Multimodal Learning Analytics

Paulo Blikstein
Stanford University Graduate School of Education and (by courtesy) Computer Science
520 Galvez Mall, CERAS 232 Stanford, CA – 94305 – USA
paulob@stanford.edu

ABSTRACT

New high-frequency data collection technologies and machine learning analysis techniques could offer new insights into learning, especially in tasks in which students have ample space to generate unique, personalized artifacts, such as a computer program, a robot, or a solution to an engineering challenge. To date most of the work on learning analytics and educational data mining has focused on online courses or cognitive tutors, in which the tasks are more structured and the entirety of interaction happens in front of a computer. In this paper, I argue that multimodal learning analytics could offer new insights into students' learning trajectories, and present several examples of this work and its educational application.

Categories and Subject Descriptors

D.3.3 [**Programming Languages**]

General Terms

Algorithms, Measurement, Human Factors.

Keywords

learning analytics, multimodal interaction, constructivism, constructionism, assessment.

1. INTRODUCTION

New high-frequency data collection technologies and machine learning analysis techniques could offer new insights into learning in tasks in which students have ample space to generate unique, personalized artifacts, such as a computer program, a robot, a movie, an animation, or a solution to an engineering challenge. To date most of the work on learning analytics and educational data mining has focused on online courses or cognitive tutors, in which the tasks are more structured and scripted, and the entirety of interaction happens in front of a computer. In this paper, I argue that multimodal data collection and analysis techniques ("multimodal learning analytics") could bring novel methods to understand what happens when students can generate unique solution paths to problems, interact with peers, and act in both the digital and the physical worlds.

Assessment and feedback is particularly difficult within those tasks, and has hampered many attempts to make them more prevalent in schools. Therefore, these automated approaches would be particularly useful in a time when the need for scalable project-based, interest-driven learning and student-centered pedagogies is growing considerably [e.g., 5]. Both K-12 and engineering education [10, 11], within a transformed societal and economic environment, now demand higher level, complex problem-solving rather than performance in routine cognitive tasks [15]. These approaches have been advocated for decades [9, 12, 16, 18] but failed to become scalable and prevalent, and came under attack during the last decade [e.g., 13, 14]. Automated, fine-grained data collection and analysis could help resolve this tension in two ways. First, they could give researchers tools to examine student-centered learning in unprecedented scale and detail. Second, these techniques could improve the scalability of these pedagogies since they make both assessment and formative feedback, which are more complex and laborious in such environments, feasible. They might not only reveal students' trajectories throughout a learning activity, but they also would help researchers design better scaffolds.

At the same time, in the well-established field of *multimodal interaction*, new data collection and sensing technologies are making it possible to capture massive amounts of data in all fields of human activity. These techniques include logs of computer activities, wearable cameras, wearable sensors, biosensors (e.g., skin conductivity, heartbeat, and EEG), gesture sensing, infrared imaging, and eye tracking. Such techniques are enabling researchers to have an unprecedented insight into the minute-by-minute development of several activities, especially when they involve multiple dimensions of interaction and social interaction. However, these techniques are still not popular in the field of learning analytics. In this paper, I propose that *multimodal learning analytics* could coalesce these multiple techniques in order to evaluate complex cognitive abilities, especially in environments where the process or the outcome are unscripted.

In addition to their "unimodality," traditional assessment privileges product over process, and are divorced from actual learning activities. In computer science courses, for example, it is common to have students write pseudo code on a paper exam for assessing their programming proficiency. The proposed work looks to advance the field's capability to understand and utilize forms of assessment that are more closely tied to the actual practice of the respective disciplines. I want to study patterns in how students of different ages and expertise levels complete tasks such as programming, building a robot, designing a device, or conducting a scientific investigation. This work is informed by many of the current lines of research within educational data mining and learning analytics. For example, Rus et al. [20] makes extensive use of text analytics within a computer-based application for science learning, using expert-generated answers as a baseline. Beck and Sison [6] have demonstrated a method to assess reading proficiency combining speech recognition and probabilistic monitoring. D'Mello et al. [8] designed an application that could use spoken dialogue to recognize the states

LAK '13 April 08 - 12 2013, Leuven, Belgium
Copyright 2013 ACM 978-1-4503-1785-6/13/04…$15.00.

of boredom, frustration, flow, and confusion. Other researchers [1, 3] have been using machine learning in many contexts to measure students' learning and affect, in particular, focusing on their trajectory within a learning environment [2]. The goal of this paper is to be a proof of concept of novel assessment techniques along several modes. I will describe examples of work in multimodal learning analytics, in three different areas: analysis of students learning to program, learning to build mechanical structures, and inventing machines.

2. STUDENTS' PROGRAMMING

2.1 Programming modes

In this first example, I focus on students learning to program a computer using the NetLogo language. Hundreds of snapshots for each student were captured and analyzed. I will briefly describe the results of the study [4] and some prototypical trajectories, and discuss how they relate to students' programming experience. Nine students in a sophomore-level engineering class had a three-week programming assignment. Students had different levels of programming expertise at the beginning of the activity. The task was to write a computer program to model a scientific phenomenon in materials science. Every time student ran, saved, or compiled their code, a snapshot of their code was stored. The analysis consisted in counting the amount and pattern of changes (additions/deletions) in the code, by calculating the number of characters and lines of code changed during the three-week assignment, as well as the frequency of compilation. At every inflection point in the character-count curve, I visually examined the data to examine the snapshots before and after the point and inspect the types of changes that were being made to the code. The data analyzed suggested three coding profiles:

- "Copy and pasters:" characterized by a step-shaped growth curve, alternating *plateaus* with few code changes (looking for code / adapting the code), and rapid growth in code size (pasting the external code) – mostly observed in novice coders.
- "Self-sufficients:" characterized by a linear curve of steady increase in code size and almost no use of external code – mostly observed in more expert programmers.

I conducted a second analysis on the frequency of code compilation. Instead of expert/novice differences, the data showed that compilation frequency could be used as a proxy for how 'stuck' students were in the assignment. Despite the small sample size of this study, the analysis of open-ended programming projects enabled the identification of patterns with real-world significance and important implications for design. Each coding strategy and profile might demand different support strategies: advanced students ("self-sufficients") might require detailed and easy-to-find language documentation, whereas "copy and pasters" need more working examples with transportable code. In fact, it could be that expert programmers find it enjoyable to figure the solutions out themselves, and would dislike to be helped when problem solving. Novices might welcome some help, since they exhibited a much more active help-seeking behavior.

2.2 Programming styles

In order to explore these questions further, we extended the data collection to a much larger population. The second example is about a fully automated system to capture snapshots of students' code during large-scale programming assignments, and the use of several computational techniques to understand their progression and learning throughout an introductory undergraduate course in computer science. Again, seemingly erratic and highly personal trajectories of students generating a computer program in fact contain identifiable patterns, states, and transition probabilities, and can be successfully utilized to both understand the programming process, predict future performance, and potentially provide real-time assessment to instructors (about their effectiveness) and students (about their learning). Additionally, when this data is combined with student specific help seeking logs, we can construct a complete picture of the learner in terms of achievement, motivation and previous experience.

We researched the creation of computer programs using 10,000 code snapshots from 74 students enrolled in an introductory undergraduate programming course, using two methods for the discovery of patterns in the data. The first method was a greatly enhanced version of the pattern detection used in the previous study, using machine-learning techniques such as cluster analysis and dynamic time warping. All code updates were characterized based on their size and type, by identifying the number of lines and characters added, removed and modified between successive snapshots. This data was then used to construct a sequence of multidimensional vectors. This process was repeated for each assignment that the student completed. Using this sequence of vectors, the progress of each student was compared over the course of the class, by comparing their sequence of updates on a given assignment to their sequence of updates on the previous assignments. Students were then clustered based on the similarity of their update sequences. Six clusters emerged from the data.

Figure 1 shows the different programming processes that students follow, and it resulted in different outcomes on the midterm and final examination. Of particular interest is the contrast between cluster five and its peers. Figure 2 goes deeper into the differences in examination grades, as students in certain clusters initiated help requests at very different frequencies. Moreover, it is possible to observe how help seeking changes over time for different groups.

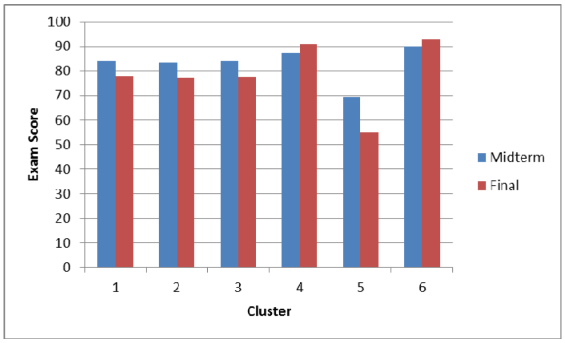

Figure 1 – Avg. Midterm and Examination Score by Cluster

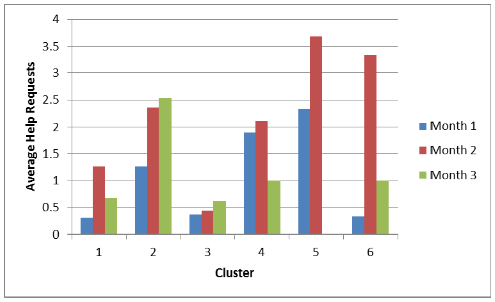

Figure 2 – Average Help Requests by Cluster and by Month

In further analysis, instead of simply clustering categories of code updates, the code snapshots were mapped into a state machine, using Hidden Markov techniques. The results of this work confirmed that indeed it was possible to cluster students in terms of robust behaviors that are dependent on expertise, and that those behaviors were correlated with their grades in the assignments and exams [19]. We will further discuss the significance of this data collection model in the conclusions.

3. TEXT MINING

3.1 Construction competence

A second important area for MMLA application is text mining. A variety of features can be extracted from text or transcripts, some of which I describe below. *Prosodic analysis* uses the pitch, intensity and duration of speech to infer intentionality and emotion. *Linguistic analysis* looks at several features such as pauses, filled pauses, and restarts, and can infer elements such as certainty. *Sentiment analysis* [17], using databases of terms such as the Linguistic Inquiry and Word Count (LIWC) and the Harvard Inquirer, infers sentiment from words or groups of words found in text. *Content word analysis* determines the knowledge contained in the text by using web-mined lexicons from chemistry, mathematics, computer science, material science and general science. Finally, *dependency parsing* and *n-gram* analysis can be used to inspect latent structures or meaning in the text by looking at group of words, sentences, and their relationship [7].

Many of these techniques were used in a study about engineering expertise [21]. Data for this study came from interviews of approximately 30 minutes with 15 students from a tier-1 research university (8 female, 7 male). Participants were asked to draw and think aloud about how to build various electronic and mechanical devices. Graduate and undergraduate students transcribed student speech. Prior to the interviews, the subjects were labeled as being experts, intermediates or novices in engineering and robotics, based on their major and previous experience. The data was analyzed using expectation maximization (EM), with an intra-cluster Euclidean distance objective function. Before running EM, the data was normalized and t-tests were performed to check statistical significance. The goal of the study was to try to predict the expertise level of the participants based on a combination of the extracted features.

Of particular interest to this study is the presence of several features that accurately predicted expertise based on *certainty*. Though not presented here at length, analysis of this data involved extracting n-grams from each transcript and looking for patterns across the different classes of participants. Not surprisingly, n-grams that indicated uncertainty, e.g. “don’t know” and “well, you know”, were more common among novices than among non-novices. These initial results confirm the work of Beck [6], which indicates that increasing expertise tends to increase student self-confidence. These results were further corroborated through sentiment analysis: certainty terms was much more common among experts, while understatements terms were more frequently employed by novices.

Additionally, novices exhibited a much higher frequency of disfluencies (average of 1.06 per normalized time interval) than experts (0.5), and experts had much longer and more frequent pauses. Taken together with the field observations, this suggests that novices were uncomfortable with moments of uncertainty, and thus “filled the silence” with disfluencies. However, experts would just stay in silence thinking about the problem and articulating their next utterance. This finding is yet another example of how multimodal data can potentially categorize students based on seemingly small differences in discourse and behavior. Changes in these elements could indicate learning or more familiarity with the field being explored.

A second study using text mining was designed to investigate students’ identity as engineers and scientists, and their level of identification with these professions [22]. Students participated in a weeklong robotics workshop and, subsequently, in one-on-one interviews with a member of the teaching staff. These semi-clinical interviews consisted in think-alouds of students designing different inventions and devices while drawing them on paper. For example, one of the questions was to design a piggy bank that automatically counts money as it is dropped in. All interviews were transcribed and analyzed using standard text mining techniques such as n-grams analysis. The interest here was to examine if students who did well in the robotics workshop would employ different vocabulary and linguistic structures. Indeed, the data showed that when students were describing a project they were proud of (as measured by the facilitators’ field notes), they would use “I” or “my,” in sentences such as, “my original design was…” When the same student was describing a project that they were not proud of, they would instead use “it,” in sentences such as, “it’d be programmed to turn on…” in place of “I programmed it to turn on.”

Counting word frequencies, we were able to determine that, for the group as a whole, students identified as high achievers in the workshop used “I” more than seven times more frequently than low-achievers. Again, here we can observe how even simple automated n-gram counting within transcripts can reveal meaningful elements of students’ affect.

4. OBJECT/BODY TRACKING

4.1 Tracking actions using video

Another example of the use of multimodal techniques is video and gesture tracking. As an example, I will describe a study on students’ ability in building simple structures [23]. Data was drawn from thirteen participants, each given everyday materials and asked to build a tower that could hold a mass of approximately 3 lbs. Participants were also challenged to make the structure as tall as possible. The task was designed to see how well students are able to take their intuitions about mechanical engineering and physics and translate them into a stable structure. Students were given four drinking straws, five wooden popsicle sticks, a roll of masking tape and a paper plate; and were told that they had approximately ten minutes to complete the activity. However, they were permitted to work for as long as they wanted with participation times ranging from six to 52 minutes. Of the twelve students, three were mechanical engineering graduate students and nine were high school students. Prior to the study students were classified based on their perceived level of expertise in the domain of engineering design. Three participants were labeled as experts, two as being of high expertise, five were classified as medium, and three as low.

A coding scheme was developed consisting of eleven object manipulation codes, identified through open coding of a sample of the videos, and agreed upon by a team of research assistants. This set was later collapsed into five codes, based on further coding sessions and discussions amongst coders (Table 1). The codes are entirely based on participant object manipulation, or lack thereof, and are not an attempt to interpret a student's intentions explicitly.

Table 1. Final codes for object manipulation

Class	Codes
BUILD	Building and Breaking
PLAN	Prototyping mechanism, Thinking with or without an object, Single object examination, Organizing and Selecting materials
TEST	Testing a mechanism and system testing
ADJUST	Adjusting
UNDO	Undoing

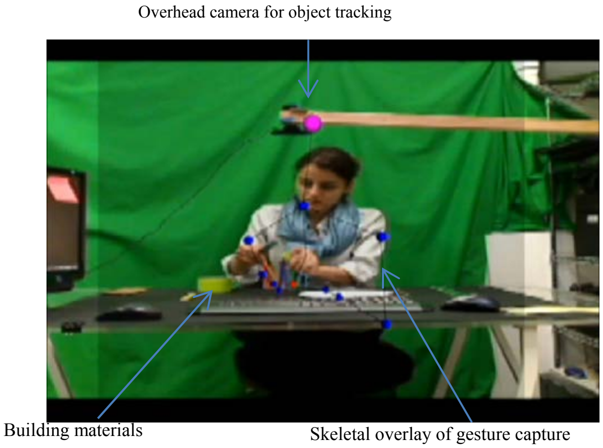
Overhead camera for object tracking
Building materials
Skeletal overlay of gesture capture

Figure 3. The capture environment used to record the audio, video and gesture data streams

Video captured the movement of objects as students progressed through the task, while gesture data, which consisted of twelve upper-body parts, recorded the students' actions. To analyze the data, many different approaches were attempted (Figure 4). First, a simple count of the number and duration of each of the codes ("single class assignment"), and seeing how well that predicted expertise based on the previously assigned expertise labels. Next, a more sophisticated cluster analysis was used, but it did not consider the sequence of codes/actions (just their quantity and duration, "non-process oriented classification"). Finally, the temporal sequence of actions was added, thus taking into consideration the sequentiality of building the object. In contrast to the non-process-oriented approach, the final object manipulation analysis algorithm was able to significantly outperform both random assignment and majority class assignment, all while preserving the process-oriented nature of the task. Figure 4 highlights the accuracy attained through the object manipulation analysis, compared to the other techniques.

Similarly, the confusion matrix is in Table 2.

Table 2. Confusion matrix of the object manipulation data, using the process-based algorithm, in terms of expertise levels

	Low	Medium	High	Expert
Low	3	0	0	0
Medium	3	1	1	0
High	0	0	2	0
Expert	0	0	0	3

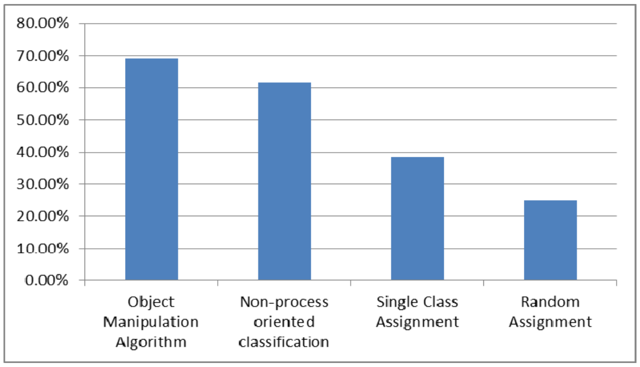

Figure 4. Comparison between different approaches to data classification

The confusion matrix shows that the algorithm worked best at uniquely clustering expert behavior, which it did at an accuracy of 100%. It also attained 100% accuracy for individuals of low expertise. However, for individuals of intermediate levels of expertise, the algorithm was less accurate. However, considering that the metric may be somewhat noisy for participants of medium expertise, it was still able to do a reasonably accurate job.

4.2 Tracking gestures using sensors

The gesture data analysis, while similar in spirit to the object manipulation analysis, involves markedly less complexity. This is preliminary work that I mention here only as an illustration of the possibilities of MMLA. Using the same setting as described in the previous study (4.1), gesture data was collected with a Kinect sensor. The hypothesis was it would be possible to cluster experts and novices in terms of the well-documented difference between the extents of two-handed coordinated movement among individuals of differing expertise. Here we consider two-handed coordinated movement to be when a participant uses both of their hands within a given action. Figure 5 shows the cumulative displacements for the right and left hand and depicts this difference. The expert's hands (right) typically move coordinated with one another, whereas the novice's hand movements are markedly asynchronous (left). These preliminary results deserve deeper examination, but are examples of low-cost techniques that could be useful when coordinated with other sources of multimodal data.

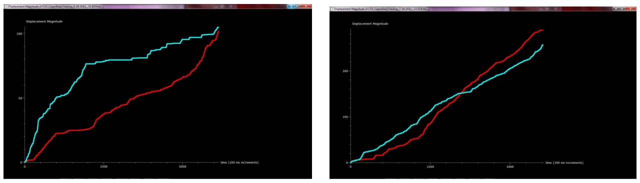

Figure 5. Novices' hands are used asynchronously (left), while experts' hands move at the same time (right).

5. CONCLUSIONS

In this paper, I presented a series of proof-of-concept studies for what I termed "multimodal learning analytics:" *a set of techniques that can be used to collect multiple sources of data in high-frequency (video, logs, audio, gestures, biosensors), synchronize and code the data, and examine learning in realistic, ecologically valid, social, mixed-media learning environments.*

The incorporation of multimodal techniques, which are extensively used in the multimodal interaction community, would allow researchers to examine unscripted, constructionist, complex tasks in more holistic ways. In particular, as a proof-of-concept, I focused on clustering participants in terms of multiple features of their actions and matching those clusters to the previously known expertise levels. First, I showed how the analysis of hundreds of programming snapshots could reveal patterns in students' programming such as 'tinkerers' vs. 'planners'. I then showed how even simple word counting and n-gram analysis techniques can reveal learners' affect and identity towards engineering, and how behavioral traces such as confidence, as well as disfluencies or pauses, can predict a subject's level of expertise. Using video analysis, I showed how the clustering of students' actions during a construction task, paired with human labeling, could also be a predictor of learning. Finally, I explored hand coordination as a possibly meaningful metric.

The goal of the paper is to be a proof of concept of novel assessment techniques along several *modes*. In all studies, I was interested in the definition of expertise and in the categorization of learners based purely on their behavior, actions, or utterances—not on the assumed level of their knowledge or their performance on extraneous tests. Many of these studies are preliminary; further studies should get deeper into the nuances of expertise, which was oversimplified for the purposes of this paper. Even with these simplifications, I was able to show that important aspects of learning "hide in the details," and both overt and tacit elements could be indicative to determine students' knowledge. The implication of this work is that, ultimately, *multimodal learning analytics* could be used to devise naturalistic assessments which would be, at the same time, social, ecologically valid, more inclusive as to the types of knowledge they measure, and enabling real-time evaluation in realistic tasks, either off or online.

6. REFERENCES

[1] Amershi, S., & Conati, C. 2009. Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining*, 1(1), 18-71.

[2] Baker, R. & Yacef, K. 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining,* 1(1).

[3] Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. 2004. Off-task behavior in the cognitive tutor classroom: when students "game the system". *Proceedings of the SIGCHI conference on Human factors in computing systems.*

[4] 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*. ACM: Banff, Canada. 110-116.

[5] Barron, B., & Darling-Hammond, L. 2010. Prospects and challenges for inquiry-based approaches to learning: OECD.

[6] Beck, J. E., & Sison, J. 2006. Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *International Journal of Artificial Intelligence in Education*, 16(2), 129-143.

[7] Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press.

[8] D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., & Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1), 45-80.

[9] Dewey, J. 1902. *The school and society*. U of Chicago Press.

[10] Dutson, A. J., Todd, R. H., Magleby, S. P., & Sorensen, C. D. 1997. A Review of Literature on Teaching Engineering Design through Project-Oriented Capstone Courses. *J of Engineering Education*, 86(1), 17-28.

[11] Dym, C. L. 1999. Learning Engineering: Design, Languages, and Experiences. *J of Engineering Education*, 145-148.

[12] Freire, P. 1970. *Pedagogia do Oprimido.* Paz e Terra, Rio de Janeiro.

[13] Kirschner, P. A., Sweller, J., & Clark, R. E. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75-86.

[14] Klahr, D., & Nigam, M. 2004. The equivalence of learning paths in early science instruction. *Psychological Science*, 15(10), 661.

[15] Levy, F., & Murnane, R. J. 2004. *The new division of labor: How computers are creating the next job market*: Princeton University Press.

[16] Montessori, M. 1965. *Spontaneous activity in education*. Schocken Books, New York.

[17] Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1): 1–135.

[18] Papert, S. 1980. *Mindstorms: children, computers, and powerful ideas.* Basic Books, New York.

[19] Piech, C., et al. Modeling how students learn to program. 2012. In *Proceedings of the 43rd ACM Symposium on Computer Science Education (SIGCSE '12)*. ACM.

[20] Rus, V., Lintean, M. and Azevedo, R. 2009. Automatic Detection of Student Mental Models During Prior Knowledge Activation in MetaTutor. In *Proc. of the 2nd Int. Conference on Educational Data Mining*. 161-170.

[21] Worsley, M. and Blikstein P. 2011. What's an Expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In *Proceedings for the 4th Annual Conference on Educational Data Mining*, Eindhoven, The Netherlands.

[22] Worsley, M. and Blikstein P. 2012. A Framework for Characterizing Student Changes in Student Identity During Constructionist Learning Activities. *Proceedings of Constructionism 2012*, Athens, Greece.

[23] Worsley, M. and Blikstein, P. 2013. Toward the Development of Multimodal Action Based Assessment. In *Proceedings for the 2013 Learning Analytics and Knowledge (LAK 2013) Conference*, Leuven, Belgium.