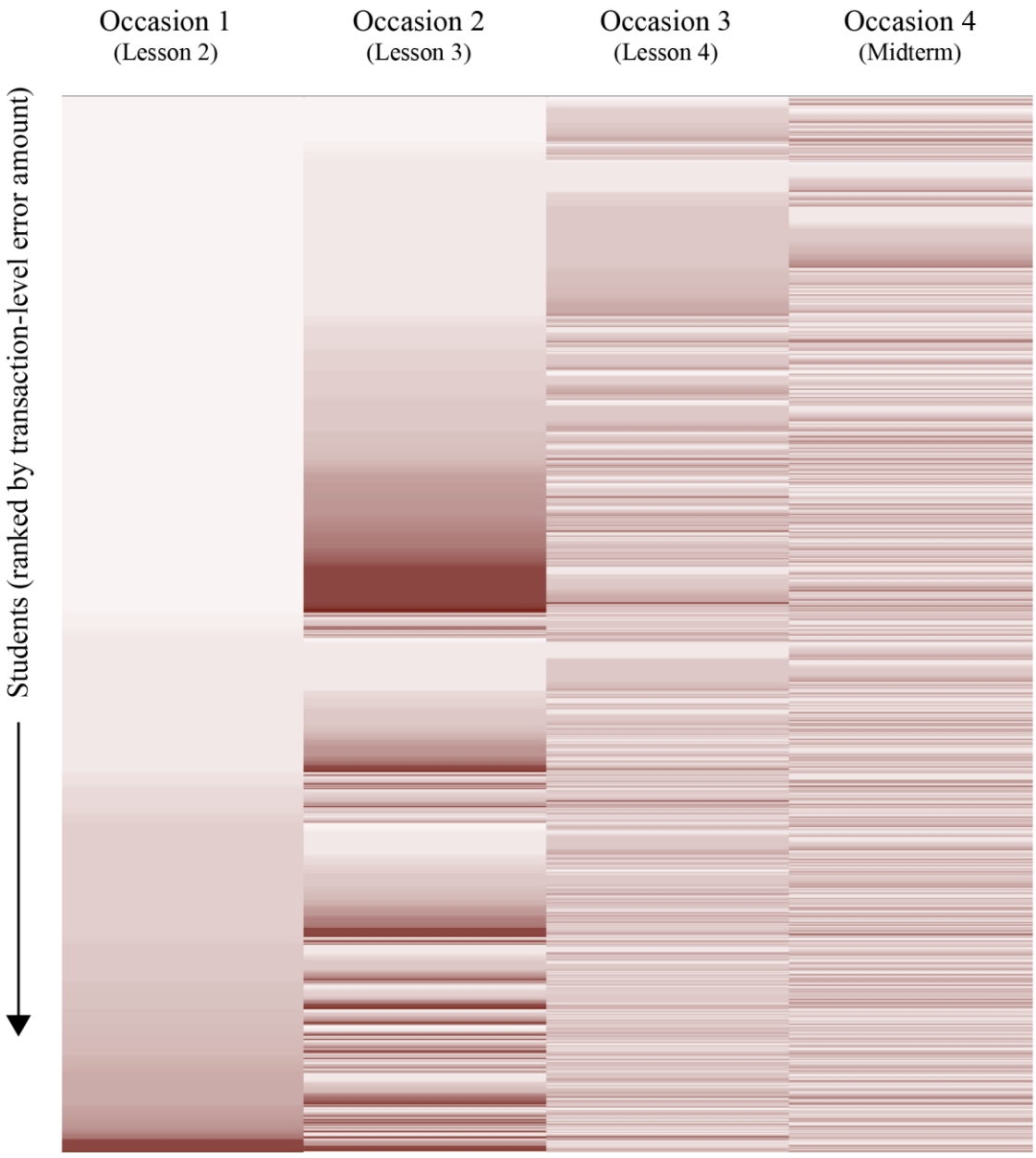

Figure 3. Heat map illustrating student error level in process solution (vertical axis) by occasion (horizontal axis). Error is sorted by occasion 1, then occasion 2, and on. Dark bands represent occasions on which students had more error in their solution.

Discussion and Conclusions

An analysis of these results supports our hypothesis that there would be a significant difference between the number of errors uncovered using transaction-level data (which trace the activity of the student in solving the problem) and the number of errors uncovered using the final submitted answer alone. Over 90% of the students seemed to hold some misconception or knowledge gap about the concept of absolute references based on analysis of the transaction-level data, although only 7% of these students failed to get the designated item correct on the midterm exam. As a result, we concluded that assessing only the final solution tends to give a false sense of adequate academic achievement for this knowledge component. The differences identify a degree of struggle student had in arriving at a final solution that would not be evident when looking at the final solution alone.

While such a finding may seem obvious—undoubtedly students will make mistakes as they struggle to work through a problem—it represents an important empirical step in the processes of educational data mining and learning analytics. If the purpose of the data analysis is to diagnose knowledge gaps that might inform remedial action, online assessments that rely only on final answers would not be adequate in identify which requisite knowledge components may have prevented a student from being able to complete the task or, more frequently, fail to identify knowledge gaps entirely. With notification only that an answer is incorrect, the instructor has no diagnostic information to guide remediation for the student (Popham, 2005). Most often specific knowledge gaps are simply ignored if a student performs sufficiently overall (Khan, 2012).

Differences in Types of Errors

Through our analysis of these data, we also discovered that not all errors are equivalent. Some errors, for example the non-critical error of including an absolute reference where it is not

needed, are present in the data but often ignored as evidence of any misconception or knowledge gap. Although including an unnecessary absolute reference does not alter the ability of the student to get a functional solution, this mistake not only indicates a potential knowledge gap, it represents an inefficient sub-optimal solution which often causes individuals to expend time and energy struggling to solve a problem we believed they had already mastered. When students lack prerequisite knowledge component it can restrict future learning that builds on it (Khan, 2012).

Other more critical errors present in the data are most often corrected by students before they submit the final solution. The challenge for instructional designers and educators is to decide which errors need to be addressed through remedial instruction and which can be overcome somewhat naturally through practice. For example, on the second occasion, about 30% of the students made the mistake of typing in a value rather than using a cell reference and absolute reference to solve the assigned problem, but the frequency of this error diminished considerably after that point. Students seemed to realize that they should use a cell reference, and thus an absolute reference, without any additional remedial intervention. This was not the case with some of the other critical errors, which tended to persist over time and (for at least 6% of the students) were never corrected in their solutions even though most of these students passed the course.

Learning Patterns Uncovered by Transaction-level Data

Our method for identifying particular knowledge gaps using transaction-level data was helpful in identifying and understanding the most common errors students encounter. Nonetheless, there did not seem to be a singular pattern in the students' learning curve. While it was expected that errors would diminish with practice (Corbett & Anderson, 1995), in this study

a student's struggle to complete the assigned problem without error did not always diminish over time as theory suggested it should. In fact, we surmise that the act of diagnosing learning difficulties must be completed at an individualized level and will differ for specific knowledge components.

As Romero, Ventura, Pechenizkly, and Baker (2011) pointed out, useful and intelligent adaption of instruction needed for personalization requires information about the individual student, not general trends of the average student. In this study, most students did not seem to struggle initially with the concept and application of absolute references. However, the step-by-step instruction provided on the first occasion (Lesson 2) did not seem to adequately establish the requisite learning for further application. Many though not all students seemed to struggle and needed to relearn the concepts on the second occasion. For possibly a variety of reasons, many students showed no indication of knowledge gaps with this concept until the midterm exam. These differences lead us to believe that decisions about remedial instruction need to be informed by data beyond final answers submitted.

Obtaining this information can be time consuming and requires technology to automate the process (Chung & Kerr, 2012; Mayer, 2009; Siemens, 2012). While the activity trace data used in this study were extensive, we did not utilize all the data that were made available through this learning management system. Additional information—including the number of steps taken to solve a problem, the time elapsed, the instructional material viewed by each student , and prior experience and achievement of individual students—are but a few of the data points that might be useful in making decisions about the need for and appropriateness of remediation.

Conclusions

This study demonstrated the value of using transaction-level data in identifying knowledge gaps for one specific knowledge component. Many courses contain hundreds of knowledge components, and not all will be a good fit for processes of capturing appropriate transaction-level data. However, for the specific knowledge component selected for this study we were able to identify the most common types of errors by using transaction-level data. Although we could not detect any specific generalizable pattern in the error data, we were able to conclude that the use of transaction-level data is superior to the use of final answer data in accurately assessing students' learning and identifying their knowledge gaps and misconceptions.

While much progress has been made, with considerable research conducted, the development of truly intelligent adaptive instruction currently remains a holy grail for instructional designers as they develop technology-enabled instructional systems (Woolf, 2010). Much remains to be accomplished. While many technology-enabled instructional systems have been created in attempts to utilize data to inform instructional adaptations, few function in real time at the process level (Siemens, 2012). Many adaptations fail to identify and properly diagnose learner misconceptions and knowledge gaps because they are not completed in real time using transaction-level data; thus, they do not lead to effective changes in instruction. As a result many students successfully complete courses (based on overall grades) without gaining the intended learning they will need to successfully accomplish more advanced learning objectives in future courses (Khan, 2012).

An increasing number of institutions expect implementation of educational data mining and learning analytics (Cummins, Johnson, & Adams, 2012). Intelligent tutoring and

remediation using these types of data will eventually function in real time, with additional applications. Analysis of transaction-level data has the potential to inform designers where to prioritize their efforts by identifying areas in which students tend to struggle and aspects of a course that need additional emphasis (Lehikoinen & Koistinen, 2014).

Future Research

Most of our efforts in this project included identifying common errors for a specific knowledge component, cleaning and coding student data, and running final data analysis. These processes were performed manually and not in real time. While this methodology was adequate for the purposes of the current study, future research needs to identify methods that will automate much of this process. At the same time, we are continuing to identify other knowledge components being taught in the course studied that would benefit from an analysis of transaction-level data.

Future research is also needed to put what we learn from data into action. While this study is informative, we are working on ways to improve the instruction and ways to automate a recommender system based on the need for remediation. To accomplish this, however, we need more information. The volume of data can be overwhelming, and data must be organized in order to be used. In addition, research on visualization of transaction-level educational data is needed. If our goal is to improve instruction by parsing through data to find patterns of misconceptions, then we must determine appropriate ways of presenting the results to instructors, designers, and students.

References

Abdous, M., He, W., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Educational Technology and Society*, 15(3), 77–88.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of the *2nd International Conference on Learning Analytics and Knowledge*—LAK '12 (pp. 267–270). New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2330666

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–16. Retrieved from http://educationaldatamining.org/JEDM/index.php/JEDM/article/view/8

Berland, M., Martin, T., Benton, T., Petrick Smith, C., & Davis, D. (2013). Using learning analytics to understand the learning pathways of novice programmers. *Journal of the Learning Sciences, 22*(4), 564–599. doi:10.1080/10508406.2013.836655

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. In Proceedings of the 1st *International Conference on Learning Analytics and Knowledge* - LAK '11 (pp. 110–116). Banff, Alberta: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2090132

Bowers, A. J. (2010). Analyzing the longitidunal K-12 grading histories of entire cohorts of students: Grades, data drive decision making, dropping out and hierachical cluster analysis. *Practical Assessment, Research & Evaluation, 15*(7), 1–18.

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4*, 379–426. doi:10.1207/s15516709cog0404_3

Campbell, J. P. (2007). Utilizing student data within the Course Management System to determine undergraduate academic success: An exploratory study. (Doctoral dissertation). Retrieved from http://docs.lib.purdue.edu/dissertations/AAI3287222/

Chung, G. (2014). Toward the relational management of educational measurement data. *Teachers College Record, 116*(11), 1-16. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17650

Chung, G. K. W. K., & Kerr, D. (2012). *A primer on data logging to support extraction of meaningful information from educational games: An example from Save Patch* (CRESST Report 814). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Collins, A., Brown, J. S., & Newman, S. E. (1987). *Cognitive apprenticeship: Teaching the craft of reading, writing* (No. 403). Champaign, IL: University of Illinois at Urbana-Champaign.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction, 4*(4), 253–278. doi:10.1007/BF01099821

Cummins, M., Johnson, L., & Adams, S. (2012). *The NMC horizon report: 2012 higher education edition*. The New Media Consortium. Retreived from http://www.nmc.org/pdf/2012-horizon-report-HE.pdf

Davies, R., & West, R. E. (2013). Technology integration in school settings. In M. Spector, M. J. Bishop, M. D. Merrill, & J. Elen (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 841-853). New York, NY: Lawrence Erlbaum.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304-317. doi:10.1504/IJTEL.2012.051816

Harlen, W. (2007). *Assessment of learning.* Los Angeles, CA: SAGE Publications, Inc.

Khan, S. (2012). *The one world schoolhouse: Education reimagined.* New York, NY: Grand Central Publishing.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2010). *The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning* (Technical Report). Pittsburg, PA: Carnegie Mellon University. Retreived from http://repository.cmu.edu/cgi/viewcontent.cgi?article=1300&context=hcii

Koh, K. H., Basawapatna, A., Nickerson, H., & Repenning, A. (2014). Real time assessment of computational thinking. In Proceedings from *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 49–52). IEEE. Retrieved from http://sgd.cs.colorado.edu/wiki/images/9/91/Paper_24.pdf

Lehikoinen, J., & Koistinen, V. (2014). In big data we trust? *Interactions, 21*(5), 38-41.

Mayer, M. (2009). *Innovation at Google: The physics of data.* Retrieved from http://www.parc.com/event/936/innovation-atgoogle.html

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching (11th ed.)*. Upper Saddle River, NJ: Prentice-Hall.

Morris, L. V., Wu, S., & Finnegan, C. L. (2005). Predicting retention in online general education courses. *The American Journal of Distance Education, 19*(1), 23–36. doi:10.1207/s15389286ajde1901

Nesbit, J., Zhou, M., Xu, Y., & Winne, P. (2007). Advancing log analysis of student interactions with cognitive tools. In Proceedings of *12th Biennial Conference of the European Association for Research on Learning and Instruction* (pp. 1–20). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Advancing+Log+Analysis+of+Student+Interactions+with+Cognitive+Tools#0

Panko, R. R. (2013). The Cognitive Science of Spreadsheet Errors: Why Thinking is Bad. In Proceedings of *2013 46th Hawaii International Conference on System Sciences* (pp. 4013–4022). IEEE. doi:10.1109/HICSS.2013.513

Panko, R. R., & Aurigemma, S. (2010). Revising the Panko–Halverson taxonomy of spreadsheet errors. *Decision Support Systems, 49*(2), 235–244. doi:10.1016/j.dss.2010.02.009

Pelligrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know.* Washington, DC: National Academy Press.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 759–772. doi:10.1109/TKDE.2008.138

Popham, W. J. (2005). *Classroom assessment: What teachers need to know (4th ed.)*. Boston, MA: Allyn and Bacon.

Romero, C., Ventura, S., Pechenizkly, M.,& Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. Baton Rouge, FL: CRC Press.

Siemens, G. (2012). Learning analytics: Envisioning a research discipline and a domain of practice. In Proceedings of the *2nd International Conference on Learning Analytics and Knowledge—LAK '12*, 4-8. New York, NY: ACM press.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265. Retrieved from http://dl.acm.org/citation.cfm?id=1435353

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computation and cognitive approaches to the communication of knowledge*. Los Altos, CA: Morgan Kaufmann Publishers.

Woolf, B. P. (2010). *A roadmap for education technology*. Retrieved from http://cra.org/ccc/wp-content/uploads/sites/2/2015/08/GROE-Roadmap-for-Education-Technology-Final-Report.pdf

ARTICLE 3:

Linking LMS Activity Data with Transaction-level Assessment Data

Linking LMS Activity Data with Transaction-level Assessment Data

Rob Nyland

Randy Davies, PhD.

Conan Albrecht, PhD.

John Chapman

Gove Allen, PhD

Brigham Young University

Corresponding Author

Rob Nyland

150 MCKB, IPT Department

Brigham Young University

Provo, UT 84602

robnyland@gmail.com

Abstract

This paper presents a case study for using rich data in a courseware environment that relies on authentic assessments. This environment, an online Introduction to Microsoft Excel course, collects detailed data regarding video and text usage along with step-by-step (transaction-level) data regarding how students complete assessments. This transaction-level data gives us a better understanding of student misconceptions in a course. In this paper, we examine the relationship between reading and video activities, and overall levels of errors on the assessments. Because the data violated assumptions for normality, Zero-inflated Negative Binomial and Negative Binomial Regressions were used to model the data. Overall, we found that these models explained less than three percent of the variance in cumulative error for a given knowledge component. We also found that reading time was the strongest predictor of students having increased error on a problem, while percentage of video watched was the best predictor of decreased error for a problem.

Linking LMS Activity Data with Transaction-level Assessment Data

Feedback, which Shute (2008) defines as "information communicated to the learner that is intended to modify his or her thinking or behavior" (p. 154), is a critical component in the teaching and learning process. This was emphasized by Hattie's (1999) meta-analysis of thousands of educational studies and their interventions. After finding that feedback had an average effect size of .65, he argued that, "the most powerful single moderator that enhances achievement is feedback. The simplest prescription for improving education must be 'dollops of feedback'—providing information [on] how and why the child understands and misunderstands, and what directions the student must take to improve" (p. 11). Not all feedback is created equal however, with some techniques being more effective than others. Hattie & Timperley (2007) attribute feedback in the form of praise, rewards, and punishment with lower effect sizes, while feedback on specific task performance was equated with higher effect sizes. Shute (2008) agrees that feedback should be specific to task performance, and in addition should be given in an objective tone, presented in manageable units, be specific and clear, and given to the student after they've attempted the solution.

While evidence suggests that feedback is important in education, there can be challenges in giving it in an online context. For instance, many online programs trade synchronous face-to-face contact in order to make learning more flexible. This however, has effects on the immediacy of feedback that is available. An example of this was shown in Kim, Liu, and Bonk's (2005) qualitative evaluation of an online MBA program. While students in the study liked the overall experience of the MBA, they wished that there were more real-time feedback and interaction in the course. There can also be issues in providing feedback at scale. This can readily be seen in the feedback that is received in Massively Open Online Courses (MOOCs)

where students are not able to receive direct and specific feedback from the instructor. In these larger scale environments, feedback is usually accomplished through auto-graded assessments, with an additional layer available in discussion forums. Regardless, good feedback requires the instructor to gather detailed information about how the student approaches a problem. An example of this includes situations where an instructor asks students to "show your work" in mathematics. By detailing the process that the student goes through to create a solution to a problem, the instructor can determine if students have any procedural bugs that may be preventing them from correctly solving the problem.

Researchers have begun looking at ways in which they can gather data created from learning environments in order to provide formative feedback to both instructors and students (Nyland, in review). Many of these learning environments collect as much student activity data as possible, process the data using data mining techniques, and present the feedback to the students in the form of dashboards or real-time recommendations. What is lacking from this research, however, is the ability to get detailed step-by-step data—which we refer to as transaction-level data (Chung, 2014)—about the student thought process as they complete learning assessments. This transaction-level data is equivalent to "showing your work", as the computer is able to see many of the intermediate steps that the student goes through in order to arrive at an answer. Through the use of this transaction-level data, the computer may be able to identify student misconceptions that might go undetected through traditional assessment methods (Davies, Nyland, Chapman, & Allen, 2015).

Data collection regarding student learning may be aided by the increased use of educational courseware. While the meaning of courseware has fluctuated over time, Feldstein

(2013) describes its current form as plug and play courses that are a combination of instructional content, a learning platform, and course design. In a traditional course, the students may be required to do several activities in order to prepare for a class, including reading their textbook, watching a video, or completing an assignment. The benefit of courseware is that most of these activities happen inside of the course, where detailed data can be collected. By understanding which course activities contribute to improved learning, the use of these data could help produce better feedback, either from the computer or from the instructor.

This paper is a case study in applying learning analytics to one of these courseware environments, an online spreadsheet course. The environment collects transaction-level assessment data, along with detailed data regarding reading and video usage. Our goal was to determine the extent to which a relationship exists between student activities and assessment performance. Because the course collects transaction-level data regarding student performance on assessments, we are better able to identify knowledge components with which the students are struggling. We hope that this research will pave the way for future research into providing feedback to students in these integrated learning environments.

Literature Review

Learning analytics and educational data mining are two relatively recent areas of research that have received increased attention in the last ten years. Learning analytics is described "as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Siemens, 2011, para. 2). Research in these fields has focused on different purposes and has used different types of educational data. In a recent literature review on learning analytics and educational data mining, Papamitsiou and Economides (2014) found that researchers had six

types of objectives: (a) student and student behavior modeling, (b) prediction of performance, (c) increase (self-) reflection and (self-) awareness, (d) prediction of dropout and retention, (e) improve assessment and feedback services, and (f) recommendation of resources. While these categories are not mutually exclusive, they are helpful in understanding the various goals of researchers within the field. As the goal of this study is to determine if LMS activity (reading content and watching videos) is linked to performance in authentic assessments, the purposes with which it most readily aligns is the prediction of performance and improvement of assessment and feedback services.

We will now review several areas of related literature with our study in four research areas: examining the links between LMS activities and overall performance, video analytics, reading analytics, and combined video and reading analytics.

Links between LMS Activities and Performance

Several studies have looked at the relationship between activities in an LMS and overall performance. Ramos and Yudko (2008) explored the relationship between discussion posts read, discussion posts posted, total page hits, and quiz scores. Using a stepwise regression, they found that only one of the variables—total page hits—predicted overall quiz score. In one of the classes, page hits explained 7 percent of the variance in overall quiz score—a relatively weak number, while in another class page hits explained 23 percent of the variance in quiz score. In the article, however, they do not discuss the alignment between the content of the page and the assessments. We feel that an understanding of alignment is critical and may change the relationship between the variables.

In another study, Agudo-Peregrina, Hernandez-Garcia, and Iglesias-Pradas (2012) identified those types of interactions and activities that were predictors of student success in the

class. Their regression found several significant predictors: First, those students who were classified as *active* (vs. passive) did better in the class. Second, students who engaged in evaluating activities—a misnomer word for activities that "have to do with completing and sending individual and group assignments, quizzes, questionnaires, or other similar tasks." (p. 3) —were more likely to perform better in the class. Finally, students who engaged in student-student and student-teacher interactions did better in the class. Importantly, student-content interactions were not a significant predictor of overall performance in the class. However, again there was no mention of the alignment of the content between course activities and overall assessment for the class.

Finally, Henrie, Bodily, Manwaring, and Graham (2015) used an intensive longitudinal approach to explore engagement in a blended learning context. Along with gathering engagement data using self-report surveys, they also gathered student behavioral data by means of LMS activity data. While their primary purpose was to examine the connection between LMS activity and engagement, they also analyzed the connection between LMS activity and performance in the process. Overall, they found that aggregated LMS activity (page views and time spent) was not a good indicator of performance in a course. To better predict engagement, the researchers categorized the type of pages in the LMS (as procedural, content, and social). They argued that the more successful students spent more time in the first weeks of class exploring the set-up of the course (such as the calendar, discussion boards, syllabus, and quizzes) and previewing future assignments. Overall, this research suggests that it is difficult to draw overall relationships between student behavior in the LMS and performance; instead, more information about the specific content of LMS pages is required.

Video Analytics

With the rise of the MOOC there have been several studies interested in looking at how students are using videos in these courses. Much of this research has been investigating what they call *clickstream* data—derived from video logs in the online courses. Aiken et al. (2014) looked at video use in a Georgia Tech introductory mechanics class. After analyzing the clickstream data for individual videos, they found a link between pausing activities on the videos and particularly salient information that was required for assessments—students were more likely to pause when there was information that they needed to study closely or use for an assignment. This finding is similar in a way to that of Avlonitis and Chorianopoulos (2014) who used pulse modeling to examine instructional video usage. They found that users tended to replay sections of the video that were information-rich and visually complex.

We could only find one study that linked clickstream data to class performance. Brinton and Chiang (2015) used data from two Information Technology MOOCs to determine if clickstream data predicted whether a student was correct on a first attempt (CFA) on an in-video assessment item (many MOOCs feature automatic pauses in the video where a machine-gradable assessment item is delivered to the student). Overall, they found that correct answers on an assessment were linked with spending more time with the video, watching a greater fraction of the video, pausing more (which the authors said was synonymous with self-reflection), and changing the playback rate more. Overall, this suggests that students who are more active with the video are more likely to do better on the assessments. While the data in the study reflects good content alignment between the assessment items and the content, the assessment items were most likely not authentic tasks—tasks that require more complex performance—as they needed

to be machine-gradable. Our goal is to examine the link between video content and authentic assessments.

Reading Analytics

While the aforementioned studies explored the connection between page views and class performance, there has been less research regarding the connection between specific reading material and performance. A few researchers have begun exploring the data that is available from online e-books. Nicholas, Rowlands, and Jamali (2010) looked at eBook use among students from 127 schools in the United Kingdom. From the project, they were able to collect several variables including: duration and time of use, book used, location of use, nature of use, and method of searching/navigating. With such a large-scale study, the researcher did not attempt to connect the data from eBook use to performance in individual classes. Rather it was exploratory in nature and the results were descriptive.

A more recent research bulletin by Horne, Russell, and Schuh (2015) wanted to "provide concrete strategies, grounded in research with e-textbook analytics, on how to use data from interactive platforms to inform decisions about supporting student learning with educational technology" (p. 1). They worked with the vendor of their eBook and were able to collect data regarding online and offline reading content and time as well as material provided by the markup features of the eBook—the number and text of highlighted sections, notes, questions, tags, and annotations. Additionally, as not much qualitative data was available from the eBooks, they augmented their study with surveys with the students and self-reported reading journals. They stated that one of the difficulties of using eBook analytics is that it is difficult to determine the quality of the reading experience – they needed to rely on the reading journals for that. Therefore, while this study began exploring the types of data that were available from eBooks,

there was no attempt to connect reading to class performance. This will be one of the goals of our study.

Combining Video and Text Analytics

A recent study Stice, Stice, and Albrecht (2015) combines both video and reading analytics in an attempt to see how these factors influence overall student performance in an introductory accounting course. The course uses a courseware-based environment that is similar to the one that is used in our study. In the course, data is collected regarding the proportion of paragraphs of text that the student read along with the proportion of videos watched. The text and videos exist side-by-side in the course and cover roughly the same content. The authors had two primary objectives: (a) determine whether usage analytics of the courseware environment predicted student success in the course and (b) determine how performance differed between students who preferred to watch videos as opposed to students who read the text. Overall, the results suggested that those students who spent more time using the courseware environment (measured as the combination of the text read and videos watched) had better overall performance on the course assessments. Additionally, they found that those students who preferred to watch videos performed significantly worse in the course than those students who preferred to read the material. While these results are interesting and a good indicator that we will be able to find good results in our study, it still only addressed usage on an overall course level. It is our goal to investigate the links between learning content interaction on the level of individual learning outcomes.

Research Purpose and Questions

The purpose of this study is to explore the links between LMS video watching and reading activities and performance on content specific authentic assessment items. Our previous

work (Davies, Nyland, Chapman, & Allen, 2015) demonstrated the ability of an LMS in collecting detailed information about a student's knowledge gaps and misconceptions regarding concepts in Microsoft Excel. In this study, we wanted to take the research further and explore the links between course activities (e.g., reading instructional material and watching instructional videos) to understand what effect they have on overall student performance and the presence of student errors. As such, our work was guided by the following research questions:

1. Is there a correlation between
    a. Reading usage and error rates?
    b. Video usage and error rates?
2. What is the unique predictive value of reading and video on errors in the presence of each other as measured by a multiple regression model?
3. What value do the reading and video resources have for self-remediation? That is, do reading and video usage decrease the number of errors once an error is identified?

Methods

Student activity data is captured by an Introduction to Excel class taught on the MyEducator platform. The data used for this analysis came from 10 sections of the course administered at two large western universities in the United States. In the course, students are given authentic assessments using Excel spreadsheets. These Excel spreadsheets are downloaded from the MyEducator platform and completed open-book without time constraints. Each spreadsheet contains a task guide that explains the problem to the student and guides them through the steps required to complete the overall task. A sample spreadsheet and task guide can be seen in Figure 1 below. These Excel spreadsheets contain a script that creates a detailed log of the process each student used to arrive at their solution to a given task. The log collects user

ID, assignment and worksheet ID, cell ID (e.g., D11), formula inputted, and the value displayed. This information is submitted into the MyEducator system when the student completes that assignment. The system then processes their grade and gives them feedback regarding their performance.

Separate from student assessment data, the MyEducator course also collects data regarding what the student has read (i.e., visits to a reading page, the number of paragraphs viewed in a section, and the time spent displaying that information) as well as what videos they've watched (i.e., video plays and total video play time). These data are available in a detailed form for each page and video in the course.

In our study, we examined the predictive relationship between student activity with learning materials on a particular topic and their performance on the corresponding assessment. Although there are many topics that we could have examined in the course, for the purpose of this case study, we looked at two specific skills. The first was absolute references (i.e., specifying that a row or column in a range should be held constant when copying a cell) and the second was the IF function (i.e., displaying the content of one of two variables based on a conditional statement). We measured error in each of these tasks on the first occasion that they came across it in the class as part of an assignment. For absolute references, this was in Lesson 2; for the IF function, this was in Lesson 3. Using a similar methodology to a previous study (Davies et al., 2015), we collected all of the transaction level data for each problem that we wanted to examine then coded each solution indicating the errors presented.

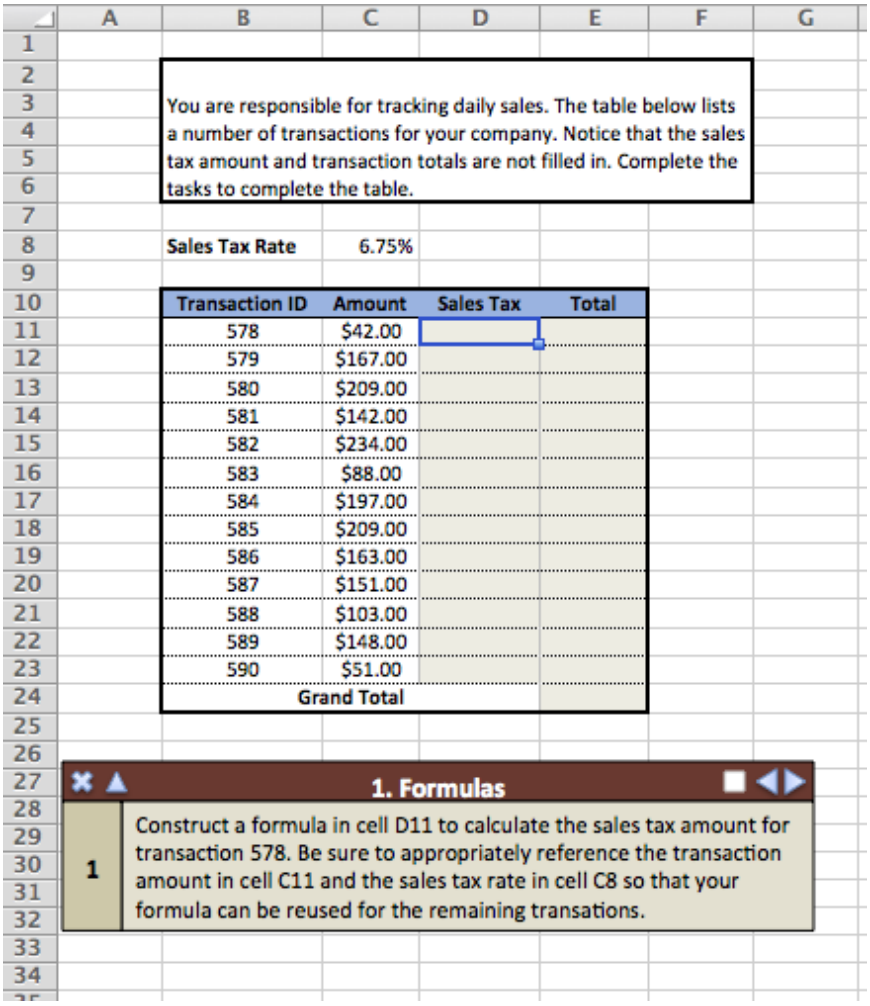

Figure 1. Sample task.

Error Coding

Cleaning and coding the data used the following procedure: First, all of the submitted answers for a specific problem were aggregated. Next, these answers were cleaned to remove blanks and unnecessary spaces. We then used a series of functions to automatically code the submitted answers for the presence of errors. In Lesson 2, we coded for evidence of four types of errors related to the use of absolute references. These included (in order of severity) (a) adding an extra absolute reference when it was not needed, (b) missing an absolute reference when it was needed, (c) using an absolute reference incorrectly (e.g., using it on the wrong cell in

the formula), and (d) typing in a value rather than a formula to avoid using an absolute reference. In a previous study (Nyland et al., in review), we found the functions that were used to automatically code submitted answers had a 92% agreement with a human coder, so we used the same functions on this set of data. Once the errors were coded, a level of error for each step was calculated using a weighted scale, along with a cumulative error level for each student. The scale was weighted according to the severity of the error. Because using an additional absolute reference on a cell is considered a minor error (i.e., it would not affect future uses of the formula) this minor issue was not factored into the error calculation. The severity weighting used for the absolute reference errors calculation are presented in Table 1.

Table 1

Identified Errors and Error Weightings for Use of Absolute References in Lesson 2

Error	Error Weighting	Example of Error
Optimal answer	0	=C11*C$8
Missing an absolute reference when it was needed	1	=C11*C8
Using an absolute reference incorrectly	2	=C$11*C8
Typing in a value to avoid using an absolute reference	3	=C11*.0675

In Lesson 3, we coded the data for evidence of errors when using the IF statement. There were three types of errors that we coded for in this instance: (a) Using the IF function with a correct condition, (b) using a valid *IF True* argument in the formula, and (c) using a valid *IF False* argument in the formula. We coded a portion of the sample manually and then compared it to the results of the function that was developed to automatically code the responses. Initial level of agreement between the human coder and the function was 94%; however, after

correcting human coding mistakes, the level of agreement was 98%. In Lesson 3, each of the errors was assigned an equal weighting as shown in Table 2.

Table 2

Identified Errors and Error Weightings for the IF Function in Lesson 3

Error	Error Weighting
Missing Function	1
Invalid IF True argument	1
Invalid IF False argument	1

Errors for lessons 2 and 3 were then aggregated at the step level for each student, and then combined across steps into a new synthesized variable: Cumulative error. Cumulative error acts as a proxy measurement for a student's misconceptions. Lesser levels of cumulative error may indicate that the student was able to solve the problem quickly, while greater levels of cumulative error indicate that the student struggled more with that concept.

LMS Data Collection

We also collected data for each student in the course regarding their interactions with learning materials. This data was limited to the sections of the course that were directly aligned with the knowledge components under analysis. Thus for absolute references, we collected data only on reading and video usage in the section of lesson 2 that taught specifically about absolute references, and the same for lesson 3. This data was collected from a SQL database using queries created in Microsoft Access. It should be noted that while we may say that a student "read" or "watched" a video, we have no way of knowing whether they actually paid attention to or comprehended the material. However, for the purpose of brevity, we will use the terms read and watch throughout the rest of the paper.

One consideration that we had to make when collecting reading time data was that we might get biased estimates of reading time. This is because reading time in the system is defined as the time difference from page load to page unload. When looking at the descriptive data for reading time, we saw that some pages were left open for days. To make estimates of reading time less biased, we followed the same method of another study (Kovanović et al., 2015) and capped the maximum reading time for a session to 10 minutes.

The final data that were collected from the system included:

- Percent paragraphs read –This is the total proportion of the paragraphs in the section that were on the screen for 5 seconds or longer
- Visits to reading pages – The total number of reading page loads that a student made
- Reading time – the sum of a student's individual reading sessions (capped at 10 minutes each)
- Percentage of video watched – This is the proportion of the video that the student played
- Number of Video Plays – The number of times the students pressed play on the video
- Video time – total time spent playing the video

The content interaction data was then combined with the transaction-level assessment data that was previously coded for error. Data was collected for every student who had completed the assignment. Overall, 3316 students completed lesson 2 and 3126 students completed lesson 3. These data were then used to answer our research questions.

Results

The relationship between LMS activity (i.e., video plays and text information views) and error rates on selected assessment items are shown in Tables 3 and 4. In lesson 2 there were two statistically (although weak) significant relationships with cumulative error: visits to reading pages, and percentage of the video watched. These relationships indicate that students who visited reading pages more often and who watched a greater percentage of the video had lower levels of cumulative error. In Lesson 3, the relationship between LMS activity and cumulative error was stronger, albeit in the opposite direction. Here students who used a greater portion of the materials in the LMS and who spent more time with them had greater cumulative error. One explanation for this might be that students who are struggling with the assessment were using the materials in the LMS more for help. This is difficult to determine from the correlation tables alone, but the next research questions seems to help us better understand this relationship.

Table 3

Correlation Table Between LMS Activity and Performance for Lesson 2

	CE	PPR	VRP	RT	PVW	NVP	TSWV
Cumulative Error	1						
Percent Paragraphs Read (PPR)	-.010	1					
Visits to reading pages (VRP)	-.046*	.509**	1				
Reading Time (RT)	.002	.555**	.670**	1			
Percentage of Video Watched (PVW)	-.105**	.436**	.364**	.533**	1		
Number of Video Plays (NVP)	-.016	.084**	.102**	.149**	.186**	1	
Time Spent Watching Video (TSWV)	-.017	.068**	.071**	.107**	.138**	.034*	1

N = 3316 * p < .05 ** p < .01

Table 4

Correlation Table Between LMS Activity and Performance for Lesson 3

	CE	PPR	VRP	RT	PVW	NVP	TSVW
Cumulative Error	1						
Percent Paragraphs Read (PPR)	.209**	1					
Visits to reading pages (VRP)	.207**	.620**	1				
Reading Time (RT)	.269**	.645**	.797**	1			
Percentage of Video Watched (PVW)	.123**	.431**	.466**	.590**	1		
Number of Video Plays (NVP)	.055**	.101**	.188**	.225**	.221**	1	
Time Spent Watching Video (TSWV)	.037*	.084**	.095**	.137**	.169**	.043*	1

N = 3126 *p < .05 **p < .01

Predictive Value of LMS Activity

We examined the predictive value of LMS activity on error rates for each of the lessons using a multiple regression. Whereas the correlation table helps us understand the relationship between two variables, regression helps us understand the unique predictive power of each variable in the presence of each other. After checking our assumptions for the data, we found that the distribution of our dependent variable, cumulative error, was non-normal and as such violated the assumptions for ordinary least squared regression (an example histogram for cumulative error in lesson 2 is shown in Figure 2). Instead, we treated cumulative error as a count variable with a non-normal distribution. To choose a model that most closely matched the data, we compared several model variations in MPlus (version 7.4)—a Poisson Regression, a Zero-inflated Poisson Regression, a Negative Binomial Regression, and a Zero-inflated Negative Binomial Regression. Negative Binomial Regressions are useful when there is over dispersion in the count variable; Zero-inflated Poisson Regressions are useful when there is an excessive amount of zeros in the count variable. The dispersion values for lessons 2 and 3 were 3.17 and 1.52 respectively, higher than the threshold of 1 for a Poisson distribution. Tables 5 and 6 show the comparison between the models using their Akaike information criteria (AIC), Bayesian information criteria (BIC), and sample adjusted Bayesian information criteria (SA-BIC) for lessons 2 and 3. Lower values indicate better model fit. For lesson 2, the negative binomial model fit the data the best; for lesson 3, the Zero-inflated Negative Binomial model fit the data the best. It should be noted as well that because the students were nested in courses, the data also violated the assumption of independence. To adjust for this, the data was clustered at the course section level in MPlus.

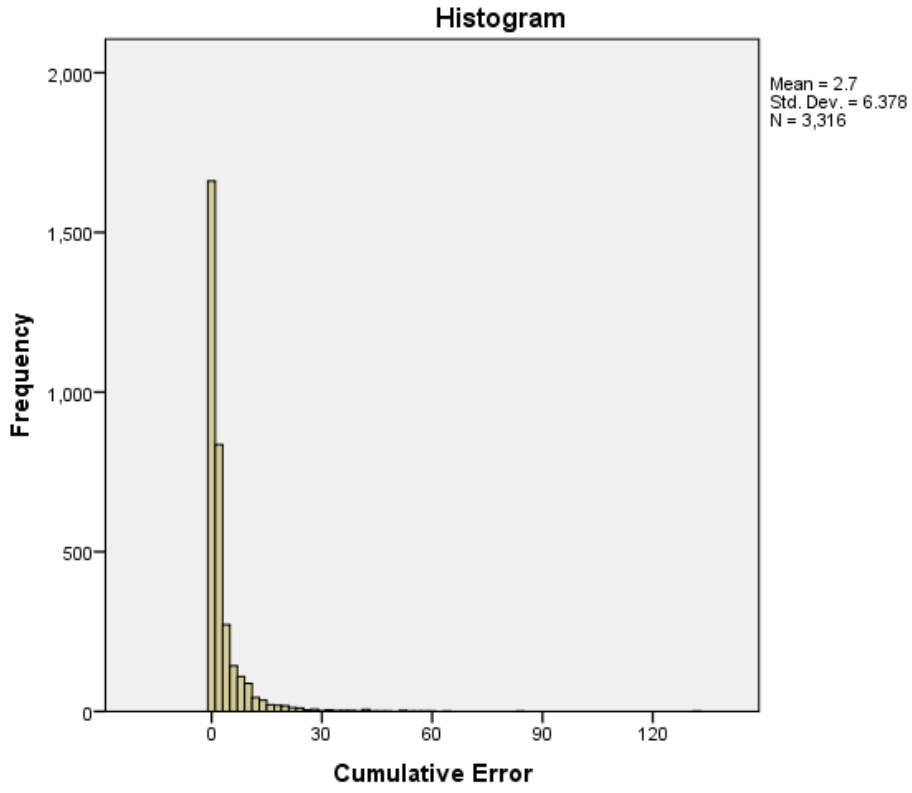

Figure 2. Histogram showing overdispersed nature of cumulative error data in Lesson 2.

Table 5

Comparison of Models for Lesson 2

	Poisson	Zero-Inflated Poisson	Negative Binomial	Zero-Inflated Negative Binomial
Akaike (AIC)	27684.43	20397.97	12711.47	12701.12
Bayesian (BIC)	27727.17	20483.46	12760.32	12792.72
SA-BIC	27704.93	20438.98	12734.90	12745.06

Table 6

Comparison of Models for Lesson 3

	Poisson	Zero-Inflated Poisson	Negative Binomial	Zero-Inflated Negative Binomial
Akaike (AIC)	54738.54	40209.04	17402.53	17227.99
Bayesian (BIC)	54780.87	40293.70	17450.91	17318.70
SA-BIC	54758.63	40249.22	17425.49	17271.04

The Zero-inflated Negative Binomial model runs two separate models, one in which the count variable is regressed on the predictor variables, and a separate logistic regression that models group membership in the zero group. This data had an excessive amount of zeros because many of the students got the answer right on their first attempt – thus receiving an error score of zero. The Negative Binomial model (without zero-inflation) only includes the first regression, not the logistic regression. The results of the analysis from the model for lesson 2 and then lesson 3 are discussed in the next two sub-sections.

Lesson 2. The Negative Binomial Regression results looks at the predictive value of each of the LMS activities on the overall value of cumulative error. The results of this regression for Lessons 2 are shown in Table 7. We've also calculated a McFadden's Pseudo R^2 value for each of the models (Hilbe, 2011) to show the amount of variation explained by the model. It should be noted overall, that the R^2 value for each of the models is low. Therefore, although we may gain certain insights from the results, they are explaining only a small portion of the variation in student error.

Table 7

Summary of Negative Binomial Regression for Lesson 2

Variable	B	SE (B)	β	*p*
Percent Paragraphs Read	.00	.00	.09	.486
Visits to Reading Page	-.05	.01	-.48	.000
Reading Time	.03	.00	.92	.000
Percentage of Videos Watched	-.01	.00	-1.09	.000
Number of Video Plays	.00	.00	-.09	.026
Time Spent Watching Video	.00	.00	-.08	.022

N= 3316
Note: β were produced using MPlus STDYX Standardization
McFadden's Pseudo R^2 = .01

Before proceeding, we should briefly discuss how to interpret a coefficient (B) in a negative binomial regression. The coefficients represent the predicted change in the log of the outcome variable (cumulative error) given that all other variables remain the same ("Stata Annotated Output Poisson Regression," n.d.). Using visits to reading pages (B = -.05, p = .003) as an example, for every increase in amount of visits to the reading pages, it is predicted that the log of cumulative error would decrease by .05. In addition to visits to reading pages, four other variables had significant predictive relationships with cumulative error in the presence of each other in lesson 2. Percentage of videos watched (B = -.01, p < .001), and number of video plays (B = -.00 , p = .034) had negative impacts on the cumulative error of students – meaning that the more students engaged in those types of activities, the less error they had. Reading time (B = .03, p < .001) was associated with increased amounts of cumulative error. Time spent watching video (B = .00, p = .022) was significant, but very small. This difference in results between reading time and visits to reading pages is interesting and may suggest that students who are more efficient with their information retrieval—making quick visits to find the pertinent

information to help them solve a problem—struggle less in the assessments. Using the standardized beta weights in Lesson 2 to compare predictors, it appears that watching the entire video had the greatest impact on the reduction of cumulative error. Those students who watched the entire video may have been more readily able to diagnose errors when they arose in the problem. On the other end, reading time was associated with the greatest increases of errors. Once again, the reason is difficult to determine from this data alone. It may be that students who had more errors spent more time in the reading pages trying to remedy those errors. Our next research question gives us further insight into this issue.

Lesson 3. In lesson 3, we used a Zero-inflated Negative Binomial Regression. As mentioned previously, this model runs two different models simultaneously: a Negative Binomial Regression, and a Logistic Regression. The results of the Negative Binomial Regression are shown in Table 8. There were only two variables that had significant predictive relationships with the amount of cumulative error. Those students who spent more time reading (B = .02, $p < .001$) and read a greater portion of the text (B = .01, $p < .001$) had increased amounts of cumulative error. In comparing the standardized beta weights, reading time once again had a greater impact on increasing levels of student error.

Table 8

Summary of Negative Binomial Regression for Lesson 3

Variable	B	SE (B)	β	*p*
Percent Paragraphs Read	.01	.00	.46	.000
Visits to Reading Page	-.02	.01	-.15	.236
Reading Time	.02	.00	.80	.000
Percentage of Videos Watched	.00	.00	-.07	.555
Number of Video Plays	.00	.01	.02	.968
Time Spent Watching Video	.00	.00	.02	.453

n= 3126
Note: β were produced using MPlus STDYX Standardization
McFadden's Pseudo R^2 = .03

In the second part of the Zero-inflated Negative Binomial Regression, a logistic regression is run to model membership in the zero group. When looking at these results, we need to keep in mind that the beta weight represents the likelihood that the student will be a part of the zero error group. For every unit increase of B, the odds that they would be in the zero groups would increase by a factor of exp(B) ("Stata Data Analysis Examples Zero-inflated Negative Binomial Regression," n.d.). So using the example of percent paragraphs read (B = -.01, $p < .001$) in lesson 3, this would mean that for every unit increase in the percentage of paragraphs read, their odds of being in the zero error group increase by exp(-.01) or .99.

For the logistic regression in Lesson 3 (Table 9), there were five significant predictors of having no error related to the use of IF statements. Students who read a greater portion of the paragraphs (B = -.01, $p < .001$), spent more time reading (B= -.06, $p < .001$), and played the videos more often (B= -.04, p =.005) were less likely to be in the zero group, while those that made more visits to the reading pages (B = .06, p = .004) and watched a greater portion of the video (B= .02, p < .001) were more likely to be in the zero error group. In examining the

standardized beta weights, number of video plays was the strongest indicator that students would make at least one error. This behavior may indicate that students are trying to correct their errors by replaying parts of the video, looking for the information that will help them. While we do not have specific evidence to suggest this, students might also be engaged in a just-in-time approach to learning (i.e., only looking at the instructional material when they encounter a problem). Since the assessment is open book, students may start the assessment and then visit the learning materials when it is needed. The results suggest, however, that watching the entire video first may be a better approach. As those students who watched a greater portion had less error overall.

Table 9

Summary of Logistic Regression for Lesson 3

Variable	B	SE(B)	β	*p*
Percent Paragraphs Read	-.01	.00	-.18	.000
Visits to Reading Page	.06	.02	.08	.004
Reading Time	-.06	.01	-.35	.000
Percentage of Videos Watched	.02	.00	.23	.000
Number of Video Plays	-.04	.01	-.44	.005
Time Spent Watching Video	-.01	.02	-.19	.427

n= 3126
Note: β were produced using MPlus STDYX Standardization
McFadden's Pseudo R^2 = .03

Remedial Value of LMS Activity

Previously we mentioned that it was difficult to understand the relationship between reading activity of students and the cumulative error on an assessment – are students using the materials more because they are having errors, or are they having more errors because they are not able to understand what the material is trying to teach? In our final research question, we

wanted to understand if the reading or video material helped students remediate errors once they had committed an error. To answer this, we took a subsample of students who had received an error score of at least 1 on the assessment. We flagged their first unsuccessful attempt to solve the problem, and then gathered all of the LMS data available for each student after that point. Because we were only looking at a subsample of a student's LMS activity, only reading time and video time were available as metrics. We also calculated any additional error that the student accumulated beyond the initial error. This cumulative error was then combined with LMS data after the error after committing their first error in a regression equation.

Once again, we examined the data to check assumptions for using a multiple regression. As it was in the previous analysis, we found that the data was extremely skewed and thus did not satisfy the assumptions for a typical linear regression. The data was loaded into MPlus, and four different models were applied (Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial regressions). Tables 10 and 11 show the model comparisons for Lessons 2 and 3. After comparing the AIC, BIC, and SA-BIC of the three models, we found that the negative binomial model was the most appropriate for lesson 2 while the zero-inflated negative binomial model was the most appropriate for lesson 3. We will now examine results from the regressions for Lessons 2 and 3 for lesson material usage after a student's unsuccessful attempt to solve the problem.

Table 10

Comparison of Models for Lesson 2 After an Error was Made

	Poisson	Zero-Inflated Poisson	Negative Binomial	Zero-Inflated Negative Binomial
Akaike (AIC)	17435.40	13375.65	7774.03	7778.57
Bayesian (BIC)	17451.64	13408.13	7795.68	7816.46
SABIC	17441.11	13389.07	7782.98	7794.22

Table 11

Comparison of Models for Lesson 3 After an Error was Made

	Poisson	Zero-Inflated Poisson	Negative Binomial	Zero-Inflated Negative Binomial
Akaike (AIC)	42661.35	31418.35	11956.25	11866.39
Bayesian (BIC)	42678.10	31451.84	11978.58	11905.46
SA-BIC	42668.57	31432.78	11965.87	11883.23

Lesson 2. The negative binomial regression shows the predictive relationship of variables with the overall count error—in this case the additional errors that a student makes after their initial error. Table 12 shows the result of this for lesson 2

Table 12

Summary of Negative Binomial Regression for Lesson 2 After Error was Made

Variable	B	SE (B)	β	*p*
Reading Time	.01	.00	1.03	.000
Time Spent Watching Video	-.00	.00	-.34	.000

n = 1657
Note: β were produced using MPlus STDYX Standardization, McFadden's Pseudo R^2 = .00

The results from lesson 2 suggest that time spent reading (B =.01, $p < .001$) was associated with a greater amount of error. This agrees with the overall data from the entire

sample of students, that there is a relationship between spending time on a reading page and struggling with the concepts that the assessments are covering. While we cannot specifically determine whether the content of the reading pages caused increased error on the part of the students, it did not seem to help them to resolve their problems quickly.

Lesson 3. In Table 13, we can see that Lesson 3 results are similar. Here reading time (B = .03, p <.001) was associated with increasing amounts of error after the initial error has been made. Again, we cannot determine whether the clarity of the materials is contributing to continual struggle by the students, but it appears that the use of the learning materials in this lesson is not helping students self-remediate.

Table 13

Summary of Negative Binomial Regression for Lesson 3 After Error was Made

Variable	B	SE (B)	β	p
Reading Time	.03	.00	.98	.000
Time Spent Watching Video	.00	.00	.10	.496

N=1964

Note: β were produced using MPlus STDYX Standardization, McFadden's Pseudo R^2 = .02

Table 14 displays the results of the logistic regression that models group membership in the zero group. Here we see that students who spent additional time reading after they committed an error were less likely to be in the group that had no additional error (B = -.06, p > .001). Additionally, students who spent more time watching the video after their initial error, were also less likely to be members of the zero group (B= -.01, p = .034).

Table 14

Summary of Logistic Regression for Lesson 3 After Error was Made

Variable	B	SE (B)	β	*p*
Reading Time	-.06	.01	-.37	.000
Time Spent Watching Video	-.01	.01	-.16	.034

N=1964
Note: β were produced using MPlus STDYX Standardization, McFadden's Pseudo R^2 = .02

Discussion

The goal of this research was to examine the predictive value of LMS activity data on transaction-level assessment data. Overall, we found that there were varying levels of correlation between LMS activities and the assessment data. The correlation levels on both occasions were quite small (less than .3). This likely indicates that there are other factors that have a greater impact on student performance (e.g., previous knowledge or experience). In our study, we found much higher correlations for lesson 3 than lesson 2. This may be because students have differing levels of familiarity with course topics and thus using the course materials is more important in some instances than others. For example, in lesson 2 students may have felt that they understood cell references—possibly because it was an introductory topic—and thus only briefly skimmed over this material before attempting the assessment. The unfamiliarity of the content of lesson 3, on the other hand, may have compelled students to spend more time with the learning materials.

When we looked at the relative predictive value of the different LMS activities on the transaction-level errors made in the assessments we found some statistically significant results. We should note that these findings are hampered by one large caveat—that in all of the cases the LMS data is only explaining less than three percent of the variance in cumulative error. Putting

this limitation aside for the moment, we do think that we found some intriguing findings in our results.

In lesson 2, we found that two of the independent variables had a strong association with cumulative error: reading time and percentage of videos watched. Those students who spent more time on reading pages had increased amounts of cumulative error, while those students who watched a greater portion of the video had less error. While the exact nature of the statistical relationship between reading time and increased error is unclear, we suppose that students spent more time viewing pages in an attempt to remediate their unsuccessful attempts to solve the assignment task. As they answered the problems incorrectly, they turned to the learning resources for remediation. This conclusion was supported by examining the LMS activity of students after they had attempted to solve a problem unsuccessfully. However, it is also possible that the readings did not immediately provide the help they need. That the students left reading pages open, looking for material that was going to help them, but still continued to struggle. These results may be an indication that the reading material might be revised to make it easier for students to find the information needed to remediate their problem. Additionally, the course could offer more opportunities for students to practice individual skills before they get to the summative assessment. This may minimize the amount of referencing a student needs to do to the learning material during the task, as they would be more confident of their own skills.

Our research also identified another factor that led to increased student success on assessments: watching a greater portion of the instructional video. Overall, we have a sense that students used a just-in-time approach to learning. They visited learning resources when they encountered a problem on their assessment. Our findings suggest that it may be a good practice for the student to watch the entire video before moving onto the assessment. The videos in the

course contain walk-through solutions to problems that are very similar to those found in the assessments and thus may promote better transfer when the student is presented with a problem on the final exam. If this trend continues through other instances in the course, this might be valuable feedback that could be generated by the computer to the student.

In our final research question, we looked more specifically at remedial actions on the part of the students – does their LMS activity after they have committed an error affect their level of error? We found a similar story across lessons 2 and 3: That reading time after the error occurred did not lead to decreases in cumulative error. If anything, increased reading time only made the problem worse. This also suggests that reading materials may not be organized in such a way so that students can quickly get remedial help for their task – or that the student is not well equipped to extract the necessary material needed to fix their problem. The students' use of video after receiving an error had no perceptible effect.

Conclusions

The purpose of this study was to act as a case study in using these types of data in a courseware system to see what kind of insights they might provide. Here we combined detailed analytics regarding the materials that a student read and videos that they watched with aligned transaction-level assessment data. First, we found that there was a slight, but varying correlation between reading and video activities in an LMS, and the overall error that a student had on a performance assessment. These levels may vary according to the students' familiarity with the material, and the difficulty of the assessment. Second, we examined the predictive relationship between reading and video activities and overall error, in the presence of each other. While the regression models explained three percent of the variance or less, we found some interesting relationships. Mainly that time on reading pages led to predicted increases in error levels, and

watching a greater portion of the video led to predicted decreases in error level. While they would still need to be validated by additional occasions, these results have implications for both the design of the course and possible feedback generated from the system.

In our last research question, we examined the remedial value of the learning resources. We found that once a student had committed an error on an assessment, additional time with the learning resources did not contribute a decreasing amount of error. This indicates that students were most likely using other resources to self-remediate.

Limitations

We should mention that our study has several limitations, the first being that the LMS metrics are imperfect. When a student is on a page, we have no way of knowing whether or not they are actually reading the content; when they are watching a video, we don't know if they are watching intently or if they are trying to multitask with doing something else. As such, we do not doubt that there is a lot of measurement error in this data (though we have tried to mitigate it as much as possible). As metrics get more sophisticated in the future, we should have a better sense of students' time on task.

The next limitation is the specificity of the data—it is very unlikely that other forms of educational data will be able to be captured in exactly the same way that it was in the Intro to Excel course. Nonetheless, we still feel that online performance assessments and simulations will become increasingly common in education, and these forms of technology have the ability to captured detailed transaction-level data that might help us better understand student learning processes and misconceptions. We hope that the processes that we have followed in the study will provide ideas regarding how other researchers might handle data in their own projects.

Areas for Future Research

This paper was a first attempt to link LMS data with detailed assessment data in our course. Based on the results, we feel that it provides many opportunities for future research. One of the things that we would like to do in the future is gather contextual data from the students regarding how they remediate their problems. This study only used LMS data, but clearly, from the amount of variance that was explained, there are other things that are happening outside of the system which determine their amount of error on the problem. In a future study, we would like to collect survey data regarding previous knowledge, use of outside resource, use of classmates, etc.

This study has also created many questions about how students approach their assessments in the course. We have a sense that many students are using the learning materials in a just-in-time fashion; however, additional analyses on the data are needed to confirm and explore this.

Another area of research is looking at this relationship across more occasions in the course. This research has covered roughly only two of hundreds of occasions for assessment. We would like to examine if this relationship changes based on the type of problem and knowledge components covered.

References

Agudo-Peregrina, A. F., Hernandez-Garcia, A., & Iglesias-Pradas, S. (2012). Predicting academic performance with learning analytics in virtual learning environments A comparative study of three interaction classifications. Proceedings of *Computers in Education (SIIE), 2012 International Symposium* (pp. 1–6).

Aiken, J. M., Lin, S., Douglas, S. S., Greco, E. F., Brian, D., Caballero, M. D., & Schatz, M. F. (2014). Student use of a single lecture video in a flipped introductory mechanics course. In *Physics Education Research Conference Proceedings 2014* (pp. 19-22).

Avlonitis, M., & Chorianopoulos, K. (2014). Video pulses: User-based modeling of interesting video segments. *Advances in Multimedia*, *2014*. doi:10.1155/2014/712589

Brinton, C. G., & Chiang, M. (2015). MOOC performance prediction via Clickstream Data and Social Learning Networks. In *IEEE Conference on Computer Communications (INFOCOM), 2015*.

Chung, G. (2014). Toward the relational management of educational measurement data. *Teachers College Record*, *116*(11), 1-16. Retrieved from http://www.gordoncommission.org/rsc/pdfs/chung_toward_relational_management.pdf

Davies, R., Nyland, R., Chapman, J., & Allen, G. (2015). Using transaction-level data to diagnose knowledge gaps and misconceptions. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (pp. 113–117). doi:10.1145/2723576.2723620

Feldstein, M. (2013). MOOCs, courseware, and the course as an artifact. Retrieved from http://mfeldstein.com/moocs-courseware-and-the-course-as-an-artifact/

Hattie, J. (1999). Influences on student learning [pdf]. Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.8465&rep=rep1&type=pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Henrie, C. R., Bodily, R., Manwaring, K. C., & Graham, C. R. (2015). Exploring Intensive Longitudinal Measures of Student Engagement in Blended Learning. *International Review of Research in Open and Distributed Learning*, *16*(3), 131–155.

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge: Cambridge University Press.

Horne, S. Van, Russell, J., & Schuh, K. L. (2015). *Assessment with E-Textbook analytics working with vendors and obtaining analytics*. Retreived from https://library.educause.edu/~/media/files/library/2015/2/erb1501-pdf.pdf

Kim, K. J., Liu, S., & Bonk, C. J. (2005). Online MBA students' perceptions of online learning: Benefits, challenges, and suggestions. *Internet and Higher Education*, *8*(4 SPEC. ISS.), 335–344. doi:10.1016/j.iheduc.2005.09.005

Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R. S., & Hatala, M. (2015). Penetrating the black box of time-on-task estimation. Proceedings of the *Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (pp. 184–193). doi:10.1145/2723576.2723623

Nicholas, D., Rowlands, I., & Jamali, H. R. (2010). E-textbook use, information seeking behaviour and its impact: Case study business and management. *Journal of Information Science*, *36*(2), 263–280. doi:10.1177/0165551510363660

Nyland, R. (Manuscript *in Review) . R*eview of data-enabled formative assessment.

Nyland, R., Davies, G., Chapman, J., & Allen, G. (Manuscript in Review). Transaction-level Learning Analytics in online authentic assessments.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology and Society, 17(4), 49–64.

Ramos, C., & Yudko, E. (2008). "Hits" (not "discussion posts") predict student success in online courses: A double cross-validation study. *Computers and Education*, *50*(4), 1174–1182. doi:10.1016/j.compedu.2006.11.003

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. doi:10.3102/0034654307313795

Siemens, G. (2011). Learning and academic analytics. Retrieved from http://www.learninganalytics.net/?p=131

Stata annotated output Poisson Regression. (n.d.). Retrieved from http://www.ats.ucla.edu/stat/stata/output/stata_poisson_output.htm

Stata data analysis examples Zero-inflated Negative Binomial Regression. (n.d.). Retrieved from http://www.ats.ucla.edu/stat/stata/dae/zinb.htm

Stice, J. D., Stice, E. K., & Albrecht, C. (Manuscript in Review). Do introductory accounting students choose to study by reading text or by watching video lectures?, 1–32.

DISSERTATION CONCLUSION

The aim of this dissertation was to explore the benefits of using detailed educational data to provide feedback to instructors as students as they engage in the learning process. In the first article, we reviewed literature regarding existing systems that use educational data to provide feedback to students and instructors. Overall, we found that research is making strides toward collecting richer data from learning interactions. While some systems have only collected machine gradable assessment data (Buchanan, 1998; Lin & Lai, 2014; Wang, 2008), other systems are collecting richer student activity data (such as content visited, time spent on activities, and system generated feedback). By capturing this richer data, researchers were able to use data mining methods to process the data and provide recommendations to instructors and students, based on the patterns uncovered. The critique that we had of these systems however, is that many provided feedback that was either not understandable to the user (such as Chen and Chen (2009), who presented association rules directly to 9 to 11 year old students) or did not provide them with information that would lead to a direct action. We feel that more exploratory work needs to be done to understand what types of information would be helpful to motivating teachers and students to accomplish their goals.

In the second article, we explored a specific learning system that collects rich educational regarding a student's problem solving process in Microsoft Excel. Our goal was to test the benefits of using detailed transaction-level data when compared to final answer data alone. We examined student misconceptions related to the use of absolute references, a foundational skill when learning about spreadsheets. Overall, we found that the transaction-level data was better at detecting persistent student errors in the problem solving process. By using the transaction-level data across time in the course, we were able to see that students were continually struggling with

the proper use of absolute references, when ideally those problems should resolve themselves across time. By identifying these instances of student struggle, the system could provide feedback that is more specific to students about their performance.

In the final article, we explored the relationship between persistent errors in the transaction-level data, and other activity in the course. Namely, we wanted to understand if there was a predictive relationship between reading learning materials, watching instructional videos, and performance on assessments. We discovered that while a relationship between these factors could be identified, the amount of variance explained by the LMS activity data was very small (three percent or less). This suggests that there are other factors that may better predict student performance. Despite the fact that our model did not explain a large amount of the variance, we did gain some interesting results from the results. Namely, that there was a positive predictive relationship between the amount of time spent on reading pages within the course and the amount of cumulative error that the student committed. This suggests that there is something about the reading process or study processes of students who keep their reading pages open for longer that contributes to an increased amount of error. We also found a negative relationship between the percentage of the learning video watch and cumulative error. This suggests that the more that a student watches the entire video; the less likely they are to commit errors on their assessments.

Based on our results, what is the future of using educational data to provide better feedback to students regarding their performance? As reviewed in article 1, we are sure to see educational systems which will collect an increasing amount of data about student activities in the course—we may even see some that collect data on a level of detail that is comparable to what we have seen in the Introduction to Excel class. Moreover, while these data might be better

equipped to detect student's misconceptions with concepts, we still need to understand the cases in which we should provide feedback to the student to motivate them to action. In article 3, we wanted to understand if we could provide feedback based on patterns that occur in the course. While we found some relationships between the LMS data and performance on the assessments that might lead us to make recommendations to the students (such as "watch the entire video before attempting the assessment"), the small amount of variance explained by the LMS data, may lead us to ask questions about student behaviors that are happening outside of the system.

Therefore, we leave this dissertation with a tempered view of the future of learning analytics. While data may help us see small patterns that can improve student performance, they may not be the panacea that people think they are. Instead, we need to take a more holistic approach to improving education.

DISSERTATION REFERENCES

Buchanan, T. (1998). Using the World Wide Web for Formative Assessment. *Journal of Educational Technology Systems*, *27*(1), 71–79. doi:10.1016/s0360-1315(00)00049-x

Chen, C. M., & Chen, M. C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education*, *52*(1), 256–273. Retrieved from https://www.lib.byu.edu/cgi-bin/remoteauth.pl?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2008-17099-026&site=ehost-live&scope=site

Hattie, J. (1999). Influences on student learning [pdf]. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.8465&rep=rep1&type=pdf

Lin, J. W., & Lai, Y. C. (2014). Using collaborative annotating and data mining on formative assessments to enhance learning efficiency. *Computer Applications in Engineering Education*, *22*(2), 364–374. doi:10.1002/cae.20561

Siemens, G. (2011). Learning and Academic Analytics. Retrieved from http://www.learninganalytics.net/?p=131

Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The Future of Data-Enriched Assessment. *Research & Practice in Assessment*, *9*, 5–16.

Wang, T.-H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education*, *51*(3), 1247–1263. doi:10.1016/j.compedu.2007.11.011