Blending Learning Analytics and Embodied Design to Model Students' Comprehension of Measurement Using Their Actions, Speech, and Gestures

Avery H. Closser[a,*], John A. Erickson[b], Hannah Smith[a], Ashvini Varatharaj[c], Anthony F. Botelho[d]

[a] *Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01612*
[b] *Western Kentucky University, 1906 College Heights Blvd, Bowling Green, KY 42101*
[c] *University of California, Santa Barbara, 3432 University Drive, Santa Barbara, CA 93111*
[d] *University of Florida, 1221 SW 5th Ave, Gainesville, FL 32601*

Abstract

Although interdisciplinary collaborations are becoming increasingly common, researchers typically use data analysis methods specific to their field in order to uncover how students learn. We present affordances of integrating theories of embodied cognition and design with machine-learning methods to study student learning in mathematics and inform the design of embodied learning activities. By increasing such collaborative research efforts, learning scientists can incorporate regularization in computational models and ultimately draw reliable conclusions to further inform theory and practice through the design of technology-augmented learning activities. To illustrate this point, we explored students' conceptual understanding of measurement since limited research has identified measurement estimation strategies that should be emphasized in classroom instruction. By uniquely applying machine-learning methods to a small, multimodal dataset from a study on student behavior in mathematics, we identified behavioral profiles, patterns in speech, and specific actions and gestures that are predictive of performance. These findings may be used to inform the design of embodied learning activities for measurement. We discuss the contribution of these findings to the field of embodied design, and the affordances and

*Corresponding author
Email address: aeharrison@wpi.edu (Avery H. Closser)

challenges of conducting collaborative research in the learning sciences.
Keywords: Embodied Cognition, Regularization, Natural Language Processing, Mathematics Learning

1. Introduction

Learning is complex and multifaceted, necessitating research across fields to study learning in different contexts (e.g., formal and informal environments, online and offline settings). On one hand, cognitive scientists have employed a range of study designs (e.g., observational, experimental, and qualitative) and statistical methods on data from a variety of sources (e.g., assessments, coded behaviors) to study learning driven by cognitive theories. On the other hand, learning analysts have refined the use of machine learning methods to make inferences about student learning and provide recommendations for instruction and learning environments from large amounts of recorded data. Unique goals and efforts from both learning analytics and the cognitive sciences have been informative to the learning and educational research communities; however, as integrative theories of embodied cognition and design emerge, and more multimodal data collection becomes feasible, synergistic efforts are necessary to advance the study of learning.

Recent work across the learning sciences has emphasized cross-disciplinary collaborations with active exchanges of field expertise to study learning more broadly. Namely, there has been a push to utilize the field of learning analytics [1] as a bridge between disciplines within the learning sciences [2]. Notably, the field of multimodal learning analytics (MMLA; e.g., [3, 4]) has demonstrated the advantages, and value, of integrating multimodal data with machine-learning methods to draw inferences about learning from multiple sources collected across disciplines within the learning sciences. Methods of MMLA often allow researchers to capture more dimensions of learning processes to measure latent constructs with greater accuracy; constructs of learning may be operationalized in varying ways and a greater number of measured dimensions of such constructs allows researchers to better separate important factors from random noise. Similarly, more research grounded in embodied cognition [5, 6, 7, 8] and embodied design [9, 10, 11] has resulted in new learning technologies that utilize multimodal sensors for feedback and data collection. As such, combining research efforts between cognitive scientists and learning analysts is necessary to advance the study of student learning and the continuous improvement of learning technologies grounded in embodied design. Such a blending of efforts, just as is an advantage of MMLA, provides additional dimensions with which to explore learning processes to build better definitions, measures, and interventions to positively impact student learning.

Currently, we advocate for collaborative efforts between the cognitive sciences and learning analytics fields. We contend that to effectively garner deeper insights into learning and contribute impactful recommendations for design, interdisciplinary collaborations must be embraced. To illustrate this point, this project applies machine-learning methods to a relatively small dataset from a qualitative study with two goals. First, we examine how elementary and college students reason about measurement and display understanding through task-related behavior. Second, we discuss the affordances and challenges of applying machine-learning methods to observational, behavior-focused data sets, demonstrating how regularization can support the blending of learning theory with learning analytics. Through this exploratory work, we aim to encourage more collaborative research informed by theories of embodied cognition and design by illustrating the affordances of utilizing learning theory and machine-learning methods together to advance educational research.

1.1. Embodiment, Mathematics Learning, and Design

Theories of embodied cognition (e.g., [5] [6] [7] [8]) share the philosophical standpoint that thinking does not occur within a black box; rather, our physical, sensorimotor experiences in the world reflect and influence our cognitive processes, including mathematical thinking and reasoning [5]. From these theories, mathematics learning can be modeled as a multimodal, cyclical process impacted by the reciprocal relationship between perception and action. Specifically, Alibali and Nathan [12] argue that mathematical cognition is "based in perception and action, and it is grounded in the physical environment." An individual's learning environment shapes their perceptions, and that, in turn, informs their cognitive processes to act on and in their environment, then influence mathematical skills and thinking. This theoretical perspective has informed educational research on students' physical behaviors (i.e., actions, language, and gestures) as they relate to mathematics reasoning and learning.

For instance, gestures are a primary example of behaviors which contribute to mathematics learning. Distinct from actions, which effect change on the environment, gestures are primarily hand movements that complement speech to simulate actions and perceptual states [13] [14]. Student gestures have been shown to improve their abilities to process new mathematics concepts [15] and indicate their readiness to learn concepts they are unable to express verbally [16] [17] [18]. In addition to impacting reasoning and learning about math, gestures and actions reveal cognitive processes and insights into student knowledge, attitudes, and beliefs that may not necessarily be reflected in speech [19] [20].

More broadly, a large body of research has examined the relations between student behavior, cognitive processes, and learning in different contexts (i.e., the relations between student actions, speech, gestures) [21] [22] [23] [20] [24]. Previous research on the coupling between speech and physical movements in communication has shown a stronger relationship between speech and gesture than between speech and action in the context of language production and language comprehension [23] [24]. However, this line of inquiry has not been extended to explore the relationship between action, speech, and gesture as they relate to mathematics learning. Similarly, limited research has examined which types of behavior could be most indicative of students' understanding and implicit cognitive processes during embodied mathematics activities.

More broadly, a large body of research has examined the relations between student behavior, cognitive processes, and learning in different contexts (i.e., the relations between student actions, speech, gestures; e.g., [21, 22, 23, 20, 24]).

Previous research on the coupling between speech and physical movements in communication has shown a stronger relationship between speech and gesture than between speech and action in the context of language production and language comprehension [23, 24]. Similarly, Congdon, Kwon, and Levine [25] discovered an interaction between students' conceptual understanding of measurement and whether they benefited from the use of actions and gestures. Specifically, they found that first graders with higher prior knowledge benefited from the use of actions and gestures whereas students who displayed lower knowledge did not benefit from producing gestures. This finding suggests that the relation between behavior types and performance on measurement tasks may be moderated by knowledge such that instruction for novices may benefit from focusing on productive actions rather than eliciting meaningful gestures associated with measurement skills. However, limited research has examined which types of behavior could be most indicative of students' understanding and implicit cognitive processes across different ages beyond first grade and by examining different behaviors at a fine-grained level of analysis.

1.1.1. Embodied Design

With a large body of research demonstrating the powerful role of the body and action in cognition, such as through the coupling between gesture and mathematical cognition, recent attention has turned to designing learning technologies that support learning and reasoning through movement-based activities. This area of research, Embodied Design (ED), builds on theories of embodied cognition to investigate how students interact with embodied learning activities; this research informs theories of teaching and learning as well as the development and refinement of embodied learning activities in a bidirectional relationship [9, 11]. ED research is growing increasingly important for evaluating and refining math learning activities and educational technologies as more of them incorporate ED principles through perceptual and embodied experiences (e.g., [26, 27]; see Abrahamson et al. [10] for an overview). Particularly relevant to our work, Abrahamson and colleagues [10] note that a major principle of ED is that of "abstr-action" which states that student actions (both spontaneous and externally directed) can facilitate insights and understanding of mathematical concepts. Our study utilizes an ED approach to extend research on embodiment and to inform the design of embodied learning activities that effectively support the development of measurement skills through guided and meaningful actions that connect to the mathematics concepts at hand. Instructional and technology-augmented activities designed from principles of embodiment present exciting opportunities to study student learning through a rich array of behavioral data (e.g., through video recordings, Kinect sensors, joint-tracking, etc.) that would not be present in computer-driven or paper-and-pencil activities and constitute the application of machine-learning methods to meaningfully integrate and interpret multimodal data (e.g., [28]). One of the largest affordances of integrating ED theory and learning analytics is the added regularization to models and analyses.

1.2. Regularization: An Affordance of Blending Embodied Design and Learning Analytics

Efforts to promote interdisciplinary research across sub-fields of learning sciences have become increasingly common as teams utilize educational technologies to collect behavioral and sensorimotor data related to learning. As a result, interdisciplinary teams have formed to utilize machine-learning methods in work that is grounded in cognitive theories of embodiment and may have been approached from a qualitative or traditional statistics approach in the past. For instance, the Mathematics Imagery Trainer (MIT; [29, 30]) was designed to support students' conceptual understanding of proportional equivalence by providing hands-on opportunities for students to develop new sensorimotor schemes. Designed in alignment with theories of embodied cognition, research around the technology has expanded to apply machine-learning methods to the multimodal behavior data to identify students' strategies while using the MIT. Pardos et al. [31], for example, used expert-developed labels to classify student strategy in deep learning models utilized deep learning models. Similarly, Ou et al.

[32] combined clustering and regression analyses to study student interactions with the MIT and Tancredi and colleagues [33] applied a nonlinear analysis model to examine changes in students' perceptuomotor behavior to inform future iterations of the technology to support student learning through embodied interactions. All of these examples from research on the MIT demonstrate how expert knowledge was incorporated into the respective machine-learning analyses to develop models of student action (e.g., the expert-generated labels utilized in Pardos et al. [31]) and to inform analysis decisions (e.g., the selection and interpretation of clusters observed in Ou et al. [32]) in a way that extends beyond the capacities of qualitative or traditional statistics approaches to provide new insights into student learning.

The importance of applying theory and domain knowledge in data analyses has recently been acknowledged in learning analytics [2] [34] and we argue that one of the primary reasons to leverage learning theory and ED research in learning analytics is the concept of regularization. *Regularization* is the introduction of knowledge or other information to better-structure a problem or improve the generalizability of a model or approach [35]. In this way, regularization is the process of restricting the search space of possible methods through practices including normalization and dimensionality reduction, or by incorporating constraints on model selection or training procedures.

The importance of applying theory and domain knowledge in data analyses has recently been acknowledged in learning analytics (e.g., [2, 34]) and we argue that one of the primary reasons to leverage learning theory and ED research in learning analytics is the concept of regularization. *Regularization* is the introduction of knowledge or other information to better-structure a problem or improve the generalizability of a model or approach [35]. In this way, regularization is the process of restricting the search space of possible methods through practices including normalization and dimensionality reduction, or by incorporating constraints on model selection or training procedures. As an example, consider the common challenge in clustering analyses in choosing number of clusters by which to group the given data; in practice, this is normally accomplished by observing and interpreting characteristics of resulting clusters at different cluster sizes (i.e. different values of 'K' in the K-means clustering method, for example) and selecting the number that produces the most interpretable groups. This process is an example of regularization in practice, as the decision is being made based on incorporated domain knowledge. Clustering as a process, as well as related methods including factor analyses and principal component analyses, are themselves examples of regularization as well. The choice to apply these methods emerges from a hypothesis or understanding that latent classes exist within a given data and that some of the variance therein is correlated within these groupings; accounting for these groupings within a machine learning model or analysis can help improve the chances of results generalizing to new contexts by explaining variance and removing noise that may confound such methods.

In the fields of machine learning, statistics, and even learning analytics, the practice of regularization is applied to reduce the chances of overfitting a model to a given dataset; this application of regularization is often considered essential when applying complex modeling methods, such as deep learning, to relatively small datasets. For example, many researchers and practitioners of machine learning are familiar with L1 and L2 regularization methods for their common usage within methods such as ridge regression [36] and Lasso [37]. These methods of regularization impose a cost when training a model on the magnitude of learned coefficients with the intuition that lower-valued coefficients are more likely to generalize across applications; that is, these methods help to perform feature selection within the model. Despite the wide usage of regularization in these contexts, there is often a connotative disconnect between common *methods* of regularization and the general *concept* of regularization. For instance, it is important to recognize that L1 and L2 methods introduce regularization when training models, but there are many other ways to incorporate regularization into an analysis.

Regularization occurs implicitly and explicitly at many stages of analysis, typically by introducing some form of external knowledge. Adding theory and domain knowledge to determine appropriate modeling procedures, including or excluding certain variables and interactions, or applying methods of grouping observation samples or reducing covariate dimensionality are all manners through which regularization is commonly introduced into an analysis. Particularly, when guided by domain knowledge, additional information can add statistical power, strengthen resulting claims, or even allow for analyses that would otherwise be infeasible [32] [31]. Therefore, ED research may greatly benefit from cross-disciplinary efforts to address impactful research questions through the application of theory-driven learning analytics methods.

1.3. Measurement and Measurement Instruction

Perhaps even unknowingly, we apply measurement skills and concepts daily. Therefore, it is crucial for elementary curricula to focus on developing an understanding of measurement concepts as well as building strategies for application beyond physical measurement [38]. However, students struggle to learn procedural and conceptual measurement skills, taking months or years to advance from one level to the next of learning trajectories [39] and even maintaining misconceptions about measurement tools and strategies through sixth grade [40, 41]. This is problematic because measurement skills is a foundation for more advanced critical skills in mathematics such as quantitative reasoning, arithmetic, and proportional reasoning, one of the key concepts of the Common Core Standards for Mathematics for middle school [42, 43, 44].

1.4. The Current Study: Exploring Students' Measurement Strategies

This project aims to advance research on student behavior, reasoning, and learning by using machine-learning methods to analyze the interplay between learners' actions, speech, and gestures while completing measurement estimation tasks. Since physical behavior during problem solving [22] and gestures [20] reveal implicit knowledge, we hypothesize that observing students' physical actions while problem solving will also reveal valuable implicit knowledge of measurement concepts. Our goal is to use machine-learning methods to discern how student behavior is indicative of student performance and, consequently, conceptual understanding in the context of measurement, in order to inform the design of future activities and instructional support for students to develop procedural and conceptual measurement skills.

To do so, we use data from a larger project in which college and elementary students estimated physical dimensions (i.e., height, width, length) of geometric objects and then explained their strategy and reasoning. The sessions were videotaped for further analysis on how physical and verbal behaviors, including actions, gestures, and speech, reveal students' understanding of measurement and estimation. Here, we analyze all three aspects of learners' behavior (action, speech and gesture) to identify those that may reveal whether a student understands concepts of measurement or whether they might be struggling and need additional support. We aim to identify different behaviors and behavioral strategies indicative of knowledge to inform the future design of embodied games for measurement. We analyze the behavior of both college and elementary students to explore how students at different levels of knowledge and expertise approach measurement estimation tasks and express their understanding of measurement. Specifically, we explore:

1. Are there common behavioral profiles among students and, if so, what do they suggest?
2. How do students' verbal reasoning about measurement tasks vary by age and measurement accuracy?
3. Do students' actions, speech, or gestures best predict performance on measurement estimation tasks?

2. Materials and Methods

2.1. Participants

We used video data collected from 51 participants. In the fall of 2018, 29 college students (59% female; 38% male; 3% non-binary) from a northeastern university participated in a study for course credit where they completed a series of measurement estimation tasks. Following a similar protocol, a second study was conducted with 33 elementary students (ages 8-11; grades 3-6) at a local after-school program. Of the 33 elementary students, we obtained and analyzed video data from 22 students for our final sample (55% female, 45% male). Of these students, six were in third grade, ten in fourth grade, and five in fifth grade (grade level was unreported for one student).

2.2. Procedure

A similar procedure, asking participants to complete a series of measurement tasks, was followed for both populations with the caveat that the study was shortened for elementary school participants. The eight tasks that were completed by both the college and the elementary students were used in the following analyses (see Appendix 8 for the list of tasks). College students who participated in the study were interviewed by graduate and undergraduate research assistants individually for 30 minutes. Research assistants were informed of the purpose of each study beforehand and trained in advance on the protocol to follow in interviews. Data collection for the elementary student study was completed by research assistants in 15-minute, one-on-one sessions at an after-school program. Informed consent was obtained for all participants prior to beginning the study.

Participants were informed that they would be completing different measurement tasks. For each task, participants were offered, though not required to use, an unmarked 6-inch or 12-inch dowel as a tool to estimate different dimensions (length, width, height or diameter) of geometric objects including prisms, spheres, and cylinders of various sizes, with dimensions ranging from two to 24 inches. For example, participants were presented with a 24” cylinder and asked to estimate its height. After verbally providing an answer for each task, participants were asked to explain how they arrived at that answer and were free to gesture during those explanations. No restrictions were placed on behavior during the answer explanations, allowing participants to demonstrate and pick up objects freely. Participants did not receive any accuracy feedback throughout the study. Participants were unrestricted in the amount of time they could spend on any given task and (barring participants who left the after-school program early) all participants provided an answer for all tasks during the interviews.

2.3. Behavioral Measures Code Book

Participants' video data was analyzed using a behavioral measures code book designed to provide quantitative data about the actions, language, and gestures observed by students as they complete estimation tasks and explain the strategies they used [45]. This code book builds off previous work on gesture analysis with the intent to capture behavioral markers of students' 1) actions while problem-solving, 2) speech used to explain measurement strategy, and 3) gestures displayed with speech, through video footage to study how different behavior types afford different information about student knowledge and reasoning. To that end, we compiled all of the coded behavior into one dataset in which the codes distinguish whether each behavior was observed while students were solving the task or explaining their process afterwards.

The coding book consists of 35 items ranging from binary indicators of a present behavior to categorical items based on participants' actions, language, and gestures. There are 11 features based on participants' actions while completing each measurement estimation task. Binary features include whether the participant: 1) *used a dowel* while measuring; 2) *used an external tool*, such as a pen; 3) *used an autonomous tool* (e.g., finger); 4) *used a placeholder* while measuring an object; 5) *used a start-point marker* to designate where they began measuring; 6) *used an end-point marker* to designate where they stopped measuring a dimension; 7) *double-checked* their answer; and whether the participant 8) *decomposed* the problem into smaller measurement tasks. Additionally, *proximity* indicated whether the participant was physically near (within a foot), moderate (roughly 1-2 feet), or far (over two feet) from the object while measuring it. Lastly, *perspective* was coded as eye-level, high, or birds-eye, relative to the position of the object. Students' time-to-answer was also recorded as the duration between the task instructions and the students' final answer (M = 25.40 seconds, SD = 22.67 seconds) prior to providing an explanation of their problem-solving strategy.

Transcripts of student explanations of their measurement strategy were also recorded. These transcripts include participants' verbal explanations of their strategy after providing an answer for each task. These explanations range from short phrases (e.g.,"I don't know, I guessed") to more explicit explanations. For example, below is the verbal exchange between the researcher and a participant for one task:

> *Researcher:* "Can you estimate the height of the cylinder?"
>
> *Participant:* "Twenty-three and a half inches"
>
> *Researcher:* "How did you reach that answer?"
>
> *Participant:* "I had the six-inch wooden dowel and used its length and put my finder there to try to measure it. Then once I got to the end I kind of approximated the length of an inch and it wasn't fully the length of it, it was kind of over a bit, so I figured that was about half an inch."

Alongside this transcript, the actions demonstrated by the participant prior to saying "twenty-three and a half inches" were video analyzed and coded to depict the presence of different actions used by the participant to measure the cylinder. Specifically, the participant *used a dowel* (D), used their finger to indicate a *start-point marker* (S) as well as an *end-point marker* (E), and used their finger as a *placeholder* (P) while moving the dowel along the cylinder from right to left to estimate its height (Figure 1).

Lastly, each gesture produced by participants was recorded and coded as being one of five gesture types, defined based on previous work with gestures. *Deictic* gestures indicate objects, people and locations through point or reaching [46] [47]. *Spatial* gestures, a subset of iconic gestures, depict spatial relations [48]. *Kinetographic* gestures, another subset of iconic gestures, indicate actions [48]. *Pictographic* gestures, also a subset of iconic gestures, are used to depict a referent object [48]. *Metaphoric* gestures occur when an individual creates a

Figure 1: Example Measurement Estimation Task and Coded Action Features.

physical representation of an abstract idea or concept [49] including dimensions such as length, width, or height.

In order to establish agreement, four coders met regularly over two months to collaboratively code all college participant videos and discuss points of controversy. Once the four coders obtained 75% agreement (3 out of 4 coders) across 80% of items on seven given cases, the coders individually coded the remaining video data while continuing to meet and resolve discrepancies. For the elementary student data, three coders met regularly for a month to collectively code all tasks together for full agreement on all tasks.

2.4. Dataset

The behavioral data was coded and analyzed at the task level using the eight measurement tasks completed by participants in both the college and elementary sample. The final dataset included a total of 378 tasks (213 college, 165 elementary).

2.5. Approach to Analysis

To address our research questions, we applied three methodologies that leverage the processed data and coding labels within a machine learning-centered approach. While early in this work we argue for the benefits of blending machine learning and theories of embodied cognition and design, it is important to emphasize the broadness of that argument; there is often misconception or even disagreement as to what constitutes *machine learning* as opposed to other more traditional methods. In this work, we use this term inclusively, encompassing methods of deep learning, natural language processing, and nonlinear modeling as well as more traditional methods including clustering and regression (deep learning, after all, is just a form of regression), all where the computer is helping further research by learning and reporting from data. In this way, many of these methods are already being utilized to study cognition and learning, particularly in exploring the relationships between covariates and outcome of interested (e.g. through the application of regression models), but it is through even further integration of these methods at different stages of the research process that they may provide even greater benefit.

This work exemplifies this by applying a "discovery with models" approach [1] to identify whether students' actions, speech, or gestures most strongly correlates with observed measurement performance in the task. First, we utilized the expert-coded labels within a k-means cluster analysis to identify behavior profiles that emerged across the participants. Second, we applied state-of-the-art natural language processing techniques within another cluster analysis to identify differences in applied verbal strategies as students explained their approach to each measurement task. Finally, to address the third research question, we built regression models that observe how the profiles of student behavior (actions and gestures) and verbal strategies correlate with performance on task. These methods are described in detail below.

2.5.1. Identifying Behavior Profiles with K-means Clustering

From the coding procedure, we derived 44 binomial variables to describe student behavior within each task. These constructed features capture binary-coded representations of categorical coding labels (e.g., applied strategy, the use or non-use of tools such as a dowel) and binned representations of ordinal and continuous metrics recorded during the task (e.g., time to answer, number of gestures). These variables included verbal codes of students' explanations as they were informative of student behavior (e.g. the use of counting) while participating in the task.

Initially reported by Harrison, Smith, Botelho, Ottmar and Arroyo [50], we applied a k-means clustering method using Jaccard distance (for binomial variables) to identify the emerging behavior profiles of the students working on each of the measurement tasks. While clustering, we did not include an indicator of student age group (college or elementary) in order to identify those behaviors that were distinctive to each group as well as those that were shared; it was hypothesized and expected, in this case, that some clusters would be comprised of primarily one age group or the other, with notable findings being the cases where there were comparable distributions within clusters indicating that the represented behaviors were not age-dependent.

2.5.2. Examining Verbal Strategies with NLP and K-means Clustering

To explore students' verbal strategies, we applied several methods of natural language processing (NLP) to the transcribed text of the student explanations recorded in describing their approach to each task. The purpose of this analysis is to construct measures that represent what students verbalize while participating in the study as distinctive from their actions; while verbal strategies are likely to be correlated with action (e.g. a student is likely to verbally reference a tool if that tool was used), the choice of words in describing such actions may vary, providing insights into the students' conceptual understanding of measurement that simple codings of action are not able to fully represent; this NLP analysis, therefore, is conducted using the raw transcripts of student explanations, independent of the verbal codes described previously.

In alignment with common practices in NLP, several pre-processing steps were applied to first "tokenize" the data (i.e. identify individual words and break up contractions such as "couldn't" into "could" and "n't") and remove overly common stop words such as "the", "and", "a", and so forth [51]. Additionally, we applied stemming to commonly reduce each word to its root form. We then used a median split of task performance to observe words commonly used by high- and low-performing students during the measurement task as shown in Table 1.

Most Frequently Used Words			
High Performance		**Low Performance**	
inches	134	like	27
like	91	inch	24
dowel	78	count	20
two	76	one	17
one	66	measure	13
three	64	two	13
measure	60	would	12
use	59	three	12
six	54	five	11
stick	52	n't	10

Table 1: Counts of most frequently said words by high- and low-performing students

While it is apparent that both high- and low-performing students used numeric language, the most notable difference is the prominence of "dowel" and "stick" in the high performing group and that of "counted" in the low performing group. The position of these two words suggests that the high-performing students used verbal strategies that referred to the measured objects while the low-performing students more often described their approach to the task by counting. Further exploration revealed, unsurprisingly, that the low-performing group consisted of mostly elementary students. This difference in verbal strategy, likely aligning with their actions (i.e. the use or non-use of dowels as a measurement tool), began to highlight different approaches to the tasks, and we examined these differences further through a cluster analysis.

Utilizing complex deep learning models, we applied SBERT [52], a method designed to build semantic representations of sentences, pre-trained on a large set of data collected from Wikipedia and BooksCorpus. Researchers have made many such models openly available to allow others to utilize such tools without facing the technical challenges of training these models on such large language datasets. While uninterpretable by humans, the method produces a large, 768-dimensional-feature vector that represents the full student explanations in regard to their semantic meaning. The vectors are produced in such a way that two different explanations of similar semantic meaning will be closer together in the generated embedding space. This type of embedding is well-suited for a cluster analysis, as the feature vector can help identify emerging groups of similar explanations without additional coding or labelling efforts. Similar to our first analysis, we applied k-means clustering to the generated vectors across all student explanations.

2.5.3. Modeling the Role of Action, Speech, and Gesture with Regression Analyses

Our final analysis sought to observe whether recorded student actions, speech, or gestures most correlated with performance, as operationalized through student estimation error on the measurement task illustrated in Figure 1. Along with the features derived from the coding procedure, the students' estimation errors were also calculated and used as a dependent measure of performance on task in the analyses described in the following sections. Specifically, the magnitude of inverted estimation error was z-scored within task (c.f. Equation 1 such that all comparisons can be made with respect to the mean and standard deviations of estimation performance rather than the raw values; in the equation, the ϵ term is added as a small offset to avoid division by 0. The purpose of inverting the magnitude of error transforms the dependent variable for better interpretation of results (i.e., a higher value represents a higher performance exhibited by the student).

$$\begin{gathered} Inverse\ Error = log\left(\frac{1}{|correct - estimated + \epsilon|}\right) \\ Z\text{-}scored\ Estimation\ Performance = \frac{(InverseError - mean_{task})}{SD_{task}} \end{gathered} \tag{1}$$

Building on the previous two analyses, a linear regression model was developed for each set of observed student actions, speech, and gestures. The gen-

Cluster Label	Behaviors	N	% of College Students	% of Elementary Students	Chi-Square p-value
Low Effort	Eyeballing, no double-checking of answers, no dowel use or end-point marker	43	8.7%	13.9%	0.102
Confused	No gestures, long answer time, unknown verbal strategies	31	6.1%	10.3%	0.127
High Performance & Conceptual Understanding	Proportional action and verbal strategy, correct math reasoning, precise language	82	32.8%	4.2%	<0.001
High Effort, Low Performance	Varied long and short answer times, action-gesture-speech mismatch, many gestures	51	6.6%	21.8%	<0.001
Counters	Used counting verbal strategy, short answer time, imprecise language	63	3.9%	32.7%	<0.001
High Effort & Experience	Long answer time, correct reasoning, double-checked answers, estimation verbal strategy	124	41.9%	17%	<0.001

Table 2: Labels, prominent behaviors, and distribution of college and elementary students by cluster.

erated behavior features from the first analysis were divided into coding labels aligned to the actions taken by the student (e.g., use of tools or place-marking) and those aligning to observed gestures (e.g., use of deictic gestures, spatial gestures, etc.). These sets of features, along with the verbal clusters generated in our second analysis, were used to train the three linear models predicting the normalized estimation performance; in each model, the age group of the student was also included as a covariate. The predictions of these models represent the task performance that is explainable by the respective feature set. As such, we ensembled the three different predictions within a fourth linear regression model and observed the predictive power of each to identify which feature set is most correlated with student estimation performance while controlling for the other factors.

3. Results and Discussion

3.1. Identifying Behavior Profiles with K-means Clustering

After observing several values of "k," we found six clusters to be most representative, determined from a plot of the ratio of within- and across-sum-of-squared distances of samples to cluster centers, in addition to being interpretable for comparisons. Table 2 describes these clusters based on students' action, verbal, and gesture strategies and distribution between age groups. Notably, some profiles include both age groups (i.e., low effort and confusion) while others more distinctly comprise either elementary or college students. For instance, the behaviors represented mostly by college students in Clusters 3 and 6 include correct mathematical reasoning and proportional strategies, which have been considered expert strategies for length measurement in previous work [42]. Meanwhile, elementary students were more likely to use a form of counting (Cluster 5). This display of behavior aligns with an intermediate level called "End-to-End Length Measurer" on a hypothetical length learning trajectory by Samara and colleagues [44]. The students in Cluster 5 also show an incongruence between their actions while estimating and their speech and gestures while explaining their answer, suggesting that they were unable to articulate their problem-solving strategy (Cluster 4). Such incongruences are consistent with prior work on speech-gesture mismatch demonstrating that students who say one statement while gesturing in reference to another are likely on the verge of learning a new concept but are yet to formalize their understanding through speech (e.g., [16, 17, 18]). We compare these behavior profiles in more detail in a prior conference proceeding [50].

These behavior profiles may assist in identifying students with different degrees of conceptual knowledge about measurement based on their displayed behavior and performance during measurement tasks in classrooms (e.g., identifying students with high performance as opposed to students who may be identified as counters). Additionally, these clusters display common behaviors by high-performing students which suggest strategies to encourage in instructional support to help students develop strong measurement skills (e.g., proportional measuring strategy, start- and end-point markers).

Importantly, we were able to discern these behavior profiles through the use of k-means clustering because the method allows researchers to efficiently identify groups of cases without necessitating a priori hypotheses about which features may be the most distinguishing between groups (although such hypotheses

Most Frequent Bi-Grams						
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
feel like	six inch	look like	eight nine	point five	look like	one two
little bit	twelve inch	little bit	got top	four five	three inch	went like
went like	two inch	think like	little bit	one two	length stick	count like
four inch	four inch	kind look	one inch	three four	like inch	measured it
three four	inch stick	like inch	that twelve	two three	like three	right one
tried see	little bit	like like	three four	five six	one two	three four
circle sphere	two feet	one kind	bit would	like inch	twelve inch	two three
could have	four time	ten inch	by going	right here	two three	count sideways
go like	half inch	this one	count way	this one	cause look	dowel rod
kind like	length dowel	because like	eight inch	two point	that two	*inaudible* there

Table 3: Most frequently stated bi-grams by students in the verbal clusters.

could benefit such an approach by means of regularization). While hand-coding multimodal data and using descriptive statistics may allow researchers to also piece together different profiles of student behavior, this application of machine learning lends efficiency by automating the process of determining the appropriate number of groups and provides the opportunity to identify clusters objectively and then use the available data to qualitatively label the different behavior profiles that emerge from the clusters. As a representative grouping, clustering in this way helps reduce noise when modeling interactions and outcomes by reducing within-group variance.

3.2. Examining Verbal Strategies with NLP and K-means Clustering

Following the same method of determining a value for "k" as was performed in the first analysis, seven clusters of verbal strategies emerged (Table 3). We explored aspects of the identified groups by observing common bi-grams used by participants within each cluster; the bi-grams are viewed by looking at pairs of consecutive words, accounting for differing forms of words that may appear (e.g., "inches" and "inch" are treated as the same when counting the frequency of word pairs). While each of the clusters contained common bi-grams that exhibited various use of numbers in terms of measurement (e.g., "four inch") and evidence of counting (e.g., "three four," as well as several such references in Cluster 7), there are some other notable differences that are observed. Clusters 2, 6, and 7, for example, both reference the use of a measurement tool ("inch stick", "length stick", and "dowel rod"). Conversely, Clusters 1, 3, and 6 exhibit frequent use of the word "like," suggesting the use of imprecise language and estimation. From this, the verbal strategies appear to correspond with, unsurprisingly, the action strategy utilized by the student, but then further distinguishes students based on the precision of language used to describe their applied approach.

3.3. Modeling the Role of Action, Speech, and Gesture with Regression Analyses

The learned coefficients for each model utilizing action, verbal, and gesture features are depicted in Tables 4, 5, and 6. As can be seen in these tables, the age group identifier is consistently a strong predictor of student performance as was expected. In regard to the action model, only the use of a placeholder was found to be statistically reliable, and correlated positively with higher student performance. For the verbal model, two clusters were found to be statistically reliable; these two, from the comparison illustrated in Table 3, appear to highlight subtle differences in the precision of terms used in the measurement task. Beyond these subtle differences, however, it is difficult to draw further meaningful conclusions or interpretations pertaining to the resulting clusters of verbal strategy. This aspect of the analysis is further discussed as a challenge and limitation of the analysis in Section 4.2. Finally, observing the gestures model, the use of spatial and kinetographic gestures reliably predict student performance.

Finally, the results of the ensemble model, observing student estimation performance in regard to the combined models of action, verbal strategy, and gesture are reported in Table 7. As seen in that table, all factors were statistically reliable predictors of student performance, with the verbal and action models being the most predictive. This result suggests that these aspects are most correlated with our measure of student performance.

Variable	β	B	95% CI	t statistic	p value	
Intercept		-0.153	[-0.480, 0.174]	-0.915	0.361	
Elementary Student	-0.305	-0.609	[-0.813, -0.405]	-5.851	<0.001	***
Use Dowel	0.099	0.288	[-0.051, 0.627]	1.664	0.097	
Use External Tool	0.025	0.123	[-0.368, 0.615]	0.492	0.623	
Use Autonomous Tool	-0.060	-0.369	[-0.953, 0.215]	-1.237	0.217	
Use Placeholder	0.120	0.273	[0.030, 0.517]	2.201	0.028	*
Mark Start Point	0.061	0.138	[-0.133, 0.409]	0.997	0.320	
Mark End Point	-0.047	-0.094	[-0.312, 0.125]	-0.842	0.400	
Use Perspective	0.074	0.137	[-0.048, 0.322]	1.452	0.147	
Use Proximity	-0.092	-0.283	[-0.582, 0.016]	-1.852	0.065	
Double Check Estimate	-0.044	-0.117	[-0.374, 0.141]	-0.888	0.375	
Use Decomposition	-0.011	-0.039	[-0.370, 0.292]	-0.231	0.818	

Table 4: Features Used in Action Model

4. Implications and Contributions

We explored how students of varying ages and levels of experience approach and reason about measurement estimation tasks in order to identify successful measurement strategies through three distinct applications of machine-learning methods. We found that students' actions, speech, and gestures are all related to performance on measurement tasks. Specifically, using a placeholder while measuring, verbally articulating the use of a tool and units of measurement, and using spatial and kinetographic gestures were all behaviors of high-performing students. Together, these findings contribute to the body of research on student behavior and mathematics learning. The findings also illustrate the affordances of leveraging theories of embodiment and learning analytics together to advance our understanding of student learning and the development of learning technologies. In the following sections, we draw on examples and findings from our study to discuss affordances provided, and challenges faced, by using collaborative efforts to study student reasoning and learning through multimodal data.

Variable	β	B	95% CI	t statistic	p value	
Intercept		0.123	[-0.120, 0.366]	0.994	0.321	
Elementary Student	-0.304	-0.607	[-0.805, -0.409]	-5.998	<0.001	***
Verbal Cluster 2	0.892	0.426	[0.134, 0.718]	2.862	0.004	**
Verbal Cluster 3	-0.072	-0.144	[-0.473, 0.184]	-0.861	0.390	
Verbal Cluster 4	0.658	0.314	[-0.039, 0.667]	1.746	0.082	
Verbal Cluster 5	0.021	0.042	[-0.315, 0.399]	0.228	0.820	
Verbal Cluster 6	0.836	0.399	[0.005, 0.793]	1.985	0.048	*
Verbal Cluster 7	-0.039	-0.078	[-0.398, 0.242]	-0.478	0.633	

Table 5: Clusters Identified in Verbal Model

Variable	β	B	95% CI	t statistic	p value	
Intercept		-0.084	[-0.368, 0.199]	-0.582	0.561	
Elementary Student	-0.365	-0.729	[-0.920, -0.538]	-7.48	<0.001	***
Gesture: Spatial	0.192	0.500	[0.168, 0.832]	2.949	0.003	**
Gesture: Kinetographic	0.206	0.574	[0.229, 0.919]	3.264	0.001	**
Gesture: Deictic	0.110	0.295	[-0.050, 0.640]	1.677	0.094	
Gesture: Metaphoric	0.109	0.408	[-0.012, 0.828]	1.903	0.058	
Gesture: Pictographic	0.084	0.656	[-0.113, 1.426]	1.672	0.095	

Table 6: Features Used in Gesture Model

4.1. Affordances of Blending Embodied Design and Learning Analytics

From this project, we have identified two overarching affordances of blending ED principles with learning analytics. From a methodological standpoint, leveraging theory during data analysis incorporates regularization to promote the generalizability of results. More broadly, new approaches to study student learning, behavior, and strategy advance theories of embodiment to ultimately inform the development of educational technologies grounded in ED principles.

The blending of ED with LA, through the application of machine learning methods, provided us with the tools and measures necessary to operationalize and explore student gesture, verbal strategy, and action as they relate to measures of embodied cognition. While each of these abstract constructs are traditionally described by multiple measures, it is through machine learning that

Variable	β	B	95% CI	t statistic	p value	
Intercept		-0.391	[-0.629, -0.152]	-3.219	0.001	**
Elementary Student	0.449	0.895	[0.387, 1.403]	3.465	<0.001	***
Gesture Model	0.260	0.645	[0.124, 1.167]	2.434	0.015	*
Verbal Model	0.337	0.793	[0.380, 1.206]	3.776	<0.001	***
Action Model	0.343	0.799	[0.401, 1.197]	3.949	<0.001	***

Table 7: The ensemble model illustrating the relationship between action, verbal strategy, and gesture with respect to z-scored estimation performance.

we were ultimately able to represent each as a single measure to explore their relationships independently (e.g. we have a single numerical-valued representation of each construct of gesture, verbal strategy, and action as they pertain to the embodied cognition task); without the use of machine learning for this task, representing these constructs becomes a challenge that may require many additional hours of coding and validation to construct similar measures. In this example, theory was used as a means of selecting and applying the machine learning methods, but the resulting models themselves were then used to further our understanding of these in the context of a higher order construct operationalized through student performance on the measurement task; conversely, using a strictly data-driven approach without the use of domain knowledge would have made it difficult to extract and effectively interpret the relationships between these represented constructs (perhaps we could build a better model to predict estimation error within the task, but the blending of modeling approaches with domain knowledge allowed us to learn more from the process). As introduced earlier, this exemplifies a "discovery with models" approach [1], but is just one example of how the blending of ED and LA may help study these and similar constructs.

4.1.1. Methodological Contributions

Utilizing learning analytics approaches with learning theory to study student behavior can help incorporate regularization to increase the generalization of our findings to new contexts. This project is unique in our attempt to use a dataset from a small, in-person study with limited cases and features, demonstrating the ability to utilize machine learning techniques when paired with learning theory in order to deeply explore data on student learning. This affordance is exemplified by the clustering methods applied in the first and second analyses. These clusters provide the means to describe emerging groupings and trends in the data, similar to the work of Ou et al. [32] using a different clustering method. The dimensionality reduction provided by the methods not only increase our ability to interpret results and identify prominent groupings, but also help to reduce the impact of small variations in the data that may detract from the ability to generalize our findings to new populations. For instance, as it is unlikely that future students performing these tasks will use the same combination of words to describe their strategy, the language used by these new populations can still be mapped onto the categories of verbal strategies found in this work; with these, we would also be able to assess how well our results generalize based, for example, on if there are future explanations that do not map to any of the identified groupings.

Aside from our cluster analyses, the simple normalization performed in our final analysis also incorporates regularization and promotes the generalization of our results. The normalization of measurement performance, as in Equation 1, helps to remove student- and task-level dependencies that may differ in new populations or different tasks. These simple methods collectively improve the strength of our conclusions and improve the chances of our findings to generalize across datasets.

4.1.2. Theoretical Contributions

Leveraging theories of embodiment alongside learning analytics can advance learning theory as well as inform classroom instruction and the design of educational technologies in a cyclical relationship. We applied machine-learning methods to both substantiate previous findings and extend previous related work on the study of student behavior and gestures in mathematics, as well as inform the further refinement of existing embodied games. First, the behavior profiles that emerged in our first analysis coincided with successful measurement strategies (i.e., the use of proportional relationships) evidenced in previous research by Ayan and Isikal [42]. To extend our knowledge about behaviors that exhibit different levels of conceptual understanding, the NLP techniques revealed differences in the use of mathematically precise words and phrases used by high-performing students. Finally, through a "discovery of models" approach [1], utilizing predictive models derived from each set of student actions, speech, and gestures, we found that actions and speech are most predictive of performance on measurement tasks, extending previous work on the relation between students' actions, speech, and gestures as they pertain to other domains (e.g. [23, 24]). [24]. The ensemble model revealed that student actions and speech are more indicative of performance than gesture. These results extend the findings of Congdon et al. [19] who found that producing actions, but not gestures, were beneficial to novice learners in first grade to show a similar pattern as we observe behaviors from elementary and college students.

These findings have implications for the design of embodied learning activities for measurement; namely, to emphasize the development of students' actions and strategies for measurement as well as their ability to articulate their measurement process. For instance, we found that the use of a placeholder is positively related to performance, we infer that this measurement strategy and the underlying concept of tiling (i.e., that measuring should be done successively without gaps between units; [53]) should be emphasized more in elementary instruction through the design of embodied activities. For example, technology-driven measurement activities should implement instruction and feedback geared towards the appropriate use of a placeholder to assist students during hands-on practice.

Additionally, while previous research has shown that dynamic gestures support mathematical reasoning in geometry [54], we show that *spatial* and *kinetographic* gestures specifically are predictive of student performance in a new context, measurement. To the best of our knowledge, this is the first project to provide evidence that spatial and kinetographic gestures predict performance for measurement tasks and may be more indicative of higher understanding among students as opposed to other gesture types.

Moving forward, these findings can be used to inform the design of embodied learning activities targeting elementary-level measurement skills. The productive behaviors observed (e.g., use of a placeholder and explicit language) can be encouraged through instructional prompts and unproductive strategies can be anticipated in hints and immediate feedback features. Conversely, spatial and kinetographic gestures may be useful in instructional support to provide students. Ultimately, these findings can be used to encourage students to practice and mirror productive behaviors which are conducive to learning.

4.2. Challenges and Limitations

Alongside the affordances of considering the application of machine-learning methods to cognitive science data, we acknowledge that there remain challenges and limitations to doing so. Most notably, while applying machine-learning methods to this data allows us a fine-grained perspective on the relations between different behaviors and performance on measurement tasks, we are still unable to draw any conclusions about causality. Specifically, it is unclear from one-session interviews whether high achievement prompts students to apply productive strategies or whether applying productive strategies leads to higher achievement on such tasks. Moving forward, it would be prudent to consider including separate measures of conceptual understanding about measurement to determine whether high performance is a prerequisite to or outcome of using productive measurement strategies to inform the design of instructional activities and support.

Another prominent challenge is identifying the appropriate and efficient methods for an analysis. For example, we sought to identify whether action, verbal, or gesture strategy most correlated with student performance, and thus a simple linear model is sufficient to answer that question. Similarly, representing the complexity of student speech and language required a more complex, deep-learning approach to effectively capture verbal strategies. Linear regressions are arguably insufficient here as they would not be able to capture the syntactic and semantic meaning of words without extensive feature engineering that is afforded inherently by the utilized SBERT model [52]. In this way, it is important to plan the intended analyses around the posed research questions both to avoid potential "fishing" and also to consider those methods that best utilize the collected data and formulated learning theory.

The careful selection of methods before an analysis requires an understanding of such methods in terms of affordances and limitations. A notable challenge in applying many machine-learning methods is in the interpretation of results, though this may not always be a challenge that jeopardizes a result. The verbal strategy model again exemplifies this challenge. The resulting verbal strategy clusters, while representing different manners in which students described their approach to each task, do not clearly reveal how each grouping differs from others aside from the limited observations that can be made in Table 3; we were hesitant, for example, to label these clusters with a descriptive title due to the uncertainty that remained from these observations. We are able to identify in Table 5 that two of the verbal strategy clusters exhibit statistical reliability in their relationship with student measurement performance, and we are able to further identify in Table 7 that applied verbal strategy as a whole is reliably correlated with student performance, but these analyses are currently unable to describe the explanations as a whole within these clusters. For the current study, this is acceptable, as our research questions do not rely on our ability to fully interpret these groupings, but this certainly raises new research questions that should be addressed in future works.

Additionally, collaborations merging theories of embodiment with learning analytics should start at the initial phase of research. Like this project, researchers studying student behavior typically design studies in line with their research questions and planned analyses, such as small classroom studies or in-person interviews. Consequently, these study designs may perpetuate datasets that are not appropriate for many machine-learning methods. In that case, there may be limited room for exploratory collaborative efforts and the application of machine-learning methods. Instead, by establishing collaborative efforts from the initial phases of projects, researchers may be able to pose a wider range of research questions and appropriately design studies and methods to collect data appropriate for a wider range of analyses.

Lastly, and possibly the greatest challenge, is finding the necessary resources and support to make such collaborations feasible and valued. Many programs do not explicitly offer interdisciplinary training and those that do may still face challenges in effecting interdisciplinary collaborations. For instance, scholars with different backgrounds may have to reconcile field-based differences in perspectives (e.g., dissemination venues), language (e.g., "dummy" versus "one-hot encoded" variables), and work styles (e.g., programming language) in order to form successful partnerships and may benefit from resources or workshops to provide guidance for establishing new project teams. The effort necessary to reconcile these differences make some interdisciplinary collaborations seem less worthwhile.

4.3. Recommendations for the Field

Researchers have already advocated for more theory-informed applications of machine-learning methods and collaborations to support active exchange of knowledge and skills, particularly in learning analytics [2] [34] and between learning analytics and educational data mining (e.g., [1]). We advocate to extend this effort beyond the field of learning analytics more broadly into the learning sciences, particularly by encouraging support and training within programs and departments to provide emerging scholars with the skills to work in interdisciplinary research teams.

Echoing the charge for support at the doctoral level by Fischer and colleagues [55], we agree that more value should be placed on team science rather than individual accomplishments in doctoral training with opportunities to develop cross-disciplinary skills. Program-wide seminars, workshops, and special topics research groups could all be used to intentionally promote data analysis skills and start a dialogue about how theories of embodied cognition can impact our design, development and evaluation of educational technologies. To support this reconstruction of graduate training, pathways for interdisciplinary collaborations within or between programs should be established to provide emerging scholars with opportunities to forge new projects together and bring new perspectives to research on learning and the design, development, and evaluation of educational technologies.

Broadly, we implore learning scientists to consider the potential benefits of cross-disciplinary collaborations by blending theories of embodied cognition with learning analytics to pose new research questions and impact future research directions. For instance, cross-disciplinary collaborations can ease the historical tension between explanation and prediction research among social scientists [56] by affording more predictive methods. Likewise, learning analytics research should consider how domain knowledge may strengthen not only novel methods themselves, but also lead to more impactful outcomes that may inform methodologies in ED research. Overall, we suggest providing pathways for predoctoral scholars to engage in interdisciplinary work and leveraging collaborations to ultimately advance learning theory and inform the design of learning technologies by blending theories of embodied cognition and learning analytics.

5. Conclusion

In this paper, we applied clustering, natural language processing, and general linear modeling to a small yet rich dataset detailing student behaviors and speech during measurement tasks to identify successful measurement strategies. Our findings revealed profiles of student behavior and speech that may indicate different levels of conceptual knowledge as well as evidence that spatial and kinetographic gestures predict performance on measurement tasks. These findings advance research on gestures in mathematics learning; further, these findings can inform instructional support and technology-augmented embodied measurement activities that scaffold student learning through instructional support and feedback targeting prominent behaviors and strategies identified through our analyses. Additionally, we highlight the benefits of interdisciplinary collaborations to advance the field of learning sciences and the development of effective educational technologies grounded in ED principles. Namely, we contend that leveraging theories of embodiment and machine-learning methods affords added regularization in data analysis to increase the reliability and generalizability of findings. Moving forward, we encourage more collaborations and predoctoral training opportunities that leverage theoretical perspectives on embodiment and machine-learning methods to study student learning with a multitude of behavioral data to advance research and inform the design of educational technologies.

6. Selection and Participation of Children

Participants in this study were attendees at a local after-school program, specifically the cohort for children in grades 3-6. Prior to data collection, the staff at the after-school program collected consent forms from parents and provided demographic information about each participant, including their age, gender, and grade level. As part of the consent letter to parents, we specified that there were no known risks to participating children, or to their privacy, if they participated. Their name would never be associated with collected data and any answers they provided would not be linked to them personally. Parents who consented to their child's participation were able to specify their preferred level of data sharing (i.e., no videotaping, videotaping to be reviewed with the research team, videotaping that may be shared for dissemination among researchers, or videotaping that may be shared in public presentations). Researchers worked one-on-one with participants in a room of the after-school program for 15-20 minutes apiece, starting with a verbal assent process and confirmation that the child could cease participation at any time. After the study, each participant was compensated with a small toy and a debriefing letter was distributed to parents and guardians.

7. Acknowledgements

We are grateful for the support and feedback provided to us throughout this project from Erin Ottmar, Neil Heffernan, Ivon Arroyo, and our colleagues in the Learning Sciences & Technologies program at Worcester Polytechnic Institute.

8. Funding

This work was supported by the National Science Foundation grants #1645629, #1652579, #1822830, and #1903304.

Appendix A.

Task	Dimension	Answer
Can you estimate the height of the cylinder?	Height	24”
Can you estimate the combined length of the three cubes?	Length	14”
Can you estimate the height of the cylinder?	Height	8”
Can you estimate the diameter of the sphere?	Diameter	9 1/2”
Can you estimate the height of the cylinder?	Height	4”
Can you estimate the length of the longest side?	Length	8”
Can you estimate the diameter of the sphere?	Diameter	7 3/4”
Can you estimate the diameter of the cylinder?	Diameter	2 5/8”

Table A.8: List of measurement tasks by verbal instruction given to students.

References

[1] G. Siemens, R. S. d. Baker, Learning analytics and educational data mining: towards communication and collaboration, in: Proceedings of the 2nd international conference on learning analytics and knowledge, 2012, pp. 252–254.

[2] C. P. Rosé, Learning analytics in the learning sciences, Taylor & Francis (2018) 511.

[3] P. Blikstein, Multimodal learning analytics, in: Proceedings of the third international conference on learning analytics and knowledge, 2013, pp. 102–106.

[4] M. Worsley, 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, 2012, pp. 353–356.

[5] L. Foglia, R. A. Wilson, Embodied cognition, Wiley Interdisciplinary Reviews: Cognitive Science 4 (3) (2013) 319–325.

[6] M. Wilson, Six views of embodied cognition, Psychonomic bulletin & review 9 (4) (2002) 625–636.

[7] J. Lave, Cognition in practice: Mind, mathematics and culture in everyday life, Cambridge University Press, 1988.

[8] M. J. Nathan, Grounded mathematical reasoning, The Routledge handbook of embodied cognition 1 (2014) 171–183.

[9] D. Abrahamson, Embodied design: Constructing means for constructing meaning, Educational Studies in Mathematics 70 (1) (2009) 27–47.

[10] D. Abrahamson, M. J. Nathan, C. Williams-Pierce, C. Walkington, E. R. Ottmar, H. Soto, M. W. Alibali, The future of embodied design for mathematics teaching and learning, Future of STEM education: Multiple perspectives from researchers.

[11] D. Abrahamson, Building educational activities for understanding: An elaboration on the embodied-design framework and its epistemic grounds, International Journal of Child-Computer Interaction 2 (1) (2014) 1–16.

[12] M. W. Alibali, M. J. Nathan, Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures, Journal of the learning sciences 21 (2) (2012) 247–286.

[13] A. B. Hostetter, M. W. Alibali, Visible embodiment: Gestures as simulated action, Psychonomic bulletin & review 15 (3) (2008) 495–514.

[14] A. B. Hostetter, M. W. Alibali, Gesture as simulated action: Revisiting the framework, Psychonomic bulletin & review 26 (3) (2019) 721–752.

[15] S. Goldin-Meadow, S. W. Cook, Z. A. Mitchell, Gesturing gives children new ideas about math, Psychological Science 20 (3) (2009) 267–272.

[16] M. W. Alibali, S. GoldinMeadow, Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind, Cognitive psychology 25 (4) (1993) 468–523.

[17] R. B. Church, S. Goldin-Meadow, The mismatch between gesture and speech as an index of transitional knowledge, Cognition 23 (1) (1986) 43–71.

[18] E. A. Gunderson, E. Spaepen, D. Gibson, S. Goldin-Meadow, S. C. Levine, Gesture as a window onto children's number knowledge, Cognition 144 (2015) 14–28.

[19] E. L. Congdon, M. A. Novack, S. Goldin-Meadow, Gesture in experimental studies: How videotape technology can advance psychological theory, Organizational Research Methods 21 (2) (2018) 489–499.

[20] S. Goldin-Meadow, M. A. Singer, From children's hands to adults' ears: Gesture's role in the learning process., Developmental psychology 39 (3) (2003) 509.

[21] M. Chu, S. Kita, The nature of gestures' beneficial role in spatial problem solving., Journal of Experimental Psychology: General 140 (1) (2011) 102.

[22] R. Clark, Cognitive task analysis for expert-based instruction in healthcare, in: Handbook of research on educational communications and technology, Springer, 2014, pp. 541–551.

[23] R. B. Church, S. Kelly, D. Holcombe, Temporal synchrony between speech, action and gesture during language production, Language, Cognition and Neuroscience 29 (3) (2014) 345–354.

[24] S. Kelly, M. Healey, A. Özyürek, J. Holler, The processing of speech, gesture, and action during language comprehension, Psychonomic bulletin & review 22 (2) (2015) 517–523.

[25] E. L. Congdon, M.-K. Kwon, S. C. Levine, Learning to measure through action and gesture: Children's prior knowledge matters (2018).

[26] E. Weitnauer, D. Landy, E. Ottmar, Graspable math: Towards dynamic algebra notations that support learners better than paper, in: 2016 Future Technologies Conference (FTC), IEEE, 2016, pp. 406–414.

[27] M. J. Nathan, C. Walkington, Grounded and embodied mathematical cognition: Promoting mathematical insight and proof using action and language, Cognitive Research: Principles and Implications 2 (1) (2017) 9.

[28] S. Lee-Cultura, K. Sharma, M. Giannakos, Children's play and problem-solving in motion-based learning technologies a using multi-modal mixed methods approach, International Journal of Child-Computer Interaction (2021) 100355.

[29] M. Howison, D. Trninic, D. Reinholz, D. Abrahamson, The mathematical imagery trainer: from embodied interaction to conceptual learning, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp. 1989–1998.

[30] D. Abrahamson, D. Trninic, Bringing forth mathematical concepts: Signifying sensorimotor enactment in fields of promoted action, ZDM 47 (2) (2015) 295–306.

[31] Z. A. Pardos, C. Hu, P. Meng, M. Neff, D. Abrahamson, Classifying learner behavior from high frequency touchscreen data using recurrent neural net-

works, in: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, 2018, pp. 317–322.

[32] L. Ou, A. Andrade, R. Alberto, G. van Helden, A. Bakker, Using a cluster-based regime-switching dynamic model to understand embodied mathematical learning, in: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 2020, pp. 496–501.

[33] S. Tancredi, R. Abdu, D. Abrahamson, R. Balasubramaniam, Modeling nonlinear dynamics of fluency development in an embodied-design mathematics learning environment with recurrence quantification analysis, International Journal of Child-Computer Interaction 29 (2021) 100297.

[34] A. F. Wise, D. W. Shaffer, Why theory matters more than ever in the age of big data., Journal of Learning Analytics 2 (2) (2015) 5–13.

[35] P. Bühlmann, S. Van De Geer, Statistics for high-dimensional data: methods, theory and applications, Springer Science & Business Media, 2011.

[36] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

[37] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[38] National Council of Teachers of Mathematics, Principles and standards for school mathematics, `https://www.nctm.org/uploadedFiles/Standards_and_Positions/PSSM_ExecutiveSummary.pdf` (2000).

[39] J. E. Barrett, D. H. Clements, J. Sarama, Children's measurement: A longitudinal study of children's knowledge and learning of length, area, and volume, National Council of Teachers of Mathematics, Incorporated, 2017.

[40] G. T. Sisman, M. Aksu, A study on sixth grade students' misconceptions and errors in spatial measurement: Length, area, and volume, International Journal of Science and Mathematics Education 14 (7) (2016) 1293–1319.

[41] K. Zacharos, Prevailing educational practices for area measurement and students' failure in measuring areas, The Journal of Mathematical Behavior 25 (3) (2006) 224–239.

[42] R. Ayan, M. Isiksal-Bostan, Middle school students' proportional reasoning in real life contexts in the domain of geometry and measurement, International Journal of Mathematical Education in Science and Technology 50 (1) (2019) 65–81.

[43] N. G. A. C. for Best Practices, C. of Chief State School Officers, Common Core State Standards, National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C., 2010.

[44] J. Sarama, D. H. Clements, J. E. Barrett, C. J. Cullen, A. Hudyma, Y. Vanegas, Length measurement in the early years: teaching and learning with learning trajectories, Mathematical Thinking and Learning (2021) 1–24.

[45] A. Harrison, Exploring measurement estimation through learners actions, language, and gestures.

[46] D. McNeill, Hand and mind: What gestures reveal about thought, University of Chicago press, 1992.

[47] J. M. Iverson, O. Capirci, V. Volterra, S. Goldin-Meadow, Learning to talk in a gesture-rich world: Early communication in italian vs. american children, First Language 28 (2) (2008) 164–181.

[48] P. Ekman, W. V. Friesen, The repertoire of nonverbal behavior: Categories, origins, usage, and coding, Nonverbal communication, interaction, and gesture (1969) 57–106.

[49] M. Andric, S. L. Small, Gesture's neural language, Frontiers in psychology 3 (2012) 99.

[50] A. Harrison, H. Smith, A. Botelho, E. Ottmar, I. Arroyo, For good measure: Identifying student measurement estimation strategies through actions, language, and gesture, International Society of the Learning Sciences (ISLS), 2020.

[51] C. Fox, A stop list for general text, in: Acm sigir forum, Vol. 24, ACM New York, NY, USA, 1989, pp. 19–21.

[52] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084.

[53] R. Lehrer, Developing understanding of measurement, Tech. rep., National Council of Teachers of Mathematics, Reston, VA (Sep. 2003).

[54] C. Walkington, R. Boncoddo, C. Williams, M. J. Nathan, M. W. Alibali, E. Simon, E. Pier, Being mathematical relations: Dynamic gestures support mathematical reasoning, Boulder, CO: International Society of the Learning Sciences, 2014.

[55] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, M. Warschauer, Mining big data in education: Affordances and challenges, Review of Research in Education 44 (1) (2020) 130–160.

[56] T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: Lessons from machine learning, Perspectives on Psychological Science 12 (6) (2017) 1100–1122.