Unraveling Students' Interaction Around a Tangible Interface Using Multimodal Learning Analytics

Bertrand Schneider
Stanford University
schneibe@stanford.edu

Paulo Blikstein
Stanford University
paulob@stanford.edu

In this paper, we describe multimodal learning analytics (MMLA) techniques to analyze data collected around an interactive learning environment. In a previous study (Schneider & Blikstein, 2015), we designed and evaluated a Tangible User Interface (TUI) where dyads (i.e., pairs) of students were asked to learn about the human auditory system by reconstructing it. In the current study, we present the analysis of the data collected in the form of logs, both from students' interaction with the tangible interface as well as from their gestures, and we describe how we extracted meaningful predictors for student learning from these two datasets. First we show how information retrieval techniques can be used on the tangible interface logs to predict learning gains. Second, we explored how Kinect™ data can inform "in-situ" interactions around a tabletop by using clustering algorithms to find prototypical body positions. Finally, we fed those features to a machine-learning classifier (Support Vector Machine) and divided students in two groups after performing a median split on their learning scores. We found that we were able to predict students' learning gains (i.e., being above or belong the median split) with very high accuracy. We discuss the implications of these results for analyzing rich data from multimodal learning environments.

Keywords: Collaborative Learning, Tangible User Interface, Kinect Sensor, Machine Learning, Information Retrieval Techniques.

1. INTRODUCTION

Students' gestures have received a great deal of attention from learning scientists over the last decades. Body movements and postures can provide valuable clues about students' prior knowledge, misconceptions, and problem-solving strategies when learning new concepts. Numerous studies have unraveled links between students' understanding of various topics (e.g., Church, 1999; Abrahamson, Trninic, Gutiérrez, Huth & Lee, 2011) and specific gestures (e.g., deictic, iconic, and metaphoric gestures; Roth, 2001). More generally, there has been a plethora of studies about people's intuitive representations of everyday situations and bodily language (e.g., embodied cognition; Anderson, 2003). This line of research has provided new ways to understand how students integrate new concepts into their everyday understanding of science phenomena using gestures and, more generally, body actions.

Yet, this field of research suffers from serious methodological limitations. Most studies are qualitative by nature or require researchers to manually annotate hours of video recordings. Now that the theoretical underpinnings of the field are established, it is the right time to speed up discovery and data analysis by using motion sensors and data mining techniques. The emerging field of multimodal learning analytics (Blikstein & Worsley, accepted) might provide just the right data collection and analysis tools to tackle this problem.

The goal of this paper is to address this methodological gap by suggesting new ways to conduct research on students' body language, as well as providing new lenses to look at students' micro-behaviors during their learning process (e.g., Siegler & Crowley, 1991). In our study, we collected data on users' actions from two sources: a Microsoft Kinect™ sensor, which tracks students' body movements, and the logs generated by a table-top tangible interface. Our goal is to apply data mining techniques on those two datasets to: 1) find patterns that characterize productive groups; 2) use various data mining algorithms and explore their potential to extract predictors for learning; and 3) investigate the social component of our dataset in terms of body synchronization and proxemics. Thus our approach is mostly data-driven, in the sense that we don't have exact predictions of the type of patterns we will find. However, our analyses are grounded by several psychological and educational theories (such as proxemics: Hall, 1966; student status in small collaborative groups: Shaer, Strait, Valdes, Feng, Lintz & Wang, 2011; bimanual coordination: Worsley & Blikstein, 2013; and more generally embodied cognition: Anderson, 2003), which makes our work partially theory-driven.

In this paper we asked students to accomplish a learning task, which involved reconstructing the inner workings of the human ear using a specially designed tabletop exploratory environment. We tracked their gestures during the task using a Kinect sensor and applied state of the art data mining techniques on this dataset and the logs generated by the tangible interface. Finally we gave students pre, mid and post-tests, which provided us with learning gain scores for each participant.

The design of the system, the experimental setup, the participants, and the behavioral results (e.g., scores on the learning tests) are reported in details in a previous conference publication (Schneider & Blikstein, 2015). Some preliminary analyses of the Kinect data were also presented at EDM2014 as a short paper (Schneider & Blikstein, 2014). We extend here these previous publications with a more comprehensive literature review, an analysis of the logs collected by the Tangible User Interface (TUI), additional analyses of the Kinect logs, and an

extended discussion of the implications of our findings for mining educational datasets collected from an interactive tabletop.

The next sections are structured as follows. First, we review foundational work in analyzing students' gestures and summarize the findings of previous research. We then describe our datasets and the study that generated them. Finally, we report our analyses and results. We conclude by describing the implications of our work for automatically detecting students' body language and mention the limitations of our approach.

2. LITERATURE REVIEW

From a reductionist perspective, previous work on studying gestures in education can be divided in two groups: qualitative studies, where manual coding schemes are applied to video frames by human coders; and quantitative studies, where algorithms and more generally computational techniques are used to study students' body language. Since low-cost motion sensors have been made available to the general public only recently, the second category contains less literature. We will briefly survey both approaches (with a greater focus on computational techniques) to provide a theoretical grounding to this paper.

In terms of qualitative papers, we will keep a high level focus and will mostly describe foundational prior work. One example of an influential contribution in this domain is the one of Roth (2001) where he describes a taxonomy of gestures in education and provides related examples: 1) in collaborative learning settings, students can use *beats*, which are gestures "void of propositional or topical content yet lend a temporal or emphatic structure to communication" (e.g., a tapping motion to emphasize certain utterances); 2) they can use *deictic gestures*, which encompass all pointing behaviors (often associated with *deictic terms*, such as here, there, this, that, and so on); 3) *iconic* gestures are displayed to mimic concrete entities and events (e.g., tracing a trajectory in the air is an example of iconic gesture); and 4) finally, *metaphoric* gestures are similar to iconic ones but refer to abstract entities (for instance, closing the gap between two hands to express the idea of "approaching the limit" in mathematics is an example of a metaphoric gesture). It should be noted that those gestures are very fine-grained and sometimes require domain expertise to be detected. This is one of the great strength of qualitative analysis: semantic meaning can easily be inferred and researchers can take full advantage of contextual information. One weakness is that qualitative research is so time consuming that it limits itself to describe few, isolated learning episodes. This is a gap that computational techniques aim to fill: with sensors and algorithms, we can detect and classify *all* the gestures and body language being used by students during their learning process.

In terms of quantitative studies, one noteworthy attempt at automatically classifying gestures is the Multi-modal Gesture Recognition Challenge 2013 (Escalera, Gonzàlez, Baró, Reyes, Lopes, Guyon, & Escalante, 2013). In summary, a group of researchers provided a large video database of ~14,000 gestures (1,720,800 frames) and asked participants to predict to which 20 categories of Italian gestures each dataset belonged to. The data was recorded using a Kinect sensor, providing the audio, skeletal model, user mask, RGB, and depth images. Even though this is a challenging temporal multi-classification task, the best competitors achieved a 12.76% error rate on the test set. This is a promising and impressive result, considering the complexity of the problem. In a different study, Grafsgaard, Wiggins, and Boyer (2014) used a multimodal data stream (such as automatic facial expression recognition, Kinect depth images

and system logs) to predict students' cognitive states. They found that facial expression and gestures (e.g., hand-to-face gestures) were predictive of engagement and frustration, while facial expressions and body posture (e.g., distance from the computer screen) were predictive of learning. Finally, in engineering education, Worsley and Blikstein (2013) tackled the challenging task of finding multimodal markers of expertise. They used a construction task where students had to build structures to hold a certain weight. Using a Kinect sensor, they found that bimanual coordination was predictive of expertise: advanced students were more likely to use both hands in a more or less synchronized fashion, while novices tended to use mostly one hand. Combined together, all those results suggest that data collected with motions sensors can generate many useful predictors for students' emotional and cognitive states, as well as their learning outcomes.

Our study is also concerned with capturing the quality of students' interactions, since our participants worked in pairs. There is obviously a wealth of frameworks describing the types of gestures and utterances students exhibit while learning in a collaborative fashion. Those frameworks are rich, precise, and sometimes require contextual information and domain knowledge. For instance, there is a large body of work on behavioral convergence looking at multimodal indicators of behavioral synchronization between group members (e.g., Pentland & Heibeck, 2008). One concrete application is given by Gweon, Jain, Mcdonough, Raj and Rosé (2013) about speech-style accommodation theory applied to student's transactivity (e.g., the tendency to build on each other's ideas). In that particular example, the authors found a positive correlation between automatic measures of speech-style accommodation (captured by a Dynamic Bayesian Network model) and manually coded transactive moves. Those results suggest a promising potential in terms of automatically capturing the quality of students' interactions. This is, however, just one example where behavioral convergence has been investigated using sensors and machine learning algorithms. For space considerations, we cannot do justice to this entire body of work. Researchers interested in capturing these types of student interactions could dive deeper into this literature by starting with Pentland & Heibeck (2008) or Gweon, Jain, Mcdonough, Raj and Rosé (2013).

Finally, some believe that the goal of computational techniques is to make the task of detecting those gestures automatic. However, we believe that some of the categories described above are not suitable for the task of detecting particular types of gestures (e.g., metaphoric gestures). Some others, such as deictic and iconic gestures, may be good candidates for automatic detection. In general, we believe that merely replacing human coders with computational tools does not take full advantage of the potential offered by sensors and algorithms. Our approach is to consider computational techniques as an *augmentation* of traditional research methods: algorithms can replace some qualitative analyses, but most of them are just too complex to be replicated automatically. Thus an additional goal is to use computers to provide an alternative perspective on educational datasets, which can then be used as lenses for constructing new hypotheses and analyses that couldn't be generated with traditional methods.

The goal of this paper is much more modest than the research plan described above. As a first pass, we aim at finding simple measures predictive of collaborative learning in our datasets. In that sense, our approach is similar to the approach used by Grafsgaard, Wiggins, and Boyer: before we can tackle more challenging problems such as gesture classification, we need to conduct lower level analyses to get a sense of the relevant signals buried in our logs. Once we

more fully understand our datasets, we plan to conduct future work to tackle more complex problems such as differentiating between deictic and iconic gestures.

This literature review is succinct: our goal was to provide a high level sense of the research on automatic gesture detection and not exhaustively review every prior work that studied gestures or used a Kinect sensor. Additionally, instead of reviewing the theoretical papers that inspired our measures here, we decided to introduce those concepts at the beginning of each corresponding section for coherency.

3. EXPERIMENT AND DATASETS

3.1. EXPERIMENTAL DESIGN

In previous research, we have studied the benefits of TUIs for discovery-based learning (Schneider, Wallace, Blikstein & Pea, 2013). A TUI is an interactive tabletop environment where physical objects are tagged with fiducial markers that are detected by a camera above the table. A projector then displays additional information on the tangibles, providing an "augmented reality layer" that provides students with additional information on the concepts they have to learn (e.g., labels, connectors, animations, description of the tangibles). In a series of controlled studies, we found that TUIs can be used advantageously in a discovery-learning situation when students approach an unfamiliar topic compared to standard "tell-and-practice" method. More specifically, we found that using a TUI before, rather than after, reading a textbook chapter or attending a lecture was beneficial. This first TUI, called "BrainExplorer" (Schneider, Wallace, Blikstein & Pea, 2013), is an interactive tabletop where users can explore the way the human brain processes visual information. Students take apart a physical replica of a brain while an augmented reality system displays visual pathways between brain regions. Users can then use an infrared pen to create lesions in the brain and observe the simulated impact of their actions on the visual field of the subject. In a controlled experiment, we showed that 1) students who used BrainExplorer outperformed students who read a textbook chapter on a learning test, and that 2) students who first used BrainExplorer and then read a textbook chapter outperformed students who completed the same activities but in the reverse order (text followed by BrainExplorer). Our conclusions were that TUIs supports students' elaboration of their own micro-theories and create an engaging point of entry for exploring a new domain. A corollary of this result was that "tell-and-practice" procedures are often not the best way to introduce students to new ideas, at least in the domains that we have explored.

In the study that generated the dataset considered in this paper, our goal was to explore this effect in more depth and disentangle confounding variables present in the study described above. One alternative interpretation of our previous results is that a "novelty" effect caused higher learning gains in our treatment groups (i.e., learning from an interactive tabletop environment, followed by a standard instruction): students may have been more engaged from the start due to the novelty of the interface, and this effect may have had a contagious effect on the rest of the activity. Similarly, a standard instruction is often considered too "school-like" and may have caused participants to disengage from the entire activity; this would have caused students to not fully take advantage of the interactive system afterwards. Thus, we designed the following experiment (Schneider & Blikstein, 2015) where students from a community college were recruited as part of a psychology class and asked to discover how the human hearing system works (N=38, average age = 22.5, SD= 6.2; 25 females, 12 males). Pairs of students worked on a tangible interface called EarExplorer (Fig. 1). EarExplorer is an

interactive tabletop environment with 3D-printed tangibles tagged with fiducial markers. A projector displays an augmented reality layer by reflecting its image on a mirror held above the tabletop. A camera is attached to the mirror and detects the location of the fiducials on the tangibles (Fig. 1). The starting screen displays three elements: the outer ear, which is the starting point of the activity (top left corner, Fig. 1); the auditory cortex, which is the end point of the activity (bottom right corner, Fig. 1); and an information box (bottom left corner, Fig. 1). Eight tangibles are arranged around the projected area. Students are asked to connect the tangibles between the starting point and the ending point to let sound waves reach the auditory cortex. Each tangible can be positioned in the information box at any time of the activity to display additional information about each organ. Users can use those hints to infer the correct sequence of tangibles and learn more facts about the function of each organ.

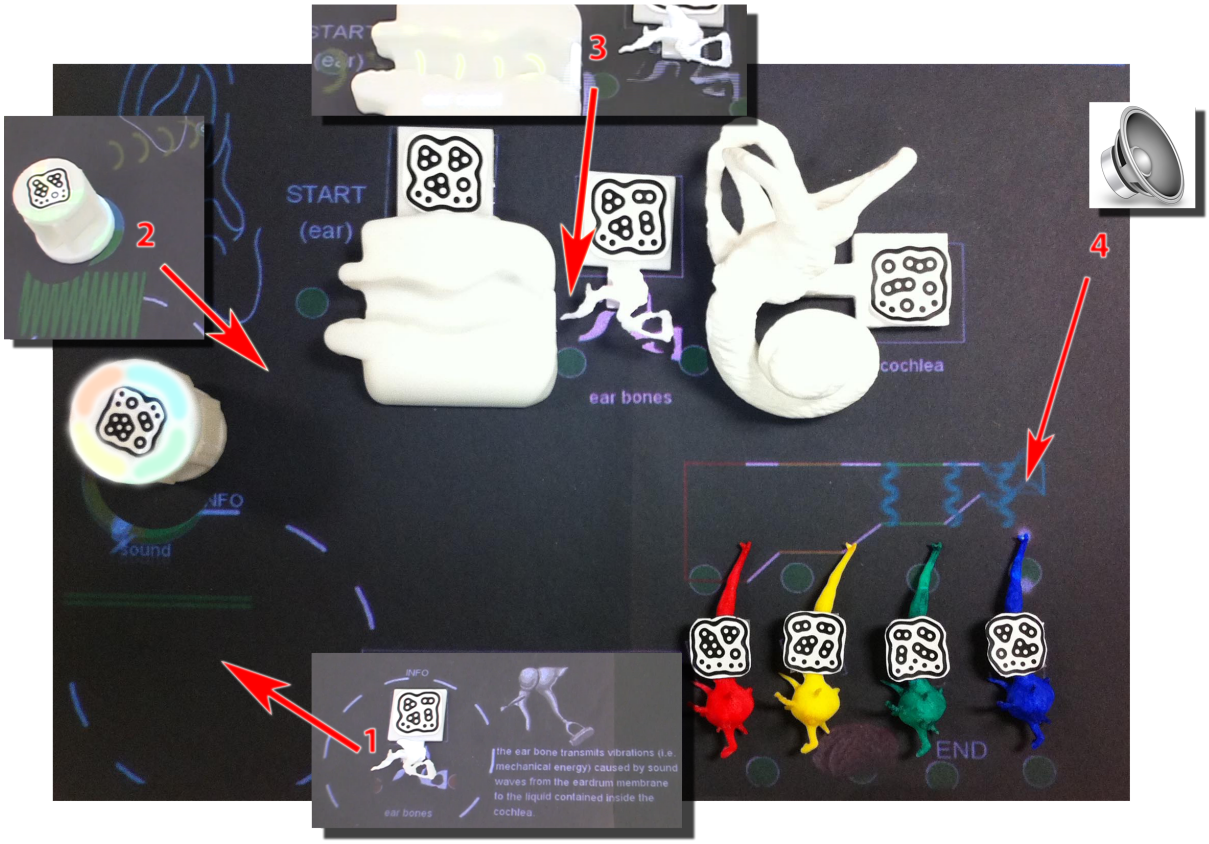

Figure 1: The EarExplorer Interface, after connecting the tangibles in the correct sequence. Students used the infobox (1) to learn about the different organs and connect them together; they then generated sounds at different frequencies with a speaker (2); sound waves travelled from the emitter through the ear canal to the ear bones (3); finally, the sound reached the basilar membrane inside the cochlea, activated a specific neuron and replayed the sound if the configuration is correct (4).

In one condition (N=18, labeled “discover”), students rebuilt the hearing system by trial and error, using resources provided by the system (Fig. 2, right side). In another condition (N=20, labeled “listen”),students followed a step-by-step recorded guidance of a professional instructional designer. The instructional designer was not aware of our research hypotheses

and was asked to make the learning material as engaging as possible (Fig. 2, left side; the red arrow points at the video). Students in this condition interacted with the TUI while they were watching the video; the video played continuously and had multiple breaks to let students experiment with the system. In this video, the instructional designer explained the function of each organ by positioning them in the information box and prompted the students to follow his instructions to rebuild the hearing system.

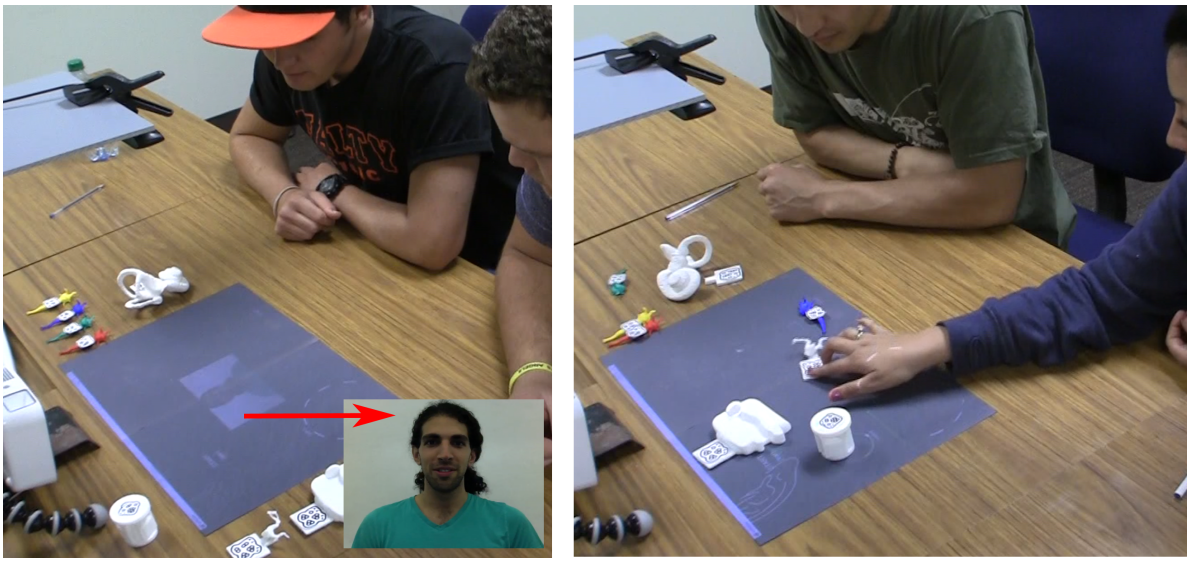

Figure 2: Students interacting with EarExplorer. The left picture shows students in the first condition ("listen") and the right picture shows the second condition ("discover"). Students in the "listen" condition followed a video tutorial, shown by a red arrow.

Students in both conditions first took the pre-test and then received a tutorial on how to use the TUI before starting the task; they spent the same amount of time interacting with the TUI (15 minutes). If they finished the task early, the experimenter asked them to keep exploring the system and try additional scenarios. After building the system, they took a mid test and then read a textbook chapter explaining sound transduction in a more formal way. Finally, students took a post-test measuring their learning gains. The learning test had six questions where students were asked to: 1) label the organs of the earing system; 2) describe various sound waves and asked which parts of the basilar membrane would vibrate at those frequencies; 3) compare the effect of various kinds of lesion (e.g. do broken ossicles have the same effect as piercing the eardrum?); 4) describe which part of the basilar membrane should be numbed to lose sensitivity to certain frequencies; 5) map the frequency range of various animals (bats, dogs, mice) inside their cochlea; 6) describe how sound is propagated from one organ to the other. Each learning test (pre, mid, post) had small variations in the questions. Students had 10 minutes to individually fill each test. Learning gains were computed by subtracting pre-test scores from the post-test.

We found that students in the first group achieved higher learning gains (Fig. 3). A MANOVA showed that participants in the "discover" group learned significantly more after the first activity: $F(1,35) = 22.11$, $p < 0.001$ and after the second activity $F(1,35) = 16.15$, $p < 0.001$ compared to the participants in the "listen" condition. Considering that the maximum score on the post-test was 15, students in the "discover" condition did not just keep their advantage from the middle-test—they were actually able to answer harder questions than students in the

"listen" condition. Those results suggest that TUIs are not a "silver bullet" for discovery learning activities. The potential "novelty effect" that they generate cannot explain the learning gains observed in previous studies. Rather, it seems that TUIs generate a sense of agency that is beneficial to learning. Given those results, our goal is to take a new look at this dataset by applying data mining and multimodal learning analytics techniques to students' body language. In the next sections, we analyze the logs generated by the tangible interface and by a Kinect™ sensor that recorded students' actions during the study.

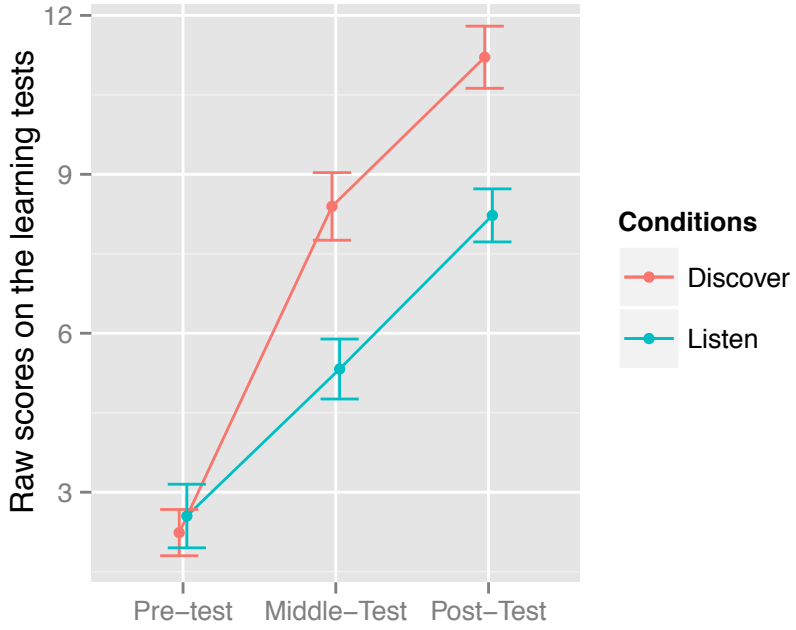

Figure 3: Results of the learning tests. The pre-test was administered at the beginning of the experiment; the middle test was completed after students interacted with EarExplorer; finally, they completed the post-test after reading a textbook chapter on the human hearing system.

3.2. DATASETS

We have collected three datasets from this experiment. The first one was manually created and contains basic information about the participants, such as their experimental condition, demographic data, GPA (Grade Point Average), learning gains, and field notes about their behavior. The second dataset was automatically generated by the tangible interface and describes students' actions when interacting with EarExplorer. There were 6 types of action: adding a tangible, removing a tangible, connecting two tangibles, generating a sound wave, playing a sound if a wave reached the brain and accessing the infobox. Table 1 shows an example of this type of log:

Table 1: Examples of actions recorded by the TUI (preceded by a UNIX timestamp) from different groups. Note: tangible #0 is the speaker, tangible #1 the ear canal, and so on.

Examples of lines recorded from TUI logs	**Explanation**
1371057396: started	Users can now interact with the interface
1371057519: fiducial added id=0	A user has added tangible #0 to the table
1371057630: fiducial removed id=0	A user has removed tangible #0

1371057632: Sound generated: 2.5363107	A user has generated a soundwave
1371057645: infobox: 2	A user has positioned tangible #2 on the info-box
1371079060: New connection: 5.0 to 9.0	A user has connected tangible #5 to tangible #9
1371079079: New wave: 4	A soundwave has reached tangible #4
1371079087: Wave success freq=4	A high-frequency soundwave has reached the brain

The last dataset was automatically collected using a Kinect™ sensor, which generated text files for each group of students. The Kinect generates 30 data points per second per user, and each data point contains the x,y,z information about 10 joints of a sitting user (head, torso, left/right shoulders, right/left elbows, right/left wrists, right/left hands). A line of a Kinect log looks like the example given in Table 2:

Table 2: Examples of the data collected by a Kinect Sensor for a sitting user.

Example of a line recorded by a Kinect sensor	**Explanation**
-0.302565, 0.0729301, 1.71089, -0.311148, 0.23276, 1.67768, -0.468162, -0.0548381, 1.76986, -0.299887, -0.0894132, 1.57217, -0.489155, -0.142518, 1.71705, -0.552245, -0.16022, 1.76821, -0.202123, -0.0309904, 1.64629, -0.223243, -0.199955, 1.55879, -0.323745, -0.0455265, 1.55587, -0.357246, 0.00594958, 1.55776, Wed Jun 12 17:10:16 2013	Comma-separated values of the x,y,z coordinates of a user's head, torso, left/right shoulders, right/left elbows, right/left wrists, right/left hands. The last field is a timestamp.

The first dataset was analyzed in a previous publication (Schneider & Blikstein, 2015). The second one (i.e., the logs from the TUI) is inspected in section 4.1. The last dataset (i.e., Logs generated by the Kinect sensor) is analyzed in section 4.2.

3.3. HYPOTHESES

The goal of this paper is to craft measures from the log files to further illuminate the results of our previous study (Schneider & Blikstein, 2015). More specifically, we are interested in isolating students' behaviors that resulted in higher learning gains and explain the difference found between our two experimental conditions. Our main hypothesis is that the amount as well as the quality of students' exploration when interacting with the TUI is correlated with their learning. We also follow an opportunistic approach and conduct additional analyses inspired from prior work that could be easily computed using the techniques we developed, more specifically on students' collaboration and coordination while interacting with the TUI.

We believe that the two datasets described above can offer complementary perspectives: the TUI logs are task-dependent and do not discriminate between participants, while the Kinect data is task-agnostic and offers fine-grained data about individuals' body posture and gestures.

We explore the following hypotheses in the sections below:

1) TUI logs: the *amount* of exploration (i.e., number of actions recorded by the TUI) is correlated with students' learning (section 4.1.1)
2) TUI logs: some particular *types* of exploration or sequences of actions (e.g., testing the system, accessing the info-box) are correlated with learning gains (section 4.1.2)

3) TUI logs: dyads (pairs of students) with high learning gains look more similar to each other in the way they explore the TUI, and look dissimilar to dyads with low learning gains (section 4.1.3)
4) Kinect logs: the *amount* of movements generated by each user is correlated with students' learning (section 4.2.1)
5) Kinect logs: some particular *types* of movements or sequences of gestures (e.g., hand movements, being in an active posture) are correlated with learning gains (section 4.2.2)

Hypotheses 1-5 are obvious first steps toward exploring the TUI and Kinect logs: we simply wanted to test whether the quantity and quality of students' exploration had a linear relationship with their learning. Hypotheses 6-8 acknowledge that the dyad's quality of collaboration had an impact on their learning and are formulated as follows:

6) Kinect logs: students' leadership can be detected from students' gestures and are associated with learning gains (section 4.3.1)
7) Kinect logs: the level of synchronization between the dyad's members is correlated with learning (section 4.3.2)
8) Kinect logs: the distance between two group members is a proxy for their level of comfort with their partner and the content taught (section 4.3.3)

Those last three measures are opportunistic, in the sense that we could easily extract them from our datasets and connect them to existing theories in social psychology: we merely computed students' bimanual coordination, body synchronization, and distance between participants as a simple proxy for the quality of their interactions and then tested their relationship to learning. Finally, in section 5 we put together all of the measures mentioned above and feed them into a machine-learning algorithm to see how well we can predict students' learning gains.

4. ANALYSES FROM INTERACTION LOGS

4.1. DATA FROM THE TANGIBLE INTERFACE

The logs generated by the tangible interface represent every action performed by the users. For instance, it records when two tangibles are connected or disconnected, when students consult the information box, and when they generate a (un)successful sound wave (for a description of the logs, see Table 1). Since those log files contain series of actions encoded as strings, we decided to use some information retrieval techniques (as suggested by previous work; Manning, Raghavan, & Schuetze, 2008), but applied to human actions ("n-actions"; Worsley & Blikstein, 2013). Our overarching goal is to describe methods for finding predictors of students' learning that are generalizable to other types of log files. In the section below, actions from the TUI logs are aggregated together (e.g., accessing the information box for the ear canal, the cochlea, or any other organ is labeled as the same action in our analyses).

4.1.1. Amount of Exploration (TUI logs)

Our first hypothesis is that the *amount* of exploration (i.e., number of actions recorded by the tangible interface) is correlated with learning. We tested this proposition by simply counting the number of actions performed by the dyad (in other words, the number of lines present in the TUI logs). We did not find a significant relationship between those two measures: $r(16) =$

0.208, p = 0.407, even though there seemed to be a positive relationship between number of actions and learning gains. There was also no significant difference between our two experimental groups: $F(1,17) = 1.43$, $p = 0.25$, Cohen's $d = 0.60$ (number of actions for the "discover" group: mean=4135.00, SD=1842.27; "listen" mean=3114.80, SD=1573.83). This suggests that merely manipulating objects in the TUI does not guarantee learning; rather, it is likely that certain types of actions are associated with higher learning gains. In the next section, we look at different types of explorative behaviors recorded by the TUI.

4.1.2. Types of Exploration (TUI logs)

Our second hypothesis is that some particular *types* of exploration or sequences of actions are correlated with learning. In order to test this hypothesis, we extracted the unigram, bigram, and trigram probabilities for each pair of students. Based on these n-grams, we looked at correlation coefficients between most frequent actions and learning gains. After selecting the three highest correlations, we tested them for significance. We found significant correlations between accessing the information box (see Fig. 1, bottom left corner) and learning gains using unigrams (number of times students accessed an info box), $r(17) = 0.47$, $p < 0.05$; bigrams (accessing the info box twice in a row) $r(17) = 0.50$, $p < 0.05$; and trigrams (accessing the info box three times) $r(17)$: 0.459, $p < 0.05$. This suggests that students who used the information box to solve the problem at hand were more likely to learn more. Additionally, students in the "discover" condition were more likely to access the info-box: $F(1,16) = 3.40$, $p = 0.08$, Cohen's $d = 0.98$ (discover mean=14.22, SD=15.76; listen mean=3.25, SD=1.20). Even though this difference is not significant, it suggests that students in the two experimental conditions behaved differently: the participants in the "discover" group may have learned more because they spent more time learning about the organs of the human system through the information box. A tentative explanation is that, because there was not a teacher to tell them how the system worked, they became more curious and freely decided how to rebuild the human hearing system.

These results might have applicability in new forms of assessment. Schwartz & Arena (2013) are currently developing a new kind of assessment called choice-based assessment, where students' choices are central to evaluating their learning trajectories. In a previous study, they found that the best predictor to students' success was not necessarily their outcomes of solving particular problems, but the extent to which they chose to access additional resources during a learning activity. For those researchers, this kind of choice is a strong feature of adaptive expertise (Hatano & Inagaki, 1986), a type of attitude where learners are able to solve previously encountered problems in an efficient way and generate new procedures for novel tasks. It is contrasted with "procedural expertise," where students learn to perfectly master cognitive or behavioral procedures but cannot transfer them to other contexts. The results above are to some extent supported by Schwartz's theory and can be considered interesting predictors for students' learning. In our case, and in general when building learning environment, it suggests that giving the *choice* to exploit additional resources is a powerful predictor for students' learning.

4.1.3. Comparing Students' Styles of Exploration

Our third hypothesis is that dyads with high learning gains look similar to each other in the way they explore the TUI and look dissimilar to dyads with low learning gains. We tested this hypothesis by computing document similarity metrics on our logs and comparing the set of actions in a session between pairs of students. More specifically, we followed the procedure

described by Manning, Raghavan, and Schuetze (2008) and transformed our logs into a term-frequencies matrix. In this matrix, each column is the count of a specific action (e.g., accessing the information box or connecting two tangibles) and each row corresponds to a group of students. We then applied a tf-idf (term frequency–inverse document frequency) transformation to the matrix to dampen the effect of common, non-informative actions and to increase the importance of rare terms. The resulting matrix contains vectors of probabilities between groups of students (rows) and each potential action (columns); two row vectors can then be compared to assess the similarity of two dyads. A common technique developed for this purpose is to compute the cosine similarity between two vectors: the result simply describes the amplitude of the angle between those vectors in a high dimensional space. Figure 4 shows all pairwise comparisons.

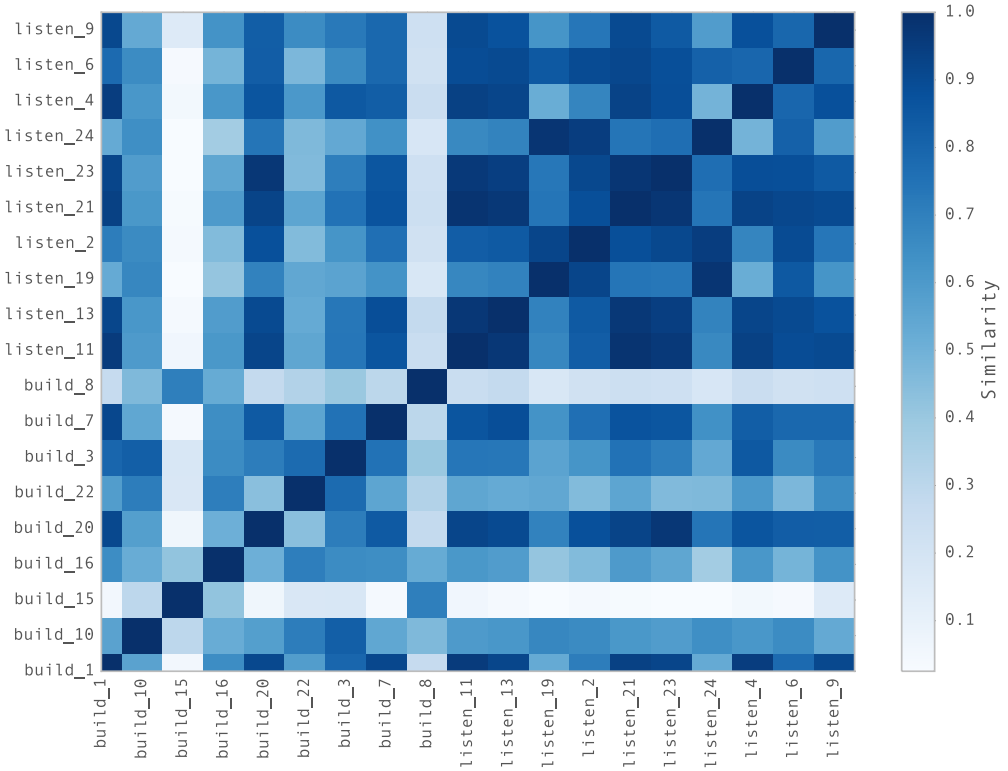

Figure 4: Cosine similarity matrix between pairs of students (computed on the logs of actions generated by the tangible interface). Light blue shows low similarity and dark blue high similarity (values range from 0 to 1, respectively).

Given the matrix of cosine similarities, the two most distinctive pairs of students from the others are group 8 and group 15 (the blue lines in Fig. 4). Interestingly, those dyads achieved the highest learning gains in our experiment (10 and 10.5, respectively). We can further observe that group 15 is very different from everyone else, because their cosine similarity scores are very low (shown in light blue in Fig. 4). Group 8, on the other hand, represents a larger spectrum of similarity: some groups are pretty similar to them (e.g., the darker shades of 3, 10, 15, 16) while others are more distinct (e.g., the darker colors). The following question naturally arises: can we use this metric as a predictor for learning? Previous studies have tried this idea on students' essays and utterances, for instance by clustering passages together to

isolate “themes” (Sherin, 2012), and have found this approach to be relatively successful in identifying misconceptions among students. In our dataset, we indeed found a significant correlation between students’ learning gains and their similarity to group 8: r(17), 0.53, $p < 0.05$ (Fig. 5). Group 8 lend itself much better as a comparison than group 15 for two reasons: 1) there is a larger spectrum of similarity scores to group 8 (more diverse shades of blue); and 2) group 15 had five times more events than all the other groups in their log file. Group 8, on the other hand, had a comparable number of actions with the other dyads. We did not find a significant correlation between students’ learning gains and their cosine similarity with group 15: $r(17) = 0.42$, $p = 0.072$, or with the only other group that achieved a learning gain higher than 10 (group 20, gain = 10.25): $r(17) = -0.111$, $p = 0.650$.

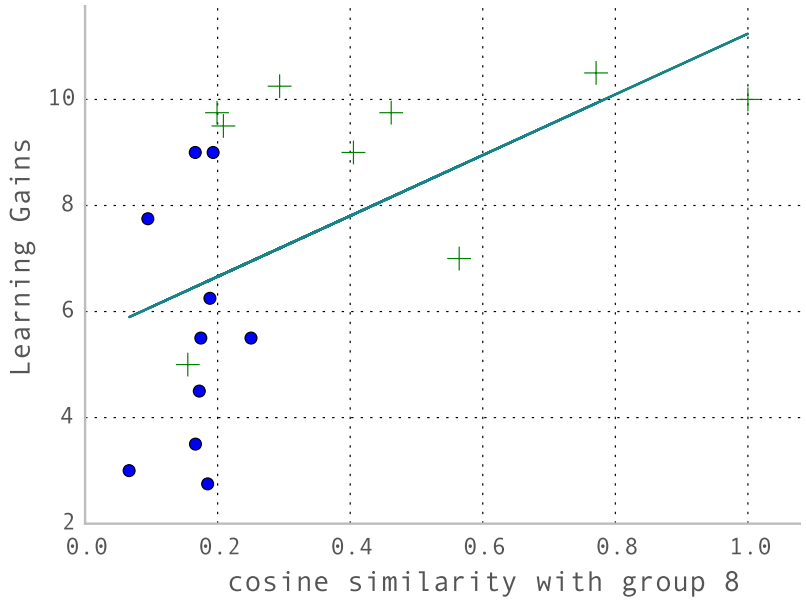

Figure 5: Scatterplot of the cosine similarity score between each pair of students with group 8 and their learning gain. Blue dots represent students in the “listen” condition and green crosses represent students in the “discover” condition.

The correlation shown in Figure 5 is not perfect: many students are actually quite different from group 8 (below 0.2 similarity on the x axis) and represent a wide spectrum of learning scores. Additionally, the students in the “discover” condition were significantly more similar to group 8: $F(1,18) = 9.67$, $p = 0.01$, Cohen's $d = 0.77$ (build mean=0.45, SD=0.27; listen mean=0.17, SD=0.05). This reinforces the fact that our experimental manipulation had a strong effect on students’ behaviors and that those differences seem to be responsible for differences in learning gains.

In summary, out-of-the-box information retrieval techniques (n-grams probabilities and cosine similarity metrics) can provide us with relatively good predictors for learning gains and for distinguishing between our two experimental conditions. From the logs collected by the tangible interface, we found that both “n-actions” associated with accessing the information box and students’ similarity with the highest scoring group on the learning test were associated with positive learning gains. This suggests that logs collected when students interact with a tangible interface have some predictive values for discriminating between proficient and non-proficient students (Worsley & Blikstein, 2013) and that our experimental manipulation had a

positive effect on productive learning behaviors. It should be noted that the methods we used are general enough to be applied to any other type of log files. In the next section, we switch our attention to the Kinect data and describe the various measures we extracted from this dataset.

4.2. KINECT DATA

In this section, we describe various measures that we extracted from the Kinect[1] logs. More specifically, we looked at: 1) the amount of movements generated by the students and the Kinect's use as a proxy for engagement; 2) prototypical body postures and students' likelihood to transition between them; 3) student's bimanual coordination and its relationship to group dynamics; 4) dyads synchronization and proxemics in small group collaboration. We predict that several of those measures should be related to learning, at least through an indirect effect via students' engagement, quality of collaboration, and cognitive states.

4.2.1 Amount of Movements (Kinect logs)

Our fourth hypothesis is that the *amount* of exploration captured by the Kinect sensor is correlated with learning. We tested this hypothesis by computing the amount of movement generated by each participant. We believe that more engaged students move their bodies to a greater extent than less engaged ones, and that the amount of physical movement is a proxy for general engagement; this engagement in turn is related to learning gains. There are two ways to compute this metric: the first is by calculating the Euclidean distance between each tracked joint in the student's body and averaging the result over time. This approach is not ideal, because it does not take into account the natural variations in limbs' lengths. An arguably better way to compute movements is to look at variations in angles between joints in body positions. We tried both approaches and sliced the data over time to get a measure for each minute. We also computed an overall score, as well as a score for each joint. We did not find any significant correlation between the measures described in this paragraph and learning gains. For instance the amount of movement computed with joint angles produced the following correlation: $r(34) = 0.079$, $p = 0.648$. On the one hand, this result is somewhat surprising: we would expect at least some of those measures to be associated with higher engagement and thus more learning. On the other hand, a movement of the hand can mean a range of different things (e.g., a sign of boredom, interest, a deictic gesture, and so on), so ultimately the results make sense. Many simple gestures are ambiguous by nature, and in our particular case we did not have enough information to correctly contextualize them. It should be noted, however, that our non-significant results do not mean that students' amount of movement is not an interesting measure; it merely shows that our simple approach was not able to isolate a signal from the noise in our data.

In the next sections, we look at more refined measures of students' movements: more specifically, we will look at hand coordination and will show how to detect prototypical body positions using unsupervised learning algorithms.

4.2.2 Types of Exploration (Kinect logs)

[1] The Kinect sensor is a motion sensing input device developed by Microsoft. It captures the skeleton (composed of 20 joints) of up to four users at a frequency of 30Hz.

Our fifth hypothesis is that some *types* of exploration or sequences of gestures are correlated with learning. In this section we describe how we used clustering algorithms to automate the creation and application of a coding scheme on body postures. Recall that most previous work in this area was conducted manually by analyzing videos frame by frame in a highly time-consuming process. If we can show that an algorithm can accomplish a similar task, it will provide researchers with a more efficient way to quickly analyze students' body language.

Our approach was to take our entire dataset (1 million entries; i.e., one entry is a line recorded by the Kinect sensor) and transform it into (joint) angles instead of positions in a three-dimensional Cartesian coordinate system. We then fed this matrix into a simple and fast clustering algorithm (K-means) that provided us with prototypical body positions. As a first step, we generated 2 to 9 clusters and visually inspected the results; we decided to keep three clusters, because the postures found were all perceptibly different, relatively easy to interpret, and there was no overlap between them. The results are shown in Figure 6.

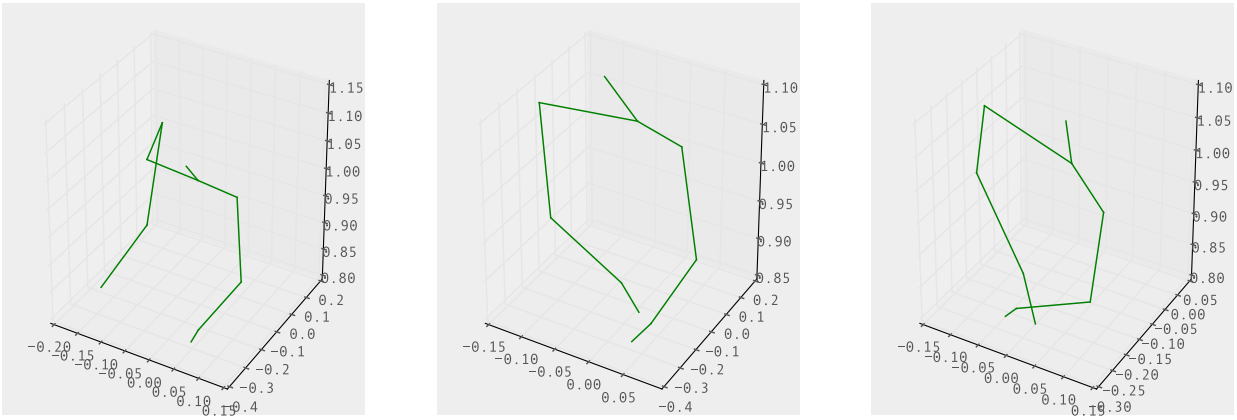

Figure 6: The results of the k-means algorithm on students' body posture (1 million data points). In this particular case, we used angles between joints instead of the standard skeleton joints provided by the Kinect sensor. The first state (left) is active, with both hands on the table; the second one (middle) is passive, with both arms crossed; the third one (right) is semi-active, with only one hand on the table.

We found the three clusters to have interesting properties. The first one (left) represents an "active" position: both arms are on the table, supposedly manipulating something or at least ready to act; the head is tilted toward the table in an attentive position. The second cluster (right) shows a "semi-active" posture: one arm is flexed, while the other one is straight on the table, probably manipulating a tangible. The last one (middle) represents a "passive" posture, where arms are both crossed and the body looks relaxed. We then used those three clusters to classify each data point into one of them based on proximity to cluster centroids and counted how many times each student was in each posture.

The way we interpret those three clusters seems to correlate with statistical measures: the "active" posture is positively associated with students' learning gains $r(34) = 0.329$, $p < 0.05$ while the "passive" one is negatively correlated with students learning gains $r(34) = -0.420$, $p < 0.05$. Additionally, we found that the number of times students *transitioned* from one posture to another was also significantly correlated with their learning gains: $r(34) = 0.335$, $p < 0.05$. Students in the "discover" condition also spent more time in the first cluster: $F(1,37) = 4.52$, $p = 0.04$ and made more transitions between clusters: $F(1,37) = 4.42$, $p = 0.04$. This

suggests not only that some postures are indicative of learning, but certain sequences of postures are meaningful predictors of learning too (which were promoted by our experimental manipulation). Previous work (Tschan, 2002) has shown that "ideal" cycles of cognition (i.e., planning, executing, and evaluating an action) are usually associated with higher performances and higher learning gains. It is possible that the results of our clustering algorithm produced a similar construct: an increased number of cycles where students think for a while (posture 1 and 2) and then execute an action (posture 3) could be interpreted as something akin to an ideal cycle of cognition described by Tschan. We also provide a graphical representation of those results (Fig. 7): three Markov Chains show the average, best, and worst students in terms of their learning gains. In those Markov Chains, node sizes represent the amount of time spent in a particular state; arrows represent the transition probability between those states (ignoring self-loops).

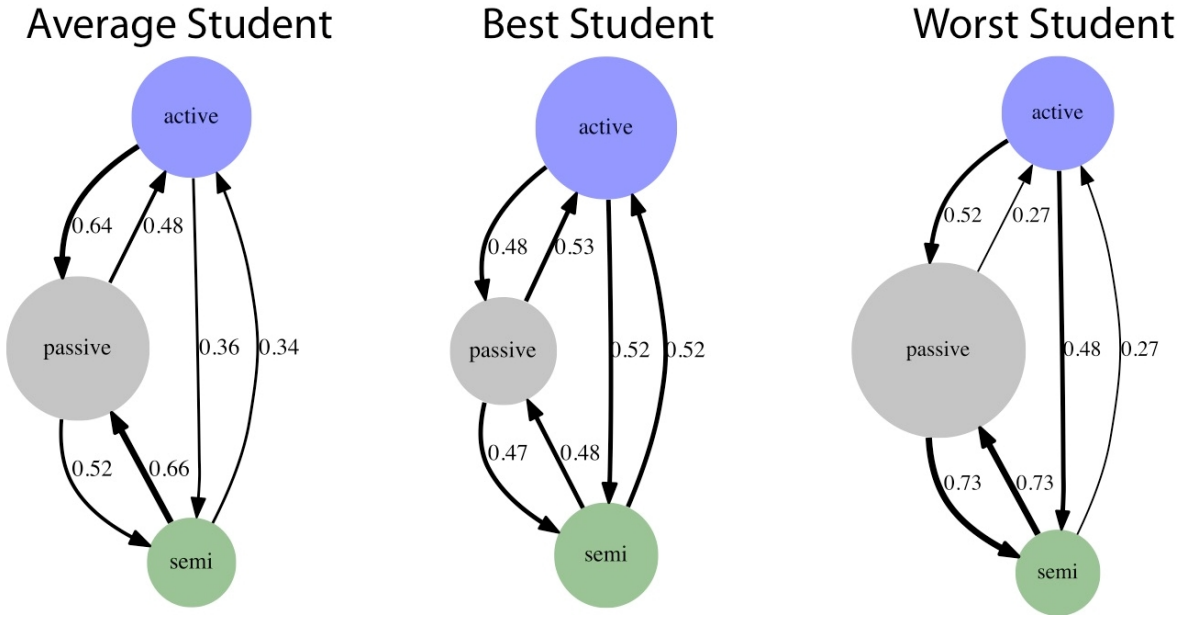

Figure 7: Markov Chains of the three states described above (active, semi-active, passive). The first shows the average student, the second and third one show the best and worst students in terms of their learning gains.

We can observe that the worst student stays mostly in a passive state; additionally, when (s)he enters a passive or semi-passive state, this student is extremely likely to keep looping between these non-productive states (from the transition probabilities, we can see that there is a 73% chance of being "trapped" in a non-active state). The average student is less likely to stay in a passive state; when entering the semi or passive state, the probability of staying there drops to ~60%. Finally, the best student displays the most balanced transitions probabilities: when (s)he is a in a particular state, there is an equal chance of shifting to an active, semi-passive and passive state. As a comparison, the transition probability of staying in a non-active state is lower than 50% (i.e., 47% and 48%). This suggests that we can potentially discriminate between high and low proficient students by looking at their transition probabilities as they go through the activity.

We should mention that we tried several approaches before finding the optimal way to cluster our dataset. We first tried to use joint *positions* in a three-dimensional space, as measured by the Kinect (i.e., the x, y, z coordinates of each joint of the Kinect skeleton: head, neck,

shoulders, elbows, arms). We found two main issues with this method: first, clusters were influenced by students' orientation toward the tangible interface (right or left side). Second, the size of their limbs (and body in general) interfered with the clustering algorithms. Longer limbs were more likely to be clustered together, and the same would happen for shorter people (shorter limbs clustered together). Finally, we also used another clustering technique (hierarchical clustering) and obtained comparable results.

4.3. ANALYSES AT THE DYAD LEVEL

We describe here additional results conducted at the dyad level, i.e., when taking both bodies into consideration. The goal is to provide new insights on collaborative learning processes by looking at body coordination when interacting with a peer.

4.3.1 Leadership Behaviors

Our sixth hypothesis is that students' leadership can be detected from students' gestures and are associated with the dyad's learning gains. In a related study, Worsley & Blikstein (2013) have shown that bimanual coordination was predictive of participants' expertise in solving an engineering problem. Based on these results, we decided to compute a similar metric for our dataset. More specifically, the idea is to compute and compare the amount of movement generated by each hand. Figure 8 shows all the graphs generated by this approach: some students barely use their left arm while others use both arms during the entire activity. It is interesting to see the variety of the graphs produced; all the students have a very distinct signature in terms of their hand movements.

To make sense of this metric, we need to introduce additional results that we found in the initial study (Schneider & Blikstein, 2015). Previous research has shown that each student working in groups can often be categorized as either being the "driver" or the "passenger" of the interaction (Shaer, Strait, Valdes, Feng, Lintz & Wang, 2011). One coder used several indicators to categorize each dyad's members: 1) who started the discussion when the experimenter leaves, 2) who spoke most, 3) who managed turn-taking (e.g., by asking "what do you think?"; "how do you understand this part of the diagram?"), and 4) who decides the next focus of attention (e.g., "so to summarize, our answers are..."; "I think we need to spend more time on this…").

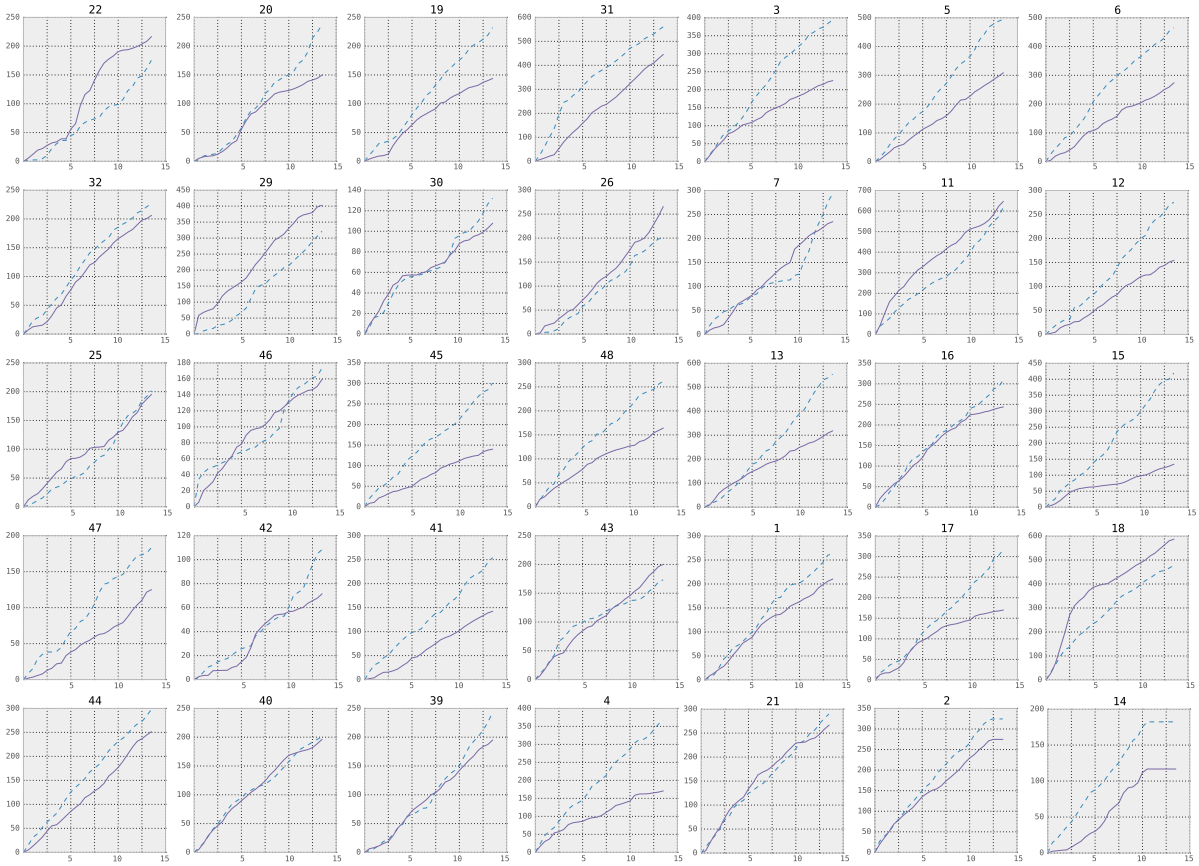

Figure 8: Hand coordination of 18 dyads (group members are shown next to each other). Dotted line represents the amount of movement generated by the right hand, and solid line represents the left hand. X-axes show time (minutes) and y-axes represent the amount of movement generated by each hand. We can see that participant #39 is bimanual and uses both hands. Participant #3 uses predominantly his right hand (dotted line).

This measure can be considered as an aggregate estimation over the whole activity of the dyad's dynamic profile. We acknowledge that subjects are likely to shift roles while working together. We also recognize that this categorization is more likely to be a continuum, and that in a few cases the difference between drivers and passengers may be subtle. Nevertheless, we decided to take the approach of classifying students in a binary way for the entire activity to simplify our dataset and present clearer results.

In our case, after making this distinction for students, we further separated them by computing a median-split on their GPA (note: there was no significant difference in terms of students' GPA between the two experimental conditions: $F(1,35) = 1.74$, $p = 0.20$, Cohen's $d = 0.34$, "discover" mean=3.52, SD=0.59; "listen" mean=3.28, SD=0.49). This resulted in four categories: a student could either be a driver or a passenger with a high or low GPA. Figure 9 shows the boxplots of each category according to their learning gains. Surprisingly, having a high GPA driver in the group does not lead to higher learning gains: $F(1,16) = 0.04$, $p = 0.84$, Cohen's $d = 0.17$ (low GPA driver: mean=7.63, SD=1.84; high GPA driver: mean=7.87, SD=2.57). On the other hand, having a passenger with a high GPA does lead to increased learning gains: $F(1,18) = 3.51$, $p = 0.08$, Cohen's $d = 1.4$ (low GPA passenger: mean=6.22, SD=2.26; high GPA passenger: mean=8.36, SD=2.43). From our observations running the experiment, this result is not totally unexpected: proficient students who do not "take control"

of the activity tended to leave more space for trial and error for their partner and suggested hints when needed. This situation resulted in increased participation and engagement from the low GPA student. In the opposite situation, the same student would stay passive and let the driver solve the problem on her/his own.

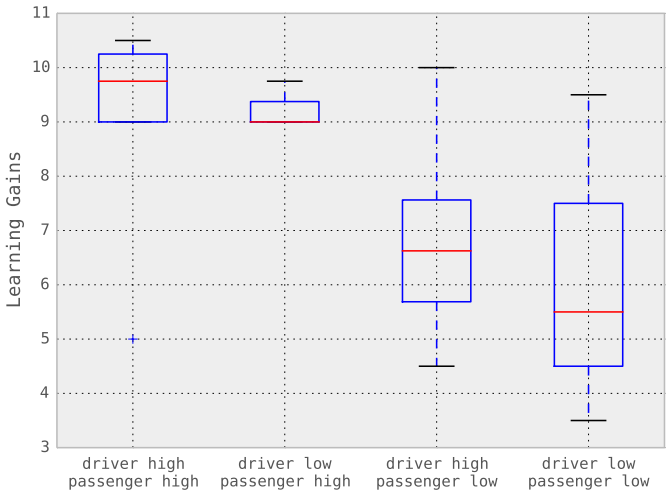

Figure 9: Boxplots of the four kinds of dyads described above: driver/passenger with high/low GPA. Y-axis shows the averaged learning gains of the dyads.

The distinction between proficient/non-proficient drivers/passengers allowed us to find interesting patterns in our data. More specifically, we found that drivers tend to use both hands while the passenger uses at most one hand. Figure 10 show the aggregated evaluation over time of the hand movements of those two types of students. Using an ANOVA, we found a significant difference between the amount of movements of each hand for the passengers at the end of the building activity: $F(1,35) = 7.66$, $p = 0.01$ (left hand mean=280.00, SD=86.30; right hand mean=205.55, SD=69.62). This difference was not significant for the drivers: $F(1,35) = 1.24$, $p = 0.27$ (left hand mean=315.38, SD=152.32; right hand mean=257.21, SD=152.27). The p-value for the passenger became marginally significant after 14 minutes: $F(1,35) = 3.18$, $p = 0.08$ and significant after 20 minutes into the activity: $F(1,35) = 4.50$, $p = 0.04$ (Fig. 10).

This result shows that we can potentially differentiate between drivers and passengers by looking at their hand movements. As a possible implication of this result, we can imagine future systems where machine-learning algorithms will make predictions about the "status" of each member of a dyad. Using many more features, we can imagine a learning environment where personalized scaffolding is provided depending on the groups' dynamic: proficient leaders can be encouraged to take a more passive role, while less proficient students would be provided with more scaffoldings and more opportunities to participate (for a similar application in displaying the amount of speech produced by each member of a group, see Bachour, Kaplan, & Dillenbourg, 2008).

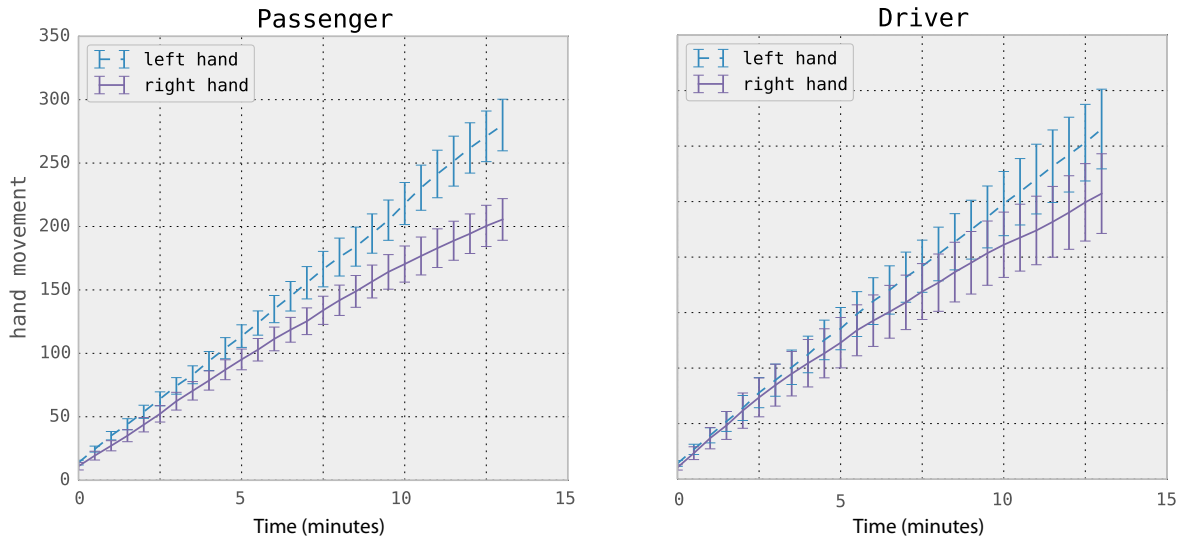

Figure 9: Bimanual coordination from Drivers and Passengers in dyads of students. X-axes show time (in minutes) and y-axes show the amount of movement generated by each hand.

4.3.2 Body Synchronization

Our seventh hypothesis is that levels of synchronization between the dyad's members are correlated with learning. In a previous study, Schneider & Pea (2013) found that students' visual synchronization (as measured by eye-trackers) was correlated with their learning gains. That is, more moments of joint attention was beneficial to establishing a common ground, which in turn positively influenced how much students learned during an activity. Other lines of research (in ethology as well as in human psychology; Chartrand & Bargh, 1999) suggest that body synchronization is associated with more productive collaborations. We were inspired by those results and decided to compute a metric for gesture synchronization using the Kinect data.

Our approach was to first take pairs of data points (one from each student) and compute the distance between them. Distance was calculated by taking the absolute value of the difference between the joint angles of each participant. Those differences were then averaged for each time point. We created graphs with time series of those data points as well as an overall measure of body synchronization. An ANOVA did not reveal any significant effect of this measure on our experimental manipulation: $F(1,17) = 0.92$, $p = 0.35$, Cohen's $d = 0.14$ ("discover" mean=0.46, SD=0.09; "listen" mean=0.42, SD=0.05). We also did not find a significant correlation between body synchronization and learning gains: $r(16) = 0.189$, $p = 0.453$. We thus conducted a second attempt that was inspired from the literature in eye-tracking studies (Richardson & Dale, 2005): it usually takes +/- 2 seconds for participants in a collaborative situation to adjust their gaze to their partner's behavior. It is possible that body language obeys the same rules. Thus, we repeated the procedure above, but this time, for each data point we looked at the minimum distance in their partner body posture +/- 2 seconds. The results were not influenced by this manipulation: $F(1,17) = 0.81$, $p = 0.38$, Cohen's $d = 0.13$ ("discover" mean=0.43, SD=0.09; "listen" mean=0.40, SD=0.05). Similarly the correlation with students' learning gains did not reach significance: $r(16) = 0.184$, $p = 0.466$ (Fig. 11). It suggests that even though gaze synchronization is a strong predictor for students' quality of collaboration and learning, body synchronization does not hold the same properties, at least in

the context of our experiment. Successful students were not more likely to coordinate their action based on their partner's behavior.

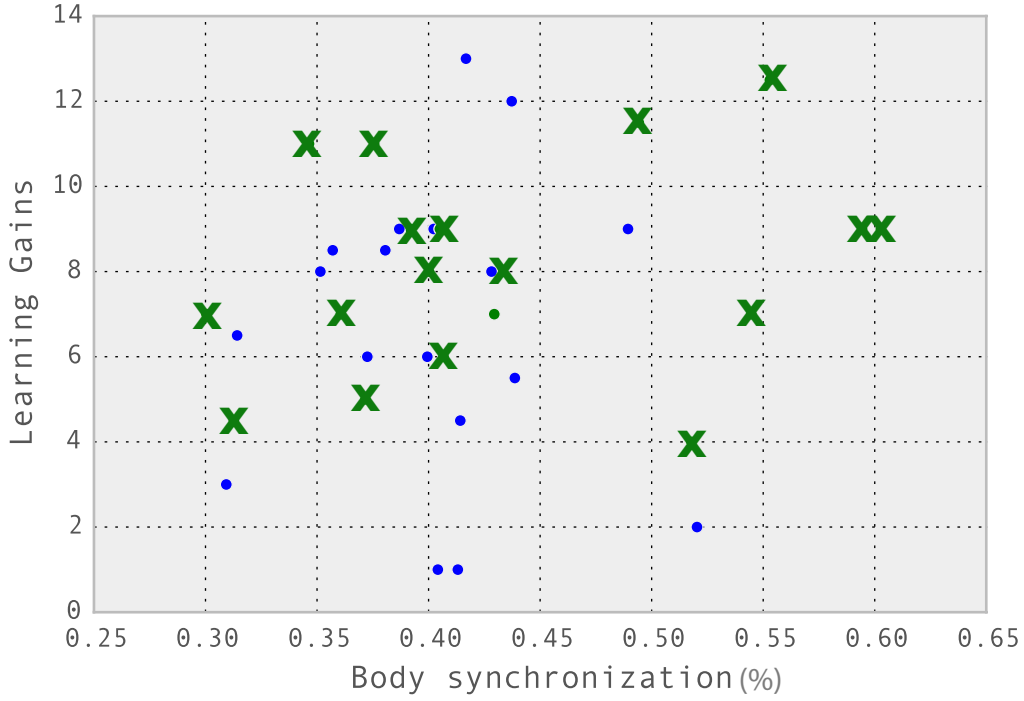

Figure 11: Relationship between body synchronization (average distance between students' bodies over time) and learning gains. Blue dots represent students in the "listen" condition and green crosses represent students in the "discover" condition.

4.3.3 Body Distance

Our eighth and last hypothesis is that the distance between two group members is a proxy for their level of comfort with their partner and the content taught. This metric was inspired by the theory of Proxemics developed by Edward T. Hall (1966). In this seminal work, he divided the distances around a person into different zones: the intimate area (less than 15 cm to 46 cm), the personal space (46 to 122 cm), the social distance (122 to 370 cm), and the public distance (370 to 760cm or more). Interestingly, in our study, students were seated at a distance that varied between the intimate and personal distance. Moving from a personal to an intimate distance is considered a violation of someone's territory if there is not implicit agreement that someone can do so. Thus, a small distance between two students can potentially characterize a productive collaboration and thus higher learning gains. Similarly, a larger distance can be an indicator of a poor collaboration.

We computed the distance between students 30 times per second by taking the rightmost joint from the student on the left side and the leftmost joint from the student on the right side of the table; we then calculated the Euclidean distance between those two points and averaged a global score for the entire activity (27000 data points). We did not find a correlation between learning and the distance between students' bodies: $r(16) = 0.377$, $p = 0.123$.

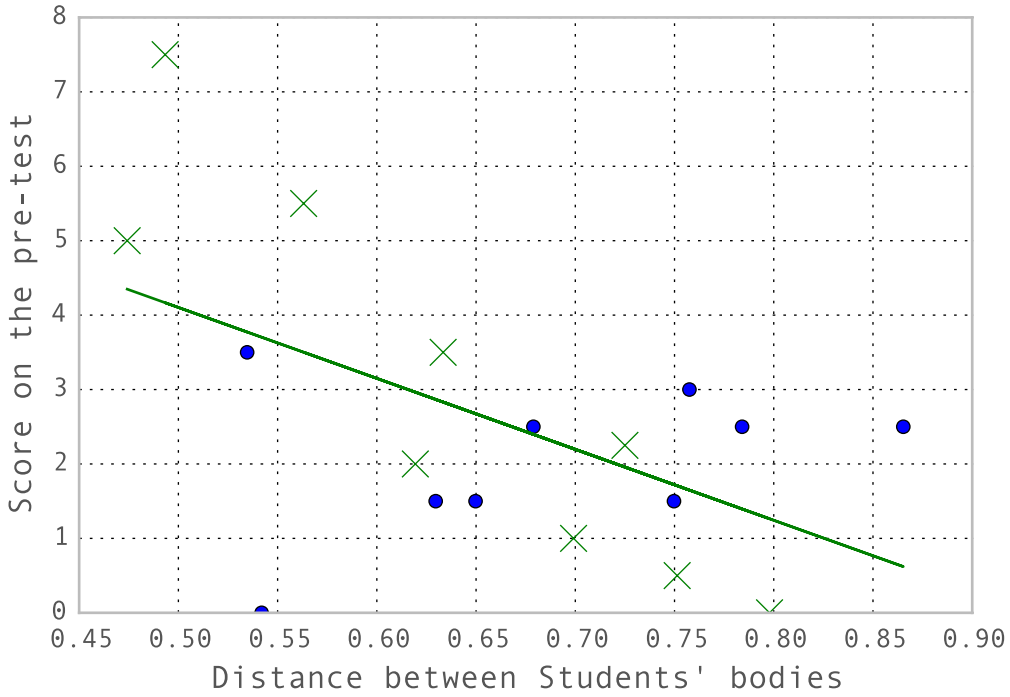

Figure 12: correlation between the distance between students' bodies and their pre-existing knowledge on the topic. Blue dots represent students in the "listen" condition and green crosses represent students in the "discover" condition.

However we found that this metric was correlated with students' pre-existing knowledge on the topic taught (i.e., score on the pre-test): $r(16) = -0.548$, $p = 0.019$ (Fig. 12). There was also no significant difference between our two experimental groups: $F(1,17) = 0.45$, $p = 0.51$, Cohen's $d = 0.12$ ("discover" mean=0.68, SD=0.09; "listen" mean=0.65, SD=0.12). While there could be multiple interpretations of this result, it suggests that students who are unfamiliar and maybe uncomfortable with the subject matter tend to establish a larger distance with their peers and possibly be more defensive during a collaborative task.

5. SUPERVISED MACHINE LEARNING

To conclude our exploration of this dataset, we decided to gather all the features mentioned above and store them into a single data frame. We then used a supervised machine-learning algorithm to see if our measures had any predictive value regarding students' learning gains. We used a median split on the learning gains to separate our students into two groups: those who fully took advantage of the activity and learned more than the other half of the students, and those who were below this cutoff. We decided to use a common and standard machine-learning algorithm to predict which student belonged to which category: Support Vector Machine (SVM). One advantage of SVM is that several kernels can be tested, which increases the chance of finding structure in a given dataset. We used a Leave-One-Out Cross-Validation procedure (LOOCV) during training. Due to the relatively small size of our dataset, we were not able to control for some structure in our data (conditions and pairs); we plan to use a larger sample in future work to control for this bias. The reader should keep in mind this limitation when looking at the results below. We selected features by starting with an empty set and

progressively chose the best features until we reached a ceiling or until the features were exhausted (i.e., forward-search feature selection). Our results are summarized in Table 3.

In terms of the features, we created the following categories: first, we used the raw and aggregated counts from the logs provided by the tangible interfaces: for instance, the number of times the information box is accessed, how many sound waves are created, how many tangibles are being used (7*2 features: tangible added, removed, connected; sound wave created, successfully reached the next organ; sound played; info-box accessed). Second, we have the cosine similarity matrix described above showing how similar groups of students are to each other (19 features). Third, we used movement data from the Kinect about joint angles: left/right shoulder, left/right elbow, left/right wrist, head and total amount of movement (8 features). Fourth, we used the unsupervised learning procedure described in section 4.2.3, created 9 clusters to provide us with additional features and counted the number of times each student spent in each position. Finally, we used the measure about body synchronization mentioned above (with and without a lag of +/- 2 sec) and the measure about body distance (n=3). This gives us 53 features in total, knowing that features at the dyad-level (such as body synchronization) are identical for each member of a dyad.

Table 3: Results of our SVM classifiers predicting students' learning on a median split. RBF stands for Radial Basis Function, and MLP for Multilayer Perceptron.

Kernel	Number of features	Top 5 features	Accuracy
Linear	7	cosine_21, infobox, successful_wave, posture_8, new_connection,	91.67%
quadratic	4	distance_bodies, tangible_added, new_wave, mov_left_wrist	91.67%
polyno-mial	8	cosine_9, distance_bodies, infobox, body_sync, body_sync_lag	91.67%
RBF	3	cosine_23, tangible_added, new_sound_wave	88.89%
MLP	6	cosine_9, posture_7, infobox, cosine_10, new_sound_wave,	100%

We achieved a 100% accuracy using a Multilayer perceptron kernel; in other words, the model could predict with 100% accuracy whether a particular student was above or below the median split computed on learning gains. This result is surprisingly high, but we should remind the reader that we are doing a very rough binary prediction (i.e., median split on learning gains) and that our dataset is relatively small (N=38). Even though we tried to prevent over-fitting by using a LOOCV procedure, it is inevitable that the model is probably mimicking our data too closely. Nevertheless, those results are promising if they can be replicated in other settings with more students.

In terms of the top features selected by each kernel, we can see that the cosine similarity metric extracted from the logs of the tangible interface is often chosen (e.g., "cosine_21" shows how similar each dyad is with group #21). The number of times students accessed the information box ("infobox") is also a strong feature for our classifier, which intuitively makes sense given the results highlighted in Section 3.1. Finally, it is interesting to see that several measures extracted from the Kinect sensor are being selected: the distance between students' bodies ("distance_bodies"), the synchronization between students' bodies ("sync") and the number of times students adopted a particular posture ("posture_7", for the cluster number seven identified by k-means). It shows that even if particular features are not significantly

correlated with students' learning gains, they still retain some predictive value for our classifiers.

6. DISCUSSION

In this paper, we presented several metrics for predicting students' learning around an interactive tabletop system. We provide a summary of our hypotheses and results in Table 4.

Table 4: summary of our hypotheses and results (X means that the hypothesis is rejected, V means that the hypothesis is not rejected)

	Hypotheses	**Results**
1	TUI logs: the *amount* of exploration (i.e., number of actions recorded by the TUI) is correlated with students' learning	X
2	TUI logs: some particular *types* of exploration or sequences of actions (e.g., testing the system, accessing the info-box) are correlated with learning gains	Students in the "discover" condition were more likely to display those behaviors.
3	TUI logs: dyads with high learning gains look more similar to each other in the way they explore the TUI, and look dissimilar to dyads with low learning gains	Students in each experimental condition were more likely to look like each other.
4	Kinect logs: the *amount* of exploration (i.e., the amount of movements generated by each user) is correlated with students' learning	X
5	Kinect logs: some particular *types* of exploration or sequences of gestures (e.g., hand movements, being in an active posture) are correlated with learning gains	Students in the "discover" condition were more likely to display those behaviors.
6	Kinect logs: students' leadership can be detected from students' gestures and are associated with learning gains	V
7	Kinect logs: the level of synchronization between the dyad's members is correlated with learning	X
8	Kinect logs: the distance between two group members is a proxy for their level of comfort with their partner and the content taught	V

First, we showed that information retrieval techniques could be used on the system's logs to predict learning gains as measured by pre and post-tests. More specifically, we found that particular *types* of exploration (e.g., consulting the information box) were associated with higher learning gains, while the amount of exploration was not. This suggests that successful groups of students were more likely to explore some very specific facets of the TUI (for instance, by reading more about the organs of the auditory system) as opposed to exploring all possible combinations of the tangibles. While this result is not surprising, it shows that we can extract useful indicators of students' learning from the logs of the TUI. We also used the best group of students as a reference point and computed cosine similarity metrics to obtain the similarity of each group with this baseline. Of course, we do not suggest that there is only one

single best way of learning, but in the context of our activity there seemed to be particular behaviors that were likely to be associated with higher learning gains (e.g., exploiting additional resources provided by the information box). In our case, we found that this measure was significantly correlated with learning. Also, the actions and behaviors of the best group emerged from the data itself and were not results of a mandated script, so examining the data from the best groups is arguably a good reference point of efficient or productive sets of actions and trajectories. Generically, with a larger group of students and more types of cognitive and non-cognitive measures, we could create a variety of productive trajectories that would work for different profiles of students (see Blikstein & al, 2014).

Second, we explored how Kinect data can inform the way we understand "in-situ" interactions around a tabletop: we found that the raw amount of movement was not a relevant predictor for our purposes; however, we found that bimanual coordination was predictive of students' leadership in a group. Even more interestingly, clustering body position with k-means provided us with interesting categories: we found that "active" positions were correlated with learning gains, "passive" positions were negatively correlated with learning gains, and that the number of transitions between those states was predictive of learning. Third, we explored students' body language on a social level: contrary to common social psychology theories, we found that body synchronization was not correlated with any of our measures. Fourth, we found that the distance between students' bodies during the activity was associated with their pre-existing knowledge on the topic taught: students with low scores on the pre-test tended to be further away from their partner compared to students who obtained a high score. We interpret this result as a sign of defensiveness regarding an unfamiliar and possibly difficult topic for them to learn about.

Finally, we gathered all measures into one matrix and run an unsupervised machine learning algorithm to roughly predict students' learning (i.e., using a median-split on their scores). We found that SVM with a multilayer perceptron kernel achieved a 100% classification accuracy using 6 only features. It is interesting to see that even though some features were not significantly correlated with our outcome of interest, they still retained some predictive value when used in combination with each other features.

There are obvious limitations in this work. Our sample is rather small (N=38) for making predictions that generalize to other settings and other groups of students, and for training a supervised machine-learning algorithm. We also found that our correlations were affected by our two experimental groups: in most cases, students in the "discover" condition behaved differently from students in the "listen" condition, which affected our results; this was visible both in our scatterplots and in the ANOVAs we performed on our various measures between our two groups. Future work should replicate those results with a larger, uniform sample. Finally it is important to mention that we did not adjust our statistical analyses for multiple comparisons (Rothman, 1990); the main reason for this is that our sample size is rather small. We decided to follow Rothman's advice that "scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings." The goal of exploratory work is not to look for causal inferences; rather, our aim was to generate hypotheses about the potential existence of one or several effects. Replications are crucial for providing more solid evidence that the findings reported in this paper represent actual effects (as opposed to being just noise).

One possibility for future work based on these findings would be to implement learning algorithm to capture data as students are working on a task and make just-in-time predictions

minute by minute. If we imagine that our predictions are already acceptable after half of the activity, we could implement a feedback loop to the learning environment (e.g., the tangible interface) and provide personalized scaffolding to different groups of students. This feedback loop could potentially also be used in a classroom to inform teachers about the status of their students.

7. CONCLUSION

Our goal with this paper is twofold: first, we introduced methods to compute meaningful measures from logs generated by a tangible interface and Kinect data; second, we correlated those measures with students' learning gains to find relevant predictors of learning. We obtained significant results using information retrieval techniques (i.e., cosine similarity metrics) on the logs and clustering methods (i.e., k-means) on the gesture data that explained differences between our two experimental groups. We also showed that our metrics were particularly useful when used as features for a supervised machine-learning algorithm.

The main implication of this work is that we found interesting predictors of learning in an ecologically-valid task: i.e., students were using an interactive tabletop that had no constraints for gestures—all actions were allowed. The task itself was very open-ended: there were multiple paths to success. This is a departure from research that uses data mining in very constricted and well-structured tasks in which students either follow a scripted procedure or have a very narrow solution space to navigate. In this work, gestures were used to automatically detect how much students learn during a particular activity, but we envision that such gains could be also correlated with other multimodal data (e.g., eye gaze movements collected with mobile eye-trackers, arousal measures gathered using galvanic skin response sensors, speech data with microphones).

The approach described in this paper opens new doors for assessing students' learning in a variety of settings. Project-based education, for instance, is used in all kinds of engineering and in K-12 classes. Being able to perform formative assessment and judging the *process* of creating a particular artifact (as opposed to merely evaluating the final product) is a powerful way to both understand and influence students' learning trajectories. Finally, most of the work about learning analytics has been conducted online and has focused its attention on click-stream data. We believe that shifting this focus from online to "in-situ" activities has the potential to provide researchers with a richer understanding of students' struggles and difficulties: by gaining access to their individual learning pathways, we can start to think about providing them with some kind of personalized assistance as they create learning artifacts in a co-located setting. This is a departure from previous research that has looked at students' processes in a qualitative way, which is an extremely difficult and time-consuming methodology. Our contribution is to make these analyses easier to conduct, easier to replicate, and to provide new ways to visualize students' progresses as they are learning new scientific ideas.

8. ACKNOWLEDGEMENTS

We gratefully acknowledge grant support from the National Science Foundation (NSF) for this work through the CAREER Bifocal Modeling grant (NSF # 1055130), as well as funding from Stanford's Lemann Center for Entrepreneurship and Educational Innovation in Brazil.

9. REFERENCES

ABRAHAMSON, D., TRNINIC, D., GUTIÉRREZ, J.F., HUTH, J., AND LEE, R.G. 2011. Hooks and Shifts: A Dialectical Study of Mediated Discovery. *Technology, Knowledge and Learning 16, 1*, 55–85.

ANDERSON, M.L. 2003. Embodied cognition: A field guide. *Artificial intelligence 149, 1*, 91–130.

BACHOUR, K., KAPLAN, F., & DILLENBOURG, P. 2008. Reflect: An interactive table for regulating face-to-face collaborative learning. *Times of Convergence. Technologies Across Learning Contexts*. Springer Berlin Heidelberg, 39-48.

BLIKSTEIN, P. & WORSLEY, M. (accepted). Multimodal Learning Analytics: a methodological framework for research in constructivist learning. *Journal of Learning Analytics.*

BLIKSTEIN, P., WORSLEY, M., PIECH, C., SAHAMI, M., COOPER, S., & KOLLER, D. 2014. Programming Pluralism: Using Learning Analytics to Detect Patterns in the Learning of Computer Programming. *Journal of the Learning Sciences*.

BRANSFORD, J. AND SCHWARTZ, D. 1999. Rethinking Transfer: A Simple Proposal with Multiple Implications. *Review of Research in Education, 24*.

CHARTRAND, T.L. AND BARGH, J.A. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology 76, 6*, 893–910.

CHURCH, R. BRECKINRIDGE. 1999. Using gesture and speech to capture transitions in learning. *Cognitive Development* 14, no. 2: 313-342.

ESCALERA, S., GONZÀLEZ, J., BARÓ, X., REYES, M., LOPES, O., GUYON, I., AND ESCALANTE, H. 2013. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 445-452). ACM.

GWEON, G., JAIN, M., MCDONOUGH, J., RAJ, B., & ROSÉ, C. P. 2013. Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer-Supported Collaborative Learning*, *8*(2), 245-265.

HALL, E.T. *The hidden dimension* (1st ed.). 1966. Doubleday & Co, New York, NY, US.

HATANO, G. AND INAGAKI, K. *Two courses of expertise.* 1986. In H.W. Stevenson, H. Azuma and K. Hakuta, eds., Child development and education in Japan. W H Freeman/Times Books/ Henry Holt & Co, New York, NY, US, 262–272.

MANNING, C.D., RAGHAVAN, P., AND SCHÜTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

PENTLAND, A., & HEIBECK, T. 2008. *Honest signals*. Cambridge, MA: MIT press.

RICHARDSON, D.C. AND DALE, R. 2005. Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and Its Relationship to Discourse Comprehension. *Cognitive Science 29, 6*, 1045–1060.

ROTH, W.-M. 2001. Gestures: Their Role in Teaching and Learning. *Review of Educational Research 71, 3*, 365–392.

ROTHMAN, K. J. 1990. No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1), 43-46.

SCHNEIDER B., WALLACE J., BLIKSTEIN P., & PEA, R. 2013. Preparing for Future Learning with a Tangible User Interface: the Case of Neuroscience. *IEEE Transactions on Learning Technologies, 6*, 2, 117-129.

SCHNEIDER, B., AND BLIKSTEIN, P. 2015. Discovery Versus Direct Instruction: Learning Outcomes of Two Pedagogical Models Using Tangible Interfaces. *International Conference on Computer-Supported Collaborative Learning*, CSCL'2015.

SCHNEIDER, B. AND PEA, R. 2013. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning 8, 4*, 375–397.

SCHWARTZ, D.L. AND ARENA, D. 2013. *Measuring What Matters Most: Choice-Based Assessments for the Digital Age*. MIT Press.

SHAER, O., STRAIT, M., VALDES, C., FENG, T., LINTZ, M., AND WANG, H. 2011. Enhancing genomic learning through tabletop interaction. *Proceedings of the 2011 annual conference on Human factors in computing systems,* ACM, 2817–2826.

SHERIN, B. 2012. Using Computational Methods to Discover Student Science Conceptions in Interview Data. *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge,* ACM, 188–197.

SIEGLER, R. S., AND K. CROWLEY. 1991. The microgenetic method: A direct means for studying cognitive development. *American Psychologist* 46.6: 606.

TSCHAN, F. 2002. Ideal Cycles of Communication (or Cognitions) in Triads, Dyads, and Individuals. *Small Group Research 33, 6*, 615 –643.

WORSLEY, M. AND BLIKSTEIN, P. 2013. Towards the Development of Multimodal Action Based Assessment. *Proceedings of the Third International Conference on Learning Analytics and Knowledge,* ACM, 94–101.