

Deciphering the Practices and Affordances of Different Reasoning Strategies through Multimodal Learning Analytics

Marcelo Worsley
Stanford University
520 Galvez Mall, CERAS 217
Stanford, CA 94305
mworsley@stanford.edu

Paulo Blikstein
Stanford University
520 Galvez Mall, CERAS 232
Stanford, CA 94305
paulob@stanford.edu

ABSTRACT

Multimodal analysis has had demonstrated effectiveness in studying and modeling several human-human and human-computer interactions. In this paper, we explore the role of multimodal analysis in the service of studying complex learning environments. We use a semi-automated multimodal method to examine how students learn in a hands-on, engineering design context. Specifically, we combine, audio, gesture and electro-dermal activation data from a study (N=20) in which students were divided into two experimental conditions. The two experimental conditions, example-based reasoning and principle-based reasoning, have previously been shown to be associated with different learning gains and different levels of design quality. In this paper we study how the two experimental conditions differed in terms of their practices and processes. The practices included four common multimodal behaviors, that we've entitled ACTION, TALK, STRESS and FLOW. Furthermore, we show that individuals from the two experimental conditions differed in their usage of the four common behavior both on aggregate, and when we model their sequence of actions. Details concerning the data, analytic technique, interpretation and implications of this research are discussed.

Categories and Subject Descriptors

K.3.1 [Computers and Education]: Computer Uses in Education – *Miscellaneous*.

General Terms

Algorithms, Human Factors.

Keywords

Learning Sciences; Computational; Constructionist; Data Mining

1. INTRODUCTION

Multimodal analysis has provided a powerful tool for studying complex human-human and human-computer interactions across a variety of domains. Furthermore, these technologies have had a strong impact on the development of multimodal interfaces that create more naturalistic, engaging and authentic environments. The development of multimodal interfaces has started to make its way into the education domain in the form of intelligent tutoring

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

MLA'14, November 12-16 2014, Istanbul, Turkey
Copyright 2014 ACM 978-1-4503-0488-7/14/11\$15.00
<http://dx.doi.org/10.1145/2666633.2666637>

systems, but has been scarcely explored in non-computer mediated environments. In this paper, we follow in the paradigm of multimodal learning analytics [1 - 3] in order to expand multimodal analysis to complex, hands-on learning environments. While multimodal analysis has long been a staple of education research, the introduction of computational multimodal techniques has received some resistance. This hesitation is understandable given the infancy of the field. In an effort to contribute to this discussion and propel the advancement of the multimodal learning analytics paradigm, we report on the affordances of employing a multimodal analysis in a complex learning environment and show how multimodal learning analytic techniques can be relevant for improving the field.

This study builds on our prior work [4], where we present two different approaches that students use in engineering design: example-based reasoning – using examples from the real-world as an entry point into solving a task; and principle-based reasoning – using engineering fundamentals as the basis for one's design. These two reasoning strategies complement prior work on analogical problem solving [5], case-based reasoning [6], mechanistic reasoning [7, 8] and expertise [9 - 11]. In [4] we described example-based reasoning and principle-based reasoning in qualitative terms, and then proceeded to use these two approaches in a controlled study (N=20) that compares how each approach impacts learning gains and performance during a collaborative hands-on activity. In that study we found that students in the principle-based reasoning condition engineered higher quality structures and also had higher learning gains from pre-test to post-test. The goal of this paper is to discover the multimodal practices that can help explain the observed differences in the two experimental conditions, learning and success. Furthermore, as a primary objective we want to show that the two experimental conditions, principle-based reasoning and example-based reasoning, are associated with markedly different processes.

In what follows we briefly present some pertinent prior literature; describe the experiment from which the data was derived; delineate the basic algorithm used to analyze the data; summarize important results; and discuss the implications of this work.

2. PRIOR LITERATURE

This paper builds on a rich body of research in educational data mining and learning analytics [12]. Through these disciplines researchers have demonstrated the ability to analyze data from a wide range of modalities. Previous work includes examples from intelligent tutoring systems that leverage: discourse analysis (e.g. [13]), content word extraction (e.g. [14]), uncertainty detection (e.g. [15]), sentiment analysis (e.g. [16 - 18]), linguistic analysis, prosodic and spectral analysis, and multi-modal analysis (e.g. [13, 19]). Additionally we leverage approaches from the multimodal

interfaces community: technological tools and research for integrating data streams [20, 21]; extracting rich context from audio-video data [22]; and behavior [23], for example. We also build on work from engineering education that has focused on studying design patterns among novices and experts (e.g. [24]).

Through the computer science and learning analytics communities we are continuously seeing new technologies and techniques for analyzing data. The challenge, however, is to leverage those technologies in a way that aligns with learning theory, and that allows us to answer important questions about learning. In terms of learning theory, this work is informed by that of [25] which qualitatively identified different multimodal learning and/or epistemic states. These states were characterized by body pose, amount of dialogue and student gaze as they worked in groups to complete a worksheet. Because of differences in context, we are not examining the same modalities, nor do we expect to see the same states. Nonetheless, the underlying assumptions around the connection between behavior, epistemology and learning remain central to the approach that we employ.

Finally, the analytic tools used to identify student behavioral states borrows strategies from our prior work in learning analytics [26 - 29]. In these papers we analyzed high frequency data by reducing the dataset down to a set of representative states that generalized to all of the participants. This is the same approach that we will employ in this paper, but differ from the prior work in that we construct states that include multimodal data, as opposed to constructing states from log-files, computer program snapshots or click-streams.

3. METHODS

To provide the reader with additional context, we briefly describe the research participants and the task that they completed before entering into a discussion of the analysis and results.

Students used common household materials: one paper plate, 4 ft. of garden wire, four drinking straws and five wooden Popsicle sticks. The objective was to use the materials provided to create a structure that could support a weight of approximately half a pound. Participants were also asked to support the weight as high off the table as possible.

Our population of students consisted of twelve 9th- through 12th-grade students and eight undergraduate students. Pairs of students were randomly assigned to either use example- or principle-based reasoning, after controlling for prior education experience. Thus, each condition had six high school students and four undergraduate students. In the example-based condition, students generated three example structures from their home, community or school in order to motivate their design. In the principle-based condition, students identified three engineering principles that conferred strength and stability to a ladder, an igloo and a bridge before embarking on the building task.

The data capture environment included: a Kinect sensor – for capturing audio, gesture and video; a high resolution web camera - to record how students moved the different materials; and an electro-dermal activation sensor – for measuring stress and/arousal. All sensor data was synchronized through the data collection software, and also verified by a research assistant.

3.1 Activity Sequence

The overall flow of activities that students completed included: a pre-test; an intervention, i.e. one of the two conditions; a preliminary design drawing; a hands-on, paired, building activity; post-test; and reflection (Figure 1).

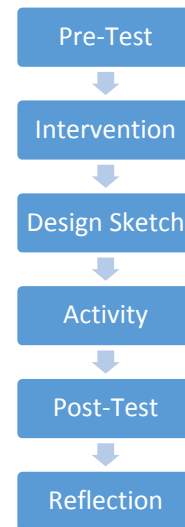


Figure 1. Overall study design

3.2 Data

Audio data. Data was derived from a combination of audio channels from an overheard web camera, and from the Xbox Kinect sensor. A custom piece of software was developed based on the Carnegie Mellon University (CMU) Sphinx Speech Recognition Toolkit. Specifically, the source code was modified to leverage the program’s voice activity detection feature. Voice activity detection is an automated means for determining when voice-based audio is being generated. Several speech recognition software solutions contain some variant of voice activity detection. The custom software provided voice detection start and stop times for all of the audio channels. Audio was considered to be present if either of the audio sources detected a voice, within a given second of time. Thus the final format of this data is a binary representation. Every second of the activity is labeled with a zero or one, for the absence or presence of audio at that time stamp. Because the audio channel captured sound from both participants this piece of data is the same for each person in a dyad.

Hand/wrist movement. Hand/wrist movement data was also generated from the Xbox Kinect sensor. Once again a custom built application was used to store three dimensional data for twelve upper body joints. The application uses native features available from the Kinect for Windows SDK, specifically, the ability to conduct skeletal tracking in the seated position. The custom application stores the data at 10 Hz. From the file generated, we utilize only the left and right wrist, hand and elbow data points. For each successive pair of data points we compute the angular displacement for the vectors that connect: left wrist and left hand; left wrist and left elbow, right wrist and right hand; right wrist and right elbow. The eventual angular displacement that is recorded is an average of the four angular displacements. Using angle as the means for comparison reduces biases introduced by participants having different sized bodies and limbs. Accordingly, for each tenth of a second in time we have stored the total angular hand/wrist displacement.

Electro-dermal Activation. Electro-dermal activation (also referred to galvanic skin response and/or skin conductance) readings were captured at 8 Hz. Processing electro-dermal activation data involved controlling for individual differences in variance, as well as individual differences in stress response. In practice, this was achieved by collecting baseline data as students completed the task of counting down by 7. We will refer to this as the “math” stress

test. As additional baseline data, students also completed a Stroop test, and had their electro-dermal activation recorded during non-task related activities. As before, each data point was time-stamped with the local date and time. Each data point was then transformed into an index value by subtracting the mean from the “math” stress test, and then dividing by the standard deviation of the “math” stress test data for that student. When we compared electro-dermal activation index values across the different activities, there were no statistically significant differences between experimental conditions for the baseline data, the Stroop test, or the math test. However, across the intervention, design phase and the building activity differences were statistically significant. This provided validation that this normalization was effective.

3.3 Algorithm

The approach follows our previous work [26, 27] on analyzing design strategies and success on hands-on engineering tasks. Here we significantly extend that work by incorporating multimodal data, as opposed to simply using hand-coded data. A visualization of the general algorithm is described Figure 2.

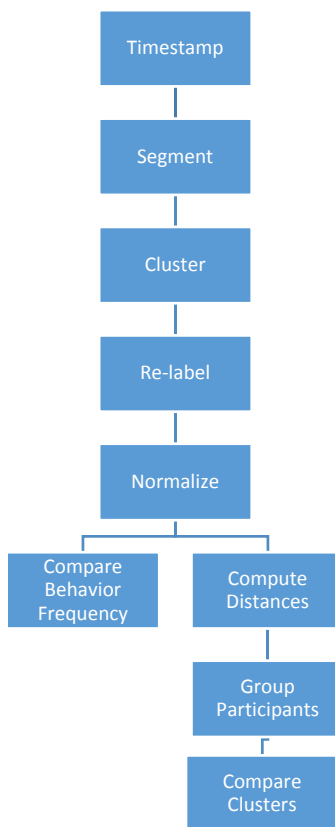


Figure 2. General algorithm

Time-stamp. The first step of extracting process data is to ensure that all data is properly time-stamped. This provides a means for synchronizing across the different modalities.

Segment. The time-stamped data is then segmented. Data is segmented every time a pair’s structure is tested. We interpret testing as representing an instance in which at least one person in the pair is eliciting feedback that will update the students on the

current stability of their structure. Testing usually takes the form of a team member placing the weight on the structure.

Segmentation always resulted in a single value for each data stream. For the audio data the value is the proportion of the “test segment” during which voice activity was detected. For the hand/wrist movement data, the value is the average total angular displacement during that “test segment.” Finally, the electro-dermal activation value is the average index value during that particular “test segment.”

As a whole, the segmentation process serves to smooth the data. Instead of having to take into account each of the spikes and troughs that may emerge from any of the data streams, segmentation allows us to look at broader trends. Noise reduction is also achieved during the following step.

Cluster. After the segmentation process, there are hundreds of unique “test segments.” Some of these will be very similar to one another, only differing by an infinitesimal amount, while others vary quite extensively from one another. The goal of clustering is to identify natural groupings among the various “test segments” and ultimately provide a common set of states, or behaviors, by which to compare individual user sequences. However, before proceeding with clustering, we first do data standardization. Namely, we adjust each value, such that all of the data in a given column has a mean of zero and a standard deviation of one. This process eliminates bias in clustering, by ensuring that each column contributes equally to the distance metric, which is this case with Euclidean distance. After standardizing the data, we used X-Means clustering to group the data points into a set of clusters that place each “test segment” with the other “test segments” that it is most similar to. Once each “test segment” has been grouped with similar “test segments,” each cluster, or group, can be described based on the average values of all of its members. These values provide the basis for determining common behavioral practices across the three analyses in Chapter 3.

Re-label. All “test segments” that are put into the same cluster are given the same name. Accordingly, each student’s sequence of “test segments” can now be represented as a list of clusters.

Normalize. In the normalization step, each student’s re-labeled sequence is lengthened to permit direct comparisons between each pair of participants. The two forms of normalization used are L-1 normalization and dynamic time warping [30]. In the case of L-1 normalization, each sequence is lengthened so that all participants’ sequences are of equal length. In dynamic time-warping, a modification of Levenshtein distance [31] is used to find the best match between pairs of sequences.

Compare Behavior Frequency. After L-1 normalization, the next step is to compare behavior frequency data across the three metrics of interest: success, experimental condition; and learning. The comparisons are based on Mood Median Tests along each of the individual clusters of “test segments.” However, instead of the traditional Mood Median Test, which computes statistical significance based on a Chi-Square distribution, we use a binomial test. These two tests were used because the data did not meet the requirements for MANOVA and violated the typical requirements of a Chi-Square Test. This step represents the conclusion of one branch of the analysis tree (left hand branch of Figure 2).

Compute Distances. Following dynamic time warping, a distance is computed between each pair of participants.

Group Participants. Based on the pair-wise distances, similar participants are forced into one of two groups using K-Means clustering. In each case a student is put into the group that contains

other students whose process was most similar to their own. Forcing the students into one of two groups was done to align with the two experimental conditions.

Compare Participant Groups. Finally, the groups are compared using a binomial test to determine the probability that individuals were randomly assigned to their specific group. Specifically, it is here that we examine the hypothesis that different groups, as partitioned by experimental condition, success on the activity, or based on post-test score, used markedly different processes from one another.

4. Results

4.1 Cluster Centroid Analysis

The first entry point for analysis is the generalized “test segment” cluster centroids that are derived from the X-Means clustering step. In particular, we found four multimodal “test segment” types (Figure 3). In discussing each cluster centroid, we will proceed from the most common “test segment” to the least common (Figure 4). Additionally, for each cluster, we created a name that captures the primary multimodal behaviors associated with that cluster of “test segments.

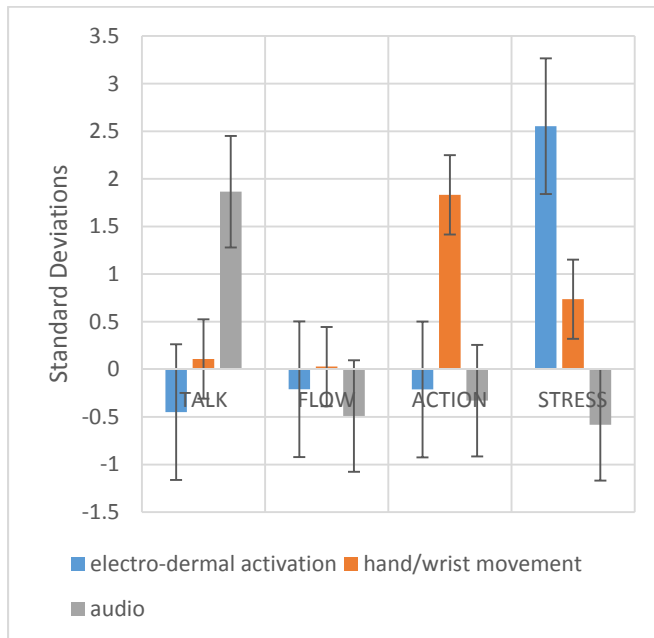


Figure 3. “Test Segment” cluster centroids for multimodal data

The most common segment, FLOW, is characterized by near or below average behavior across all three variables: audio, hand/wrist movement and electro-dermal activation; and is roughly 60% of all “test segments.” As one can likely deduce, calling this cluster FLOW is a reference to [32]. As we compare the usage across conditions, rate of success and quality of learning, the argument for calling this category FLOW will become clearer. For now, suffice it so say that this cluster represents the vast majority of all “test segments”, and that it is typified by average stress, average movement and little audio, providing initial indication that this state is in line with Csikszentmihalyi’s flow. One can picture students in FLOW concentrating on the task by carefully manipulating the materials without the need for extensive discussion, movement or stress.

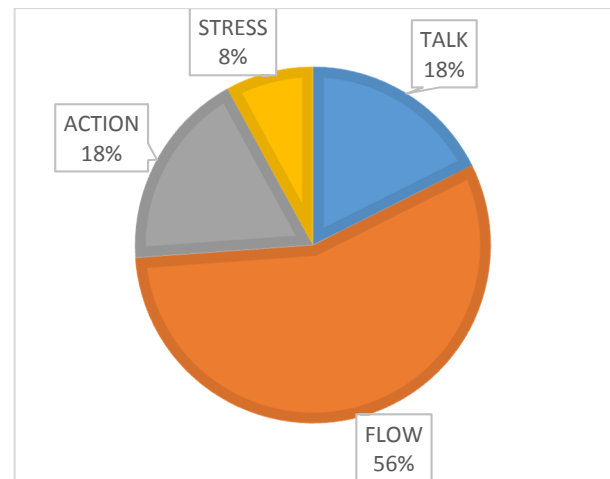


Figure 4. Distribution of Multimodal “Test Segments”

The second most frequently occurring “test segment” is ACTION, which represents 18% of all “test segments.” This behavior primarily consists of students who are currently engaging in above average hand/wrist movement. What’s more, though, is that this occurs in the absence of high electro-dermal activation, which is normally correlated with body movement. An additional point of interest is the lack of audio associated with this behavior. Students are focused on building and refraining from extensive discussion with one another. Accordingly, one might conjecture that the students are finding other means through which to communicate with one another.

After ACTION, the most frequently occurring state is TALK. This particular cluster represents approximately 18% of all “test segments.” The amount of audio in this cluster is approximately two standard deviations above the mean. Hand/wrist data is just above the mean, and electro-dermal activation is nearly half a standard deviation below average. Again, this is analogous to ACTION in that students appear to only engage one of the multimodal behaviors at a given time.

The final cluster is STRESS. This behavior is characterized by extremely large values of electro-dermal activation, as well as above average hand/wrist movement. As mentioned before, one would expect for electro-dermal activation and hand/wrist movement to correlate with one another. Hence we can anticipate that the electro-dermal activation values may be slightly inflated, but are still, more than likely, well above average. On the other hand, that the students are stressed, may be causing them to work more frantically, which would result in an increase in hand/wrist movement. STRESS accounts for approximately 10% of students “test segments.”

4.2 Process Similarity Analysis

Before discussing the specifics of the process differences we first present results from grouping students based on the similarity of their processes. Recall that this analysis is derived from the right hand-branch of Figure 2. Seven of the eight students assigned to Group A are from the principle-based condition. The inverse pattern is observed for Group B, with seven of the eight individuals in that group coming from the example-based reasoning condition (Figure 5).

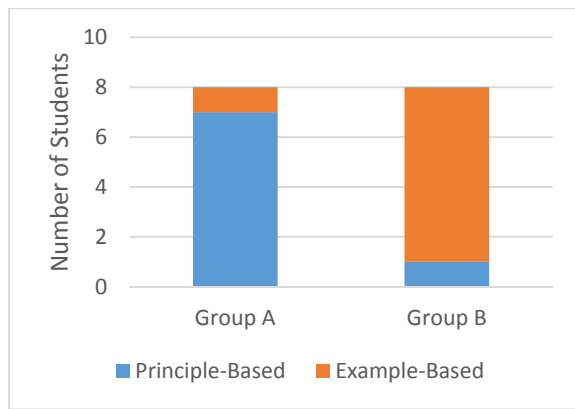


Figure 5. Composition of groups based on experimental condition as derived from process similarity

The likelihood of this happening randomly is less than 0.002, suggesting that the two conditions did, in fact, utilize markedly different processes. One question is if these results are being inflated by the dyadic nature of the task. For example, two individuals that work together are likely to mirror each other's behavior. When we control for this by removing the partner from consideration, students are still likely to be most similar to another individual from the same experimental condition.

Having observed that there are salient differences between the processes that the two conditions use, as determined through multimodal data, we now consider the nature of those differences. Recall that these differences are based on Mood Median Tests where we used a binomial distribution as opposed to a Chi-Squared distribution. Accordingly, all p-values are based on an exact binomial test.

4.3 Coarse-Grain Cluster Usage Analysis

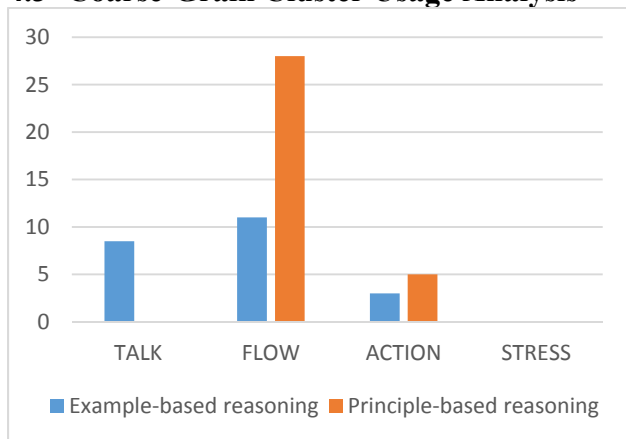


Figure 6. Median common behavior usage by condition

Figure 6 shows the median normalized frequency of cluster usage by experimental condition. From the figure it is apparent that the median value for FLOW is quite different between the two conditions. A test of statistical significance confirms that students in the principle-based reasoning condition were more frequently ($p = .007$) in FLOW than their peers in the example-based reasoning condition. No other statistically significant differences emerged between the two groups when comparing their cluster frequency.

In terms of success there are no statistically significant differences in terms of cluster, or behavior, usage (Figure 7).

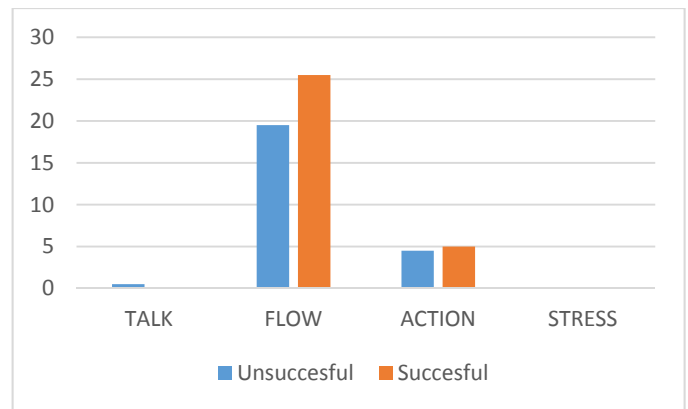


Figure 7. Median common behavior usage by success

Much like the case of success, the data does not reveal any significant differences between students who experienced positive learning, and those who experienced negative learning (Figure 8).

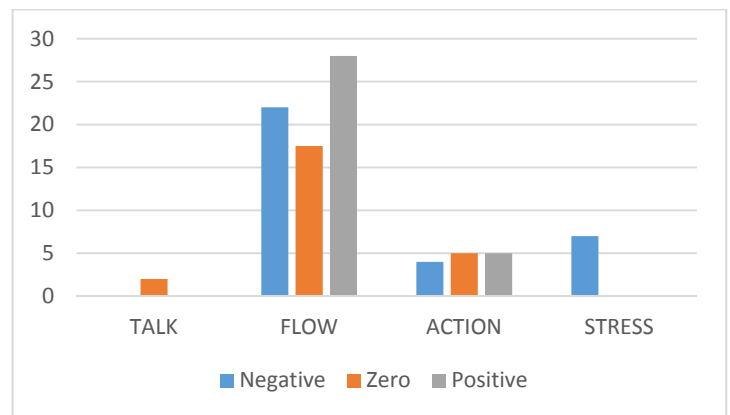


Figure 8. Median common behavior usage by learning score

4.4 Fine-Grain Cluster Usage Analysis

A fine-grained analysis indicates that there are significant differences between the two conditions during all three portions of the activity. Students in the principle-based reasoning condition are more likely to be in FLOW for the first ($p < 0.001$ (1.9×10^{-5})), second ($p = 0.007$) and third ($p = 0.007$) thirds, than their peers in the example-based reasoning condition. This suggests that the two conditions are most divergent during the first portion of the activity. In particular, many of the students from the example-based reasoning condition primarily spend the first third of the activity in ACTION.

The fine-grained analysis did not identify any significant differences between successful and unsuccessful students in terms of how frequently they used the different clusters, or common behaviors.

The fine-grained analysis did indicate that students that learned more through the activity were more likely to be in FLOW during the first third of the activity. Specifically, the students that received positive post-test scores were much more likely ($p < 0.001$ (0.0004)) to use FLOW than students who received a negative post-test score. This trend continues among the students that received a score of zero, but not at a statistically significant level.

5. DISCUSSION

In deciphering the differences between the principle-based reasoning condition and the example-based reasoning condition, the current multimodal analysis included statistically significant

differences in the process similarity metric between the experimental conditions. Furthermore, both the coarse-grain and fine-grained analyses offered important insights into identifying the elements of each condition's process that differed. Specifically, coarse and fine-grained analyses showed that students in the principle-based reasoning condition made more extensive use of FLOW than students in the example-based reasoning condition.

Important to note, however, is that FLOW was characterized by multimodal behavior, meaning that a uni-modal analysis would not have been able to identify the presence and usage of FLOW, since FLOW was based on all three modalities. This provides support for the need to employ multimodal analyses.

Differences in FLOW usage emerged during all three portions of the activity and was most pronounced during the first third. This is telling because it indicates that the differences were not merely the result of students being more or less successful on the activity. In fact, this analysis did not reveal any differences between successful and unsuccessful students at any grain size. Instead, the only other difference was observed from the learning metric. Once again, FLOW was positively correlated with student learning. From this, we can deduce that FLOW is one of the behaviors that characterized the principle-based reasoning condition, and also played a role in fostering student learning during the activity.

As we consider possible interpretations for these results we will again turn to prior work on epistemological frames. Specifically, in [33] the authors showed that multimodal behavior during cognitive clinical interviews can be correlated with three common epistemological frames: the inquiry frame; the expert frame; and the examination frame. Of particular relevance to this paper are the inquiry frame and the expert frame. The inquiry frame is typified by students whose work pattern is somewhat disjointed in that they may regularly engage in inquiry related practices. This is in contrast to the expert frame where students tend to be able to work through a given question or response without extensive hesitation. As we compare the usage of FLOW between the example-based reasoning condition and the principle-based reasoning condition, we see that these two experimental conditions may be related to the two different epistemological frames. Namely, example-based reasoning may be more closely associated with the inquiry frame, and the principle-based reasoning condition may bear greater resemblance to the expert frame. For example, students in the principle-based reasoning condition are able to stay in FLOW during a larger portion of the activity meaning that they complete less "inquiry" during the course of the building task (here we take the TALK, ACTION and STRESS behaviors as being more closely associated with acts of inquiry). Thus the theoretical framework used to motivate the initial multimodal analysis can also be used to explain the nature of the differences that we observed through our research study.

6. LIMITATIONS AND FUTURE WORK

One limitation is that we have focused on behavior level features. For example, the audio, hand/wrist movement and electro-dermal activation data was all low-level. In future work we will conduct more semantic analyses of the individual modalities before subjecting them to data fusion. For example, a series of hand/wrist movements may be interpreted as a specific gesture (e.g. pointing, shrugging, etc.). This may help identify additional commonalities, as well as highlight more of the complex multimodal behavioral patterns that students use.

Additionally, in future work we plan to explore additional problem and content domains. The current work was restricted to an engineering design context where students were issued a challenge.

While there was great variability in ways to successfully complete the activity, we would like to examine these behavioral patterns on a design task that is entirely of the student's choosing. Finally, in future research we employ new data collection techniques that offer more fine-grain capture of how students are interacting with each other and with the materials that are at their disposal.

7. CONCLUSION

Studying complex, open-ended, hands-on learning environments is be quite challenging. Researchers may spend significant time manually coding and analyzing video data with the hope of understanding the effect of a given treatment, or experimental condition. In this paper we have shown how a semi-automated analysis of multimodal data can be used for studying user behaviors during different learning treatments. We have also shown how the results derived from our study can be explained through the same theoretical lens that contributed to its framing and design. Our hope is that through this paper, researchers will develop additional insights and motivation for leveraging multimodal learning analytics to study complex learning environments in ways that can clearly connect to the learning sciences.

8. REFERENCES

- [1] Worsley, M. 2012. Multimodal Learning Analytics: Enabling the Future of Learning through Multimodal Data Analysis and Interfaces. In Proceedings of the 14th ACM international conference on Multimodal interaction (ICMI '12). ACM, New York, NY, USA. 353-356.
- [2] Blikstein, P. 2013. Multimodal Learning Analytics. In *Proceedings of the 3rd Annual Learning Analytics and Knowledge Conference*. Leuven, Belgium.
- [3] Scherer, S., Worsley, M. and Morency, L. 2012. 1st international workshop on multimodal learning analytics: extended abstract. In Proceedings of the 14th ACM international conference on Multimodal interaction (ICMI '12). ACM, New York, NY, USA, 609-610.
- [4] Worsley, M., & Blikstein, P. (2014). Assessing the Makers: The Impact of Principle-Based Reasoning on Hands-on, Project-Based Learning. *Proceedings of the 2014 International Conference of the Learning Sciences (ICLS)*, 3, 1147-1151.
- [5] Gentner, D. and Holyoak, K. J. 1997. Reasoning and Learning by Analogy, 52(1), 32-34
- [6] Kolodner, J. L. 1997. Educational implications of analogy: A view from case-based reasoning. *American psychologist*, 52(1)
- [7] Russ, R. S., Coffey, J. E., Hammer, D., & Hutchison, P. 2009. Making classroom assessment more accountable to scientific reasoning: A case for attending to mechanistic thinking. *Science Education*, 93(5), 875-891. doi:10.1002/sce.20320
- [8] Lehrer, R., & Schauble, L. (1998). Reasoning about structure and function: Children's conceptions of gears. *Journal of Research in Science Teaching*, 35(1), 3-25.
- [9] Chi, M. Glaser, Rees. 1981. Expertise in problem solving.
- [10] Nokes T J, Schunn C D and Chi M. 2010, Problem Solving and Human Expertise. In: Penelope Peterson, Eva Baker, Barry McGaw, (Editors), *International Encyclopedia of Education*. volume 5, pp. 265-272. Oxford: Elsevier.
- [11] Ahmed, S., & Wallace, K. M. 2003. Understanding the differences between how novice and experienced designers

- approach design tasks, 14, 1–11. doi:10.1007/s00163-002-0023-z
- [12] 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), 3-17.
- [13] Litman, D., Moore, J., Dzikoska, M., and Farrow. E. 2009. Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains and Modalities. *Proceedings 14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK, July.
- [14] Chi, M., Van Lehn, K., Litman, D., & Jordan, P. 2010. Inducing Effective Pedagogical Strategies Using Learning Context Features. In: *Proc. of the 18th Int. Conference on User Modeling*.
- [15] Liscombe, J., Hirschberg, J., and Venditti, J. 2005. Detecting Certainty in Spoken Tutorial Dialogues. In *Proceedings of Interspeech 2005—Eurospeech*, Lisbon, Portugal.
- [16] Craig, S. D., D'Mello, S., Witherspoon, A. and Graesser, A. 2008. 'Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive-affective states during learning'. *Cognition & Emotion*, 22: 5, 777 — 788.
- [17] D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., and Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction* 18, 1-2 (Feb. 2008), 45-80.
- [18] Conati, C. and Maclaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*. 267-303.
- [19] Worsley, M., & 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 of the Fourth Annual Conference on Educational Data Mining (EDM 2011)* (pp. 235–240). Eindhoven, Netherlands.
- [20] Fouse, A., Weibel, N., Hitchins, E. & H, J. 2011. Chronoviz A System for Supporting Navigating Time-Coded Data. In *Proceedings of CHI 2011, SIGCHI/ACM*. pp. 299-304.
- [21] Worsley, M., Johnston, M. and Blikstein, P. 2011. OpenGesture: a low-cost authoring framework for gesture and speech based application development and learning analytics. In *Proceedings of the 10th International Conference on Interaction Design and Children (IDC '11)*. ACM, New York, NY, USA, 254-256.
- [22] Schick, A., Morlock, D. Amma, C., Schultz, T. and Stiefelhagen, R. 2012. Vision-Based Handwriting Recognition for Unrestricted Text Input in Mid-air. In *Proceedings of the 14th ACM International on Multimodal Interaction (ICMI '12)*. ACM, New York, NY, USA, 217-220
- [23] Song, Y., Morency, L. and Davis, R. 2012. Multimodal Human Behavior Analysis: Learning Correlation and Interaction Across Modalities. In *Proceedings of the 14th ACM International on Multimodal Interaction (ICMI '12)*.
- [24] Atman, C. J., Cardella, M. E., Turns, J., & Adams, R. 2005. Comparing freshman and senior engineering design processes: an in-depth follow-up study. *Design studies*, 26(4), 325-357.
- [25] Scherr, R. E., & Hammer, D. 2009. Student Behavior and Epistemological Framing: Examples from Collaborative Active-Learning Activities in Physics. *Cognition and Instruction*, 27(2), 147–174.
- [26] Piech, C., Sahami, M., Koller, D., Cooper, S., & Blikstein, P. 2012. Modeling how students learn to program. In *Proceedings of the 43rd ACM technical symposium on Computer Science Education* (pp. 153-160). ACM.
- [27] Worsley, M., & Blikstein, P. 2013. Towards the Development of Multimodal Action Based Assessment. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 94–101). New York, NY, USA: ACM.
- [28] Worsley, M., & Blikstein, P. (2014). Analyzing Engineering Design through the Lens of Computation. *Journal of Learning Analytics*, 1(2), 151–186.
- [29] Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., & Koller, D. (in press). Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. *Journal of Learning Sciences*.
- [30] Rabiner, L. R., Rosenberg, A. E., & Levinson, S. E. 1978. Considerations in dynamic time warping algorithms for discrete word recognition. *The Journal of the Acoustical Society of America*, 63(S1), S79–S79.
- [31] Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- [32] Csikszentmihalyi, M. 1992. *Flow : the psychology of happiness*. London: Rider.
- [33] Russ, R. S., Lee, V. R., & Sherin, B. L. 2012. Framing in cognitive clinical interviews about intuitive science knowledge: Dynamic student understandings of the discourse interaction. *Science Education*, 96(4), 573–599.