This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



# A Multimodal Assessment Framework for Integrating Student Writing and Drawing in Elementary Science Learning

Andy Smith, Samuel Leeman-Munk, Angi Shelton, Bradford Mott, Eric Wiebe, and James Lester

**Abstract**—Science learning is inherently multimodal, with students utilizing both drawings and writings to explain observations of physical phenomena. As such assessments in science should accommodate the many ways students express their understanding, especially given evidence that understanding is distributed across both drawing and writing. In recent years advanced automated assessment techniques that evaluate expressive student artifacts have emerged. However, these techniques have largely operated individually, each considering only a single mode. We propose a framework for the multimodal automated assessment of students' writing and drawing to leverage the synergies inherent across modalities and create a more complete and accurate picture of a student's knowledge. We introduce a multimodal assessment framework as well as two computational techniques for automatically analyzing student writings and drawings: a convolutional neural network-based model for assessing student writing, and a topology-based model for assessing student drawing. Evaluations with elementary students' writings and drawings collected with a tablet-based digital science notebook demonstrate that 1) each of the framework's two modalities provide an independent and complementary measure of student science learning, and 2) the computational methods are capable of accurately assessing student work from both modalities and offer the potential for integration in technology-rich learning environments for real-time formative assessment.

**Index Terms**—Intelligent Tutoring Systems; Formative Assessment; Multimodal Assessment; Student Writing Analysis; Student Drawing Analysis.

## 1 Introduction

Assessment plays a crucial role in learning. In the classroom, teachers rely on a combination of summative and formative assessments to help monitor student knowledge, diagnose areas of misunderstanding, and refine instructional strategies [1]. Formative feedback provided during the learning process can be more beneficial than a single summative judgment at the end, which places a growing importance on accurate and timely formative assessment [2], [3]. However, designing effective assessments is challenging, particularly if they are to be minimally disruptive to learning.

Interest in investigating how student learning data can be leveraged in real-time automated formative assessment to support teachers in the classroom has increased in recent years [4]. Compared to more distal, summative assessments, the rapid, cyclical nature of formative assessment provides a unique opportunity to integrate powerful computational systems that effectively diagnose student conceptual understanding and misunderstanding as learning progresses. Previous work makes clear that the more restrictive methods traditionally used in summative assessment, such as multiple-choice questions, are limited in their ability to provide the analyses necessary for guiding real-time scaffolded instruction for students (e.g., [5]). To address this issue, recent approaches to real-time formative assessment have shown promise by leveraging the rich, multifaceted data generated by digital learning environments, including analyses of student interaction logs in open-ended learning environments [6] and analyses of interactions with course materials and online tools to predict student performance [7].

In addition to measures of student interactions and interaction logs, analyzing artifacts of student work for formative assessment also shows great promise for making accurate inferences about student knowledge. However, student artifacts can take many forms depending on the subject matter and curricular goals. Given the growing breadth of activities enabled by digital science inquiry environments, it is important to develop assessment tools that can conduct integrated assessments of student work across multiple activities and modalities. In the work reported here we focus on two modalities commonly used in science education:


- *Andy Smith is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695. E-mail: pmsmith4@ncsu.edu.*
- *Samuel Leeman-Munk is with the Cognitive Computing Group, SAS Institute Inc., Cary, NC, 27513 . E-mail: sleemanmunk@gmail.com.*
- *Angi Shelton is with the College of Education, North Carolina State University, Raleigh, NC 27695 E-mail: angishelton@gmail.com.*
- *Eric Wiebe is with the College of Education, North Carolina State University, Raleigh, NC 27695. E-mail: wiebe@ncsu.edu.*
- *Bradford Mott is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695. E-mail: bwmott@ncsu.edu.*
- *James Lester is with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695. E-mail: lester@ncsu.edu.*




xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



short-text constructed responses and learner-generated drawings.

Each of these modalities has typically been assessed individually. Short-text constructed response items have been shown to reveal cognitive processes and states in students that are difficult to infer through multiple-choice equivalents [8]. Even when it seems that items could be designed to address the same cognitive processes, success in devising multiple-choice and constructed-response items that exhibit psychometric equivalence has proven to be limited [9]. Because standards-based STEM education in the United States explicitly promotes the development of writing skills for which constructed response items are ideally suited [10], [11], the prospect of designing text analytics techniques for automatically assessing students' textual responses has become even more appealing, spurring an acceleration of research in this area [12].

In a parallel development, drawing has become recognized as a central activity in science education, particularly in lower grades. Generating drawings of science phenomena can engage students in inquiry processes and foster a deeper understanding of concepts more than simply viewing drawings [13]. A variety of studies show that instructional strategies focusing on learner-generated drawings can produce positive learning outcomes by improving science text comprehension and student engagement [14], facilitating the writing process [15], and improving the acquisition of content knowledge [16]. However, these assessments are very labor intensive to develop and evaluate. Combined with the natural affordances provided by the growing presence of tablet computing platforms, automated analytical techniques provide a promising solution for implementing drawing-based assessments at scale.

Despite progress in evaluating each of these modalities in isolation, there is limited work exploring integrated frameworks for both writing and drawing. However, initial findings show that not only do students reveal conceptual understanding through both modalities, but also that different aspects of their understanding are often distributed across the modalities [17]. Other studies have shown that student drawings can be used as a valuable source of evidence to resolve ambiguities in student writing [18].

In this work we investigated the potential of multimodal assessment by analyzing elementary student writings and drawings with a common rubric. We present automated assessment techniques that are used to investigate two research questions. First, we explored how automated tools can assess student short constructed responses and symbolic drawings with respect to human grading. We found that a convolutional neural network approach for analyzing writing and a topology-based approach for analyzing drawing closely mirror the assessments made by human graders. Second, we explored how accurately a multimodal assessment framework (Fig. 1) integrating writing and drawing assessments predict learning outcomes compared to a single modality framework. We found that not only does each modality individually predict student learning outcomes, as measured by a summative post-test, but the integrated multimodal framework outperforms both uni-modal assessments individually.

This article is organized as follows. Section 2 discusses related work in the automated assessment of student short-text constructed responses and science drawings. Section 3 introduces the LEONARDO system used to collect the student science writing and drawing corpus, as well as the coding procedure used to analyze student writings and drawings. Section 4 provides an analysis of the human-coded scores. Section 5 introduces the computational methods used to automatically assess the writings and symbolic drawings. Section 6 presents the results of the automated assessment as well as additional analysis. Finally, Section 7 discusses results and directions for future work.

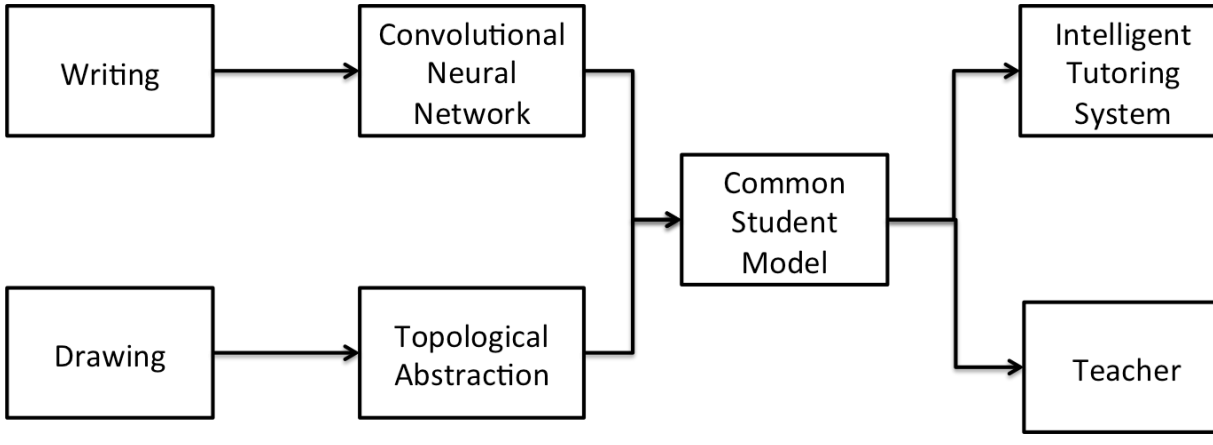


Fig 1. Multi-modal assessment framework

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



## 2 Related Work

Both the educational psychology community and the intelligent tutoring systems community have investigated the modalities of writing and drawing as types of formative assessment. However, relatively little work has explored their integration. Below we describe prior work on the analysis and interpretation of both modalities, as well as prior work in automated assessment.

### 2.1 Short-text Constructed Response Analysis

For many domains, a rich source of formative assessment is student written responses. Short answer responses can generally be characterized as requiring a response between one phrase and one paragraph of natural language that recalls knowledge not stated in the question, and is evaluated on content rather than writing style [19]. For the science classroom in particular, these written responses are often structured to encourage students to both make claims about scientific principles and provide evidence to support these claims [20]. These responses serve the dual role of both revealing underlying student mental models as well as encouraging reflection [21].

Short-text constructed items have long been a subject of interest for the intelligent tutoring systems research community, spawning a variety of approaches for automated analysis. Some of the simplest approaches to Constructed Response Analysis (CRA) are based on the assumption that the words a student uses can be used to analyze the content of his or her statement. This assumption allows early approaches, known as "bag of words" approaches, to ignore many complexities of human language in an effort to increase computational efficiency and portability across languages without the need to build or automatically learn complex grammars [22]. This makes "bag of words" approaches especially useful for cases of ill-formed text. However, for many purposes, "bag of words" approaches can also significantly reduce precision of analysis.

More recent approaches introduce greater sophistication and complexity in an attempt to capture the meaning lost in bag-of-words approaches. These techniques are characterized by their strategy of finding an alignment between sentences. A student answer is decomposed into constituent elements, generally words and short phrases, and these are annotated with linguistic features that the system uses to establish a best match with the given reference answer. Another approach gaining traction in recent natural language processing competitions combines complementary techniques that each specialize in a different characteristic of student answers to form hybrid techniques that can successfully analyze both. In this section we present these approaches and some techniques that represent each.

Latent Semantic Analysis is a widely used technique for the simple bag-of-words approach to CRA [23]. Latent Semantic Analysis makes a term-document co-occurrence matrix from a large corpus and performs Singular Value Decomposition, a process that filters noise from the data to leave only the most significant patterns. The result can be used to represent a given word or document as a vector of high level features that each represent a latent concept in the text. Comparing these vectors gives us a numeric conceptual similarity measure of the two documents or words.

As a bag-of-words technique, LSA is generally too imprecise, and therefore too permissive, in its grading. Because it cannot distinguish between "the water evaporates, leaving the salt" and "the salt evaporates, leaving the water," it will assign them the same grade. If the reference answer is looking for the concepts "salt," "water, and "evaporate," this grade will indicate a high level of student understanding even if the student has a serious misconception about the core concepts.

Beyond bag-of-words, many current techniques take advantage of deeper linguistic understanding, which they use to align concepts between a reference answer and a student answer. Both the Content Assessment Module [24] and Educational Testing Services' C-Rater [25] use a battery of linguistically sophisticated preprocessing tools to automatically annotate words in an answer with linguistic features such as its morphological stem, part-of-speech, and syntactic relationship with other words in the sentence. These features then allow the systems to map elements between the reference and student responses and assign a grade based on the resulting alignment.

Many state-of-the-art techniques work by integrating multiple approaches, often a combination of bag-of-words and alignment approaches. The philosophy behind this method is that while technique $A$ may have certain weaknesses, and technique $B$ may have other weaknesses, meta-technique $AB$ can integrate the best features of both of its constituents and achieve better accuracy than either on its own.

Educational Testing Services (ETS) uses a hybrid technique combining a bag-of-words technique with a translation evaluation technique BLEU [26] and its alignment-based system PERP [27]. ETS's technique adapts to scoring either questions present in the training data or unseen questions by making multiple copies of each feature. One copy is learned only from answers that share the same question as the answer to be graded, another from answers that share a domain, and a third from all the answers [28].

Dzikovska, Nielsen, and Brew propose an approach based on separating a reference answer into multiple facets and determining correctness of an answer for each facet to provide more fine-grained information about student understanding to intelligent tutors and human teachers [29]. Like the alignment approaches,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



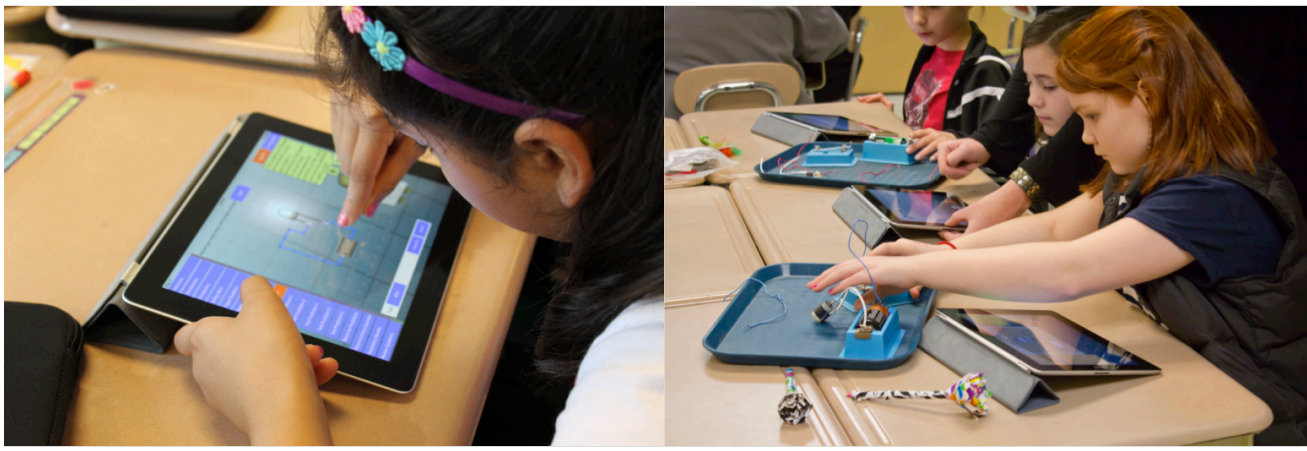

Fig 2. Students using LEONARDO on a tablet and in conjunction with physical experiments in the classroom

these facets are drawn from modified syntactic dependency parses of the reference answer, but unlike alignment approaches, the goal is not to return one overall score but to identify the student's expression of each facet individually.

The approach we use in this paper, convolutional neural networks, builds on our previous work [30], but otherwise has not been applied to short answer assessment. It has, however, demonstrated promise in several other text analytics tasks, including question identification, sentiment prediction, and semantic similarity [31]–[33].

### 2.2 Learner-generated Drawing

Unlike the well-studied areas of how people learn from writing text, viewing graphics, and reading, relatively little is known about how creating graphical representations affects learning. Van Meter and Garner [34] posit that students asked to draw a picture engage in three cognitive processes: selecting relevant information, organizing the information to build up an internal verbal model, and constructing an internal nonverbal representation to connect with the verbal representation. Others suggest that drawing can be a meaningful learning activity requiring both essential and generative processing to mentally connect multiple representations of the knowledge [35].

The benefits of learning-generated drawing are best realized by thoughtfully designing activities and situating them within a well-formulated curriculum, as the positive effects of drawing strongly depend on the quality of the learner-generated products and scaffolding [36]. The act of generating a visual representation can be a cognitively demanding task and as such, requires scaffolds to guard against excessive and extraneous cognitive load [37]. Examples of effective scaffolds for more structured drawing include providing cutout figures, guiding questions, and targeted drawing prompts [38]. Creating visual representations is also a crucial element of modelling in science education, often times combined with simulations or written explanations to help students illustrate, explain, and predict phenomena [39]. Furthermore, preliminary studies at the elementary grades show that student understanding is distributed unequally across drawing and writing in science notebooks [40].

Interpreting these visual artifacts poses significant computational challenges, with the majority of prior work focusing on entity recognition in free-hand sketches. The Mechanix system builds on this research by using free-hand recognition capabilities to convert student statics drawings into free-body equations which the system can then compare to a target solution and provide basic feedback on the students' solutions [41].

Bollen and van Joolingen's SimSketch system merges free-hand sketching with modeling and simulation of science phenomena [42]. In SimSketch, user free-hand drawings are segmented into distinct objects by the system, and then manually annotated by the user with a variety of behaviors, attributes, and labels. Students can then run a simulation based on their drawing and see the results before revising their sketch. SimSketch has been evaluated in a planetarium setting and been shown to be both a functionally useable and enjoyable system for visitors.

Another promising line of investigation for studying learner-generated drawing in educational settings is the CogSketch system [43]. CogSketch has been developed as an open-domain sketch understanding system, allowing users to annotate objects and relations in their drawings with entities and relations represented in the OpenCyc knowledge base. Short drawing activities, called Sketch Worksheets, have been built within CogSketch and used in a pilot study to collect undergraduate geology student drawings, which were then clustered using an analogical generalization engine [44]. Researchers have also used CogSketch to identify differences in the way experts and novices copy existing diagrams, comparing not only the final drawings but the process with which the drawings are created [45].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



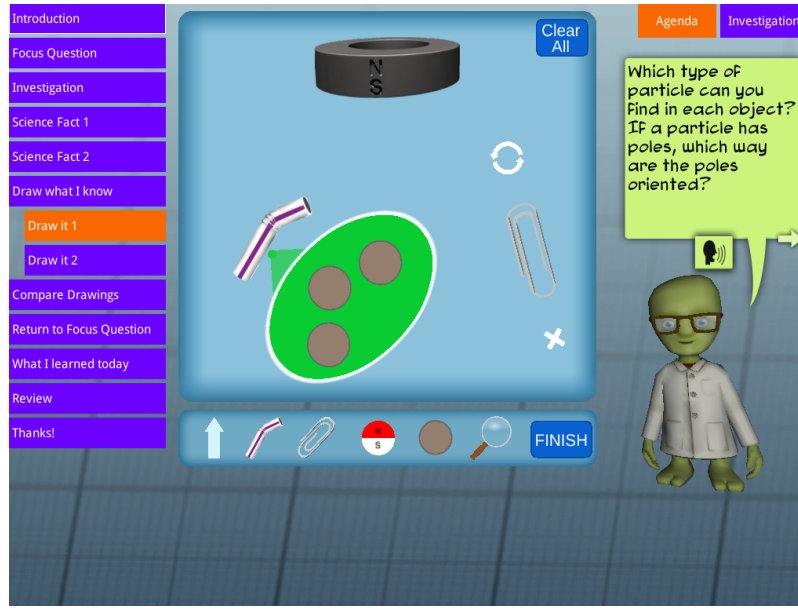


Fig. 3. Screenshot of the LEONARDO learning environment

Additionally, there has been work combining language and visual artifacts in tutoring systems. The ATLAS-ANDES system combined dialogue feedback and drawing, with students' free body diagrams potentially triggering feedback in a physics tutor [46]. More recently, a pilot study of the Design Buddy system combined drawing with Cogsketch and structured language input created with dropdown menus to provide feedback to students based on the consistency of the explanation [47]. Building on this work we explore the assessment of student writing and drawing with a common rubric. Then, utilizing the human codings of student artifacts we build automated assessment systems. Finally, we investigate how these codings can be used to compare student writing and drawing to better understand how the two modalities might be used synergistically in assessment.

## 3 System Description

In order to study student writing and drawing in an ecologically valid setting, we developed a digital learning environment modeled after science notebooks. Science notebooks are used extensively in elementary grades as a mechanism to promote and reveal reflective thought [48]. Science notebooks capture students' inquiry-based activities in a variety of forms, including both written and graphical form, potentially providing a valuable diagnostic source of student understanding and misconceptions. Unfortunately in many cases elementary teachers are trained as generalists and often have limited training specifically in science pedagogy, which poses significant challenges in effectively using science notebooks in classroom learning activities [49].

As computing technologies become more affordable, and ubiquitous in classrooms, we see opportunities to transition the paper science notebook to a virtual environment capable of leveraging the advances of intelligent tutoring systems. Over the past five years our laboratory has been developing a digital science notebook for elementary school science education called LEONARDO (Fig. 2) [30].

LEONARDO integrates intelligent tutoring systems technologies into a digital science notebook that enables students to graphically model science phenomena with a focus on the physical and earth sciences. Capable of operating on both conventional and tablet computing platforms, LEONARDO is designed to be used in the classroom in conjunction with physical experiments and is aligned with the Next Generation Science Standards for elementary school science education.

LEONARDO's curriculum is organized around focus questions that encourage students to follow the scientific method. For each focus question, students explore natural phenomena through writing and drawing about underlying scientific principles. Writing exercises are in the form of short answer questions where the student reads a question and answers it in a sentence or two. Drawing exercises consist of students creating symbolic drawings of different concepts depending on the current topic. Given the challenges of machine recognition of freehand sketch, as well as concerns of excessive cognitive load for fourth graders working on such an unstructured task, LEONARDO supports symbolic, diagrammatic drawing (Fig. 3). This can be considered analogous to existing class room activities such as students creating visual artifacts with paper cutouts, or working from a predefined glossary of symbols. During these activities, students choose from a variety of pre-authored symbols representing macroscopic and microscopic elements of a domain. Students can then add, remove, rotate, and move the elements to produce their visual representations.

To date modules have been developed for three con-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



tent areas: Electricity and Circuits, Magnetism, and Weather. These modules represent over 16 classroom hours of content. Overall various versions of LEONARDO have been used in over 35 schools spanning more than 11 states within the United States.

### 3.1 Data Corpus

The data analyzed in this study consists of student writing and drawing samples collected from a learning activity in the area of Magnetism. This particular activity guides students' investigation of what happens to magnetic particles in the presence of a magnetic field. From this activity, two writings and two drawings were collected for each student. The first writing sample was taken at the beginning of the exercise in response to the prompt, "What happens to the particles when an object is turned into a temporary magnet?", and is referred to as *Writing 1* in the results. After responding to the first prompt, students were then presented with a series of scientific explanations, as well as a brief physical experiment involving a magnet, a straw and a paperclip designed to help students determine that some materials can undergo induced magnetism and others do not. Examples of student answers are shown in Table 1 for each of the two writing prompts.

TABLE 1
EXAMPLES OF STUDENT WRITING

| Question | Answer | Score |
|---|---|---|
| **Writing 1** | When an object is turned into a temporary magnet, its particles become magnetic and attract magnetic stuff | 5 |
| **Writing 2** | The particles in the paperclip face the same way as the particles in the temporary magnet | 6 |

The students then completed two drawing activities (*Drawing 1* and *Drawing 2*) using foundational symbols. Both drawing exercises utilize the same set of symbols available to the student: paper clip, arrow, straw, magnetic particle, inert particle, and a magnifying bubble. The first drawing prompt instructed students to draw what a paperclip and straw's particles look like when far from a magnet. The second prompt asked what the particles would look like when close to the magnet. For these exercises, the magnet is placed by the system in the drawing space for the student and cannot be manipulated. Finally, the students were again presented with the focus question that began the exercise and asked to construct a written response given what they learned during the activity (*Writing 2*). The ideal answer combines the macroscopic concept that paperclips can undergo induced magnetism when near a permanent magnet with the microscopic reason being the change in the orientation of its particles.

TABLE 2
SHARED RUBRIC FOR WRITING AND DRAWING

| Rubric Item | Writing | Drawing |
|---|---|---|
| Paper clip | 0-1 | 0-1 |
| Straw | 0-1 | 0-1 |
| Magnifier | 0-1 | 0-1 |
| Particles | 0-1 | 0-1 |
| Clip Particles | 0-3 | 0-3 |
| Straw Particles | 0-2 | 0-2 |
| Magnet Particles | 0-2 | 0-2 |
| Dynamic | 0-2 | N/A |
| Semiotic | 0-3 | N/A |
| Total | **16** | **11** |

Instead of developing separate rubrics for evaluating writing and drawing, we used the common content focus of the activity to develop a shared rubric designed to evaluate student responses in both written and graphic form. The rubric evaluates student responses against several criteria. Four of the criteria concern the usage of core 'actors' from the magnetism investigation: paperclips, straws, magnifiers, and particles, and were scored on a 0-1 scale for each actor. The scale attempted to account for more than just the presence of the actor, for example positional requirements for the actors in the drawing space as well as requiring the actor to be part of a complete thought in the writings. Three dimensions were related to the accurate depiction of the particulate nature of permanent magnets, objects that could be magnetized (e.g., paper clips), and nonmagnetic objects (e.g., straws). These criteria were scored on a 0-2 scale for the magnet and straw particles, and a 0-3 scale for the paperclip particles, accounting for both the types and alignment of the particles. Coding of the written responses for the paperclip, magnet, and straw particles were scored on whether students understood the composition of the particles and whether the orientation of its respective particles was fixed or not. Written responses were scored on two additional criteria focusing on the dynamic and symbolic nature of the response. The dynamic dimension scored whether students referenced a change over time. The semiotic dimension indicated whether the nature of the written arguments was evaluated as iconic (only using words to represent concrete ideas) or symbolic (using words representing abstract concepts). These dimensions represented an implicit semantic property that couldn't be assessed in the specific drawing activities studied in this work. For instance, the first example shown in Table 1 received 1 point for mentioning that the particles become magnetic, 1 point for mentioning particles, 1 point for implying change occurred (dynamic), and 1 point for describing abstract concepts. The answer for Writing 2 received 1 point for referencing the paperclip, 3 points for describing how the particles align in the paperclip,

and 1 point for particles. The full rubric is shown in Table 2.

Students for this analysis were selected from a larger sample of implementing classrooms from the 2013-2014 school year. Given the wide range of implementations, a subset of 20 classrooms, representing 14 teachers, was selected due to a high degree of commonality in implementation protocol. Classrooms in this sample had teachers who had completed the professional development session, implemented the system over 10-14 days, had a large percentage of consenting students, and reported only minor technological issues. From this sample, the work of 95 students was analyzed for this work. These students were selected due to their completion of all evaluated written and drawing activities, as well as their completion of both the pre- and post-test assessments. To score the work, two raters coded the graphic and textual artifacts created by students in response to specific prompts in LEONARDO. Inter-rater reliability was calculated via Cohen's kappa (κ) and a protocol for drawing and writing coding using a separate 3-classroom training set featuring students not included in the analyzed corpus before coding the remaining corpus. Coders initially coded a portion of the training set and discussed differences in order to refine the coding process and resolve ambiguities in the initial rubrics. Coders then independently coded an overlapping set of drawings for each question from the three training classrooms and achieved an acceptable level of agreement for each criterion (average $\kappa = .88$) before coding the remainder of the corpus. The procedure was then repeated for the writing prompts, achieving a $\kappa = .76$, after which the remainder of the corpus was coded.

## 4 Analysis of Coded Results

After coding of the written and drawing artifacts was complete, we compared the scores across the two modalities in an effort to analyze how student knowledge was distributed across the two modalities. As shown in Table 3 below, overall students performed better on drawing tasks when expressing their conceptual understanding, scoring noticeably higher, 6.69 for drawings versus 2.48 for writings, despite the writing rubric containing more possible points due to containing two categories not scored for the drawings. Converting these scores to a percentage, students on average scored 60.8% of possible points for drawing responses versus only 15.5% for written responses. This result is perhaps not surprising and aligns with previous research showing younger students are typically much better at illustrating their understanding than providing the same detail with the written word [40]. Additionally, providing the symbols with which students draw provides more built-in support for the drawings than the support provided in writing prompts.

The relationship between drawings and writings was also investigated using a Pearson correlation. Student writing and drawing scores were weakly, but significantly correlated ($r = .299$, $p<.005$). This indicates that while proficiency in the writing and drawing tasks is related, there is also a portion of the students who are capable of demonstrating knowledge through one modality better than the other.

TABLE 3
WRITING AND DRAWING SCORES BY PROBLEM

| Prompts | N | Max | Mean | SD |
|---|---|---|---|---|
| Writing 1 | 95 | 16 | 2.96 | 1.37 |
| Writing 2 | 95 | 16 | 1.99 | 1.37 |
| Drawing 1 | 95 | 11 | 6.64 | 3.46 |
| Drawing 2 | 95 | 11 | 6.74 | 3.17 |

To further evaluate and validate the common rubric approach used to assess the artifacts, we investigated how well the scores on these activities predicted student knowledge. As a proxy for student knowledge, we used their performance on a multiple-choice summative assessment given at the end of the 5-day LEONARDO implementation. We also included the students' performance on a similar instrument given before the implementation to attempt to control for prior knowledge. The instruments were validated through both expert review, as well as a reliability analysis yielding a Cronbach's $\alpha = .77$. A multiple linear regression was conducted with student post-test score as the dependent variable, and student pre-test score, the average of the two writing scores, and the mean of the two drawing scores used as the independent variables. Results for this regression are shown in Table 4.

TABLE 4
REGRESSION MODEL OF HUMAN SCORES

| Variable | B | sig | $sr^2$ | $R^2$ |
|---|---|---|---|---|
| **Model** | | | | **.443** |
| Pre-test | .356 | .000 | .109 | |
| Human-Scored Writings | .216 | .010 | .042 | |
| Human-Scored Drawings | .332 | .000 | .090 | |

The results show a strong relationship between post-test performance and both modalities. Writing score was less predictive than drawing score; however, it nevertheless accounted for 4% of the variance independent of the other two factors and was a significant predictor even when including pre-test. Drawing score was a very strong predictor, accounting for approximately 9% of the variance independently and providing almost as much predictive power as the pre-test score.

## 5 Automated Assessment Methods

Building on the strong results from the human-coding, we sought to develop computational methods to au-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



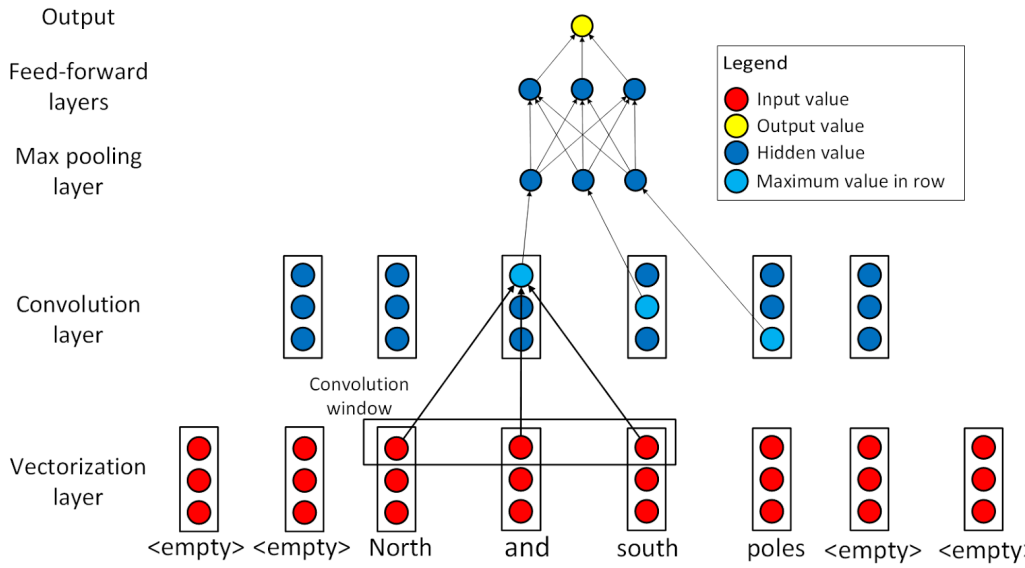


Fig. 4. A Convolutional Neural Network for short answer analysis

tomatically score student answers. We created a multimodal assessment framework that considers the two modalities: 1) student writing, which is assessed with a convolutional neural network (a type of deep learning neural network) for short answer response analysis, and 2) student drawing, which is assessed with a topology-based drawing analysis model.

### 5.1 A Convolutional Neural Network for Short Answer Analysis

To analyze students' written responses, we used a convolutional neural network with max-pooling. A *convolutional neural network* (CNN) differs from a feed-forward network in that it can evaluate inputs of arbitrary length, which is useful in language processing where statements can be anywhere from one word to pages or chapters of text. We select it over a more traditional method such as latent semantic analysis because it takes word order into account and has proven to be effective in recent applications to other text analytics tasks, such as sentiment prediction and question type classification [32]. We also select this method because it automatically learns relevant features and constructs from the text itself, thus requiring no labor-intensive human engineering of features.

Analysis of a student short answer using our CNN is a four-step process: vectorization, convolution, max-pooling, and sending the output to a shallow feed-forward neural network. Fig. 4 shows how our CNN analyzes an example answer, "North and south poles." Each of the circles in this figure represents a single value within the network, often referred to as a node. Starting at the bottom with the input layer, we represent each word as a vector of continuous values in n-dimensional space. These distributed representations of words are learned, either by the model along with its other weights [50], or ahead of time from a large text corpus using an unsupervised technique such as GLoVe [51]. Fig. 4 uses three red nodes above each word to represent a three-dimensional word vector. This array of word vectors makes up the first layer of our model.

The next layer is the convolution layer. This layer is different from a typical feed-forward neural network layer in that it is not fully connected, and the connections it has follow a pattern of shared weights. The convolutional layer is a network of a small fixed size that runs on a moving window across the input. For example, consider the input "North and south poles," where each word is represented as a node. The weight assigned to a node is dependent on where it lies in the window. In this example, we have a window size of three, so our convolutional layer operates on {<empty> <empty> North}, {<empty> North and}, {North and south}, {and south poles}, {south poles <empty>}, and {poles <empty> <empty>}. We may use any size window, but we use three because it is the smallest, and therefore computationally most simple, size that still accounts for context (one word on either side). Note that we pad the ends of our input with dummy values so that the words on the end, "North" and "poles," do not get underrepresented. Each dimension of the word vector space is convolved separately, meaning that the section of the convolution layer corresponding to a given set of three words has the same dimensionality as the word vectors themselves. The convolution layer is thus separated into $w+2$ groups of $n$ values where $w$ is the number of input words and $n$ is the dimensions of the word vector space. Because each dimension in each group is dependent on the same dimension in the three words below it, these groups are effectively hidden representations of semantic meaning, or "hidden words."

The output of the convolution layer is based on the length of the input layer. This is an issue for the following feed-forward layer because each input in the feed-

forward layer is a feature that the network learns to handle in training based on its position. If the features change positions or are added or removed it affects the network's ability to learn. We must extract a fixed number of nodes to send to the feed-forward layers. To do this we use max pooling. For each vector dimension, max pooling simply selects the highest value from among all of the hidden words. We select the maximum value because, in addition to large values naturally having more impact on the output, a trained network will have learned in its convolutional layer to give the most relevant features the highest values.

To increase the representative power of our model we can add more parallel versions of the convolutional layer that have their own weights. These are known as "feature maps." So with $f$ feature maps, an arbitrary window size of $k$, and $n$-dimensional word vectors each pass of the window takes $k$ words and makes $f$ hidden words. Each of these $f$ feature maps goes separately through max-pooling, and together we end up with $f^*n$ values to send to the feed-forward network. As with feed-forward networks, training for a convolutional neural network is done via backpropagation.

A separate CNN was trained for each facet of the rubric, for each of the questions. We built the convolutional neural networks for this task in Theano, a Python-based deep learning library [52]. For our word vectors we used the vectors available on GLoVe's website - 300 dimensions trained on 840 billion tokens in the common crawl corpus [51]. Because many students had written answers to the relevant questions but were missing data elsewhere, we also used these students' answers for training. Our dummy values that padded the ends of each sentence were simply vectors with zeros in every dimension. Words that did not appear in the vectors list we used are also represented with the same dummy value. For training, the objective function is the root mean squared error between human and machine score, which is backpropagated through the network. For more details on CNNs for sentence modeling, see Kalchbrenner et al. [32]. We trained our system using full batch RPROP [53].

Hyperparameters were selected as follows. We selected three as the smallest symmetric window size that still takes advantage of context, i.e., the previous and next word. For the number of feature maps, a sweep of one through nine showed five to be the best-performing. We used only one feed-forward hidden layer, and it has one hidden word in size, i.e., 300 values. After experimenting with the L2 regularization coefficients at 0.0025 and at 0 (no regularization), we found that it was more effective to not regularize. We then tested 50, 100, 150, and 200 as options for epochs, with 50 performing best. These hyperparameters were then used for all models.

### 5.2 Topology-based Drawing Assessment

Building on previous work on automatic assessment of symbolic drawings [30], we sought to emulate human assessment of student drawings through automated analyses of the topological relations between objects in the drawing space. We first defined a set of possible relations between objects for this domain. Because both target drawings used the same set of elements (paperclip, arrow, straw, magnetic particle, inert particle, magnifying bubble, magnet), we were able to use the same set of relations for both drawing prompts. In this domain, the relevant relationships between elements were identified as *near*, *far*, and *contains*. To limit the number of relationships generated, and to help prune irrelevant relationships, elements in the scene were assigned types, with each relation only being generated between objects of certain types. For example, *near* and *far* relations were only generated between straws/paperclips and magnet objects, but not magnifiers or particles. Next, a mapping was created between the 2-dimensional arrangement of the particles and the semantic relations. This mapping was hand-authored by defining thresholds for distance between objects (using bounding boxes and rectangle-to-rectangle distance) and checking for intersections between objects' bounding boxes. Fig. 5 shows an example student drawing and the corresponding topological network. For these questions, *far* was defined as

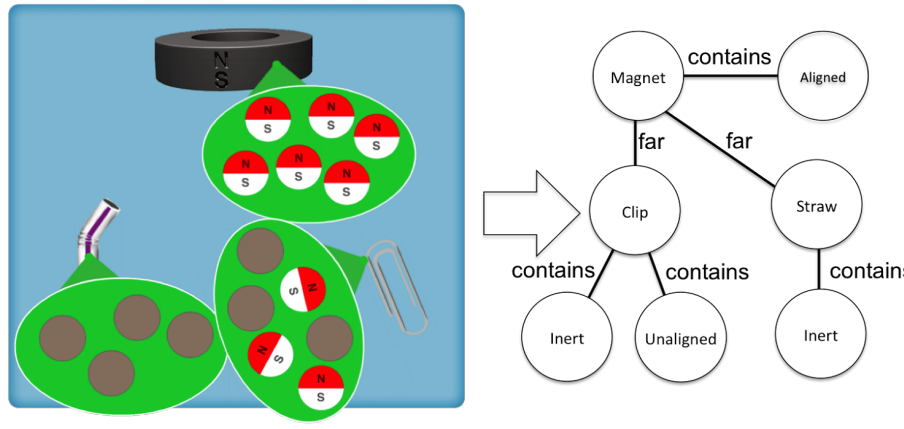


Fig. 5. Student drawing converted into semantic network

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



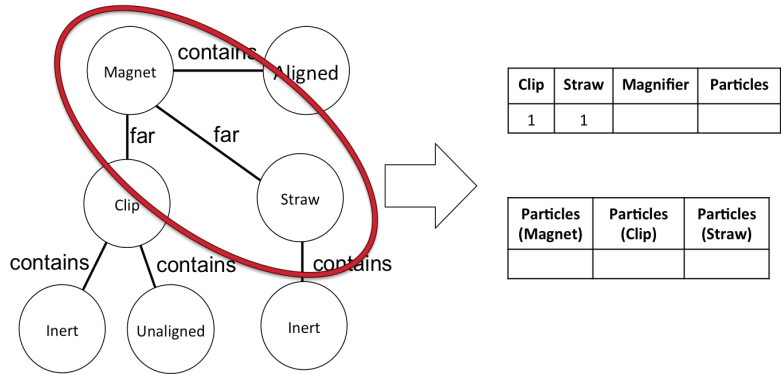


Fig. 6. Generating facet scores from the topological network

farther than 100 pixels from the magnet when calculating the rectangle to rectangle distance between the bounding boxes.

The *contains* relation is more complex and based on multiple 2D relation between objects. For the activities presented here, the particles' correctness can only be evaluated in relation to one of the macro-level objects, such as the straws, paper clips, or magnets. The system assigns particles to these objects by first determining if the particles are contained by a magnifier object. If a particle overlaps with multiple magnifiers, then it is assigned to the one whose center point is closest to the center point of the particle. After being assigned to a magnifier, the system then checks which macro-level objects the magnifier overlaps with. It does this by checking if the region representing each magnifier's magnification point intersects with any such objects. If a magnifier happens to intersect multiple objects, it is also assigned to the object whose center is closest to the center point of the magnifier region. After assigning magnifiers, a *contains* relationship is generated between the magnifier's particles and the macro-level object. Any remaining unassigned particles are then checked to see if they overlap with a straw, a paperclip or the magnet, as some students did not use the magnifier and instead placed particles directly on the objects. The orientation of any magnetic particles contained by the same object is then checked to determine if the group is "aligned," signaling that all particles are rotated to the target rotation, or "unaligned," signaling that at least one particle's rotation does not match the target rotation for this group. For example, the magnetic particles associated with the paper clip in Fig. 5 are oriented in different directions, and are classified as "unaligned," while the magnetic particles associated with the magnet are all oriented close to 0 degrees and are classified as "aligned."

After the final network is completed, it still needs to be converted into a set of scores corresponding with the rubric used in the human coding. To accomplish this task, a set of rules is authored for each facet of the rubric that corresponds to features of the topological graph. For example, in the network shown in Fig. 6 a point would be credited for the presence of a "far" edge between the magnet and the straw in the topological graph. Other facets of the rubric combine multiple rules to produce a score, such as requiring the paper clip to contain both inert particles as well as either aligned or unaligned magnetic particles. This intermediate interpretation allows for flexible scoring of the drawings depending on the specific rubric of interest, as well as providing opportunities for more fine-grained comparison of drawings for misconception identification and clustering.

## 6 Evaluation

The first step of the evaluation of our computational models was to measure how well they align with the human scorings generated in Section 3. The CNN models used to score the writings produced a continuous score for each facet of the rubric. The models were trained to minimize the root mean squared error (RMSE) for each facet. As mentioned earlier, a separate model was trained for each facet and evaluated on only that facet in test data. The models were trained using 10-fold student cross validation, so that each student's writings appeared in exactly one of the test data sets. The overall score represents the summation of all components of the rubric allowing for scores between zero and sixteen, though no student achieved a score above nine. The Pearson correlations between the human and machine scores can be seen below in Table 6 for each of the facets of the writing rubric. The *Straw Placement* and *Straw Particles* facets are marked with *N/A* since no student received points for those categories making a correlation impossible. The system was able to significantly correlate with 3 of the facets in the 1st problem, and 4 in the 2nd prompt, however the total score was significant for both prompts with a p-value $< .0001$. The $r$ values of .466 and .489 are also in range with previous systems on a similar short answer analysis task [4]. Though it may be the result of the system converging to a local optimum, the results are promising given that CNNs typically require larger datasets though is likely due to including the pre-trained word-embeddings that are trained on a massive amount of outside data. Overall, the model is underfitting the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



data, highlighting a need to investigate changes to the model architecture, or potentially replacing or augmenting the input with automatically generated features of text, such as specific word occurrence and sentence length.

A support vector regression (SVR) model was trained for each facet to help determine if the CNN structure was responsible for the performance, or if it was entirely due to the pre-trained word embeddings. The input to the model used the same pre-trained embeddings as the CNN model, padded so that all sentences were the same length, as in the CNN model. A grid search of the $\epsilon$ {0.1, 0.2, 0.5} and C {.5, 1, 2} parameters showed the best performing model to have values of .2 and 1 respectively. The SVR model only produced a significant correlation with 1 facet of Writing 1, and 2 facets of Writing 2, though in all 3 cases it was a negative correlation indicating the model did not learn well. The total score was not significant for Writing 1 ($r$ = -.06, p=.57), though was significant for Writing 2, but with a negative correlation ($r$=-.29, p = .003).

TABLE 5
PEARSON CORRELATION OF AUTOMATED WRITING SCORES AND HUMAN SCORES BY RUBRIC FACET

| Rubric Facet | Writing 1 | Writing 2 |
|---|---|---|
| Clip Placement | .253* | .237* |
| Straw Placement | N/A | -.059 |
| Particles | .006 | .241* |
| Magnet Particles | .205* | .016 |
| Clip Particles | .402** | .285* |
| Straw Particles | N/A | -.150 |
| Iconic/Symbolic | .135 | .081 |
| Static/Dynamic | -.023 | .297* |
| Overall Score | .466** | .489** |

Note. N=95; *p<=.05, **p<.0001

For the drawings, our system produced ordinal scores for each of the rubric criteria. Cohen's κ was calculated to measure agreement between machine and human scorings for each criterion of the two drawings, shown in Table 6. The machine scores showed strong agreement with human codings, producing an average κ =.89 for the first drawing and an average κ = .85 for the second drawing. This result suggests that the drawing assessment model is capable of replicating human scoring with a high level of agreement and also suggest directions for improvement in future work. For example, since the topology is generated from a list of elements placed in the drawing space, it makes no assumptions about occlusion. In several student drawings paperclips, particles, or other elements affecting the machine score were fully obscured from the image viewed by the human grader causing a mismatch in scoring. Other errors involved human coders giving credit for features that were not explicitly correct as defined by the rubric, such as particles placed near objects but not explicitly defined with a magnifier object. Additional errors were observed when human coders would penalize for extra elements in the drawing space that the topological assessment rules ignored.

We next investigated whether the machine-generated scores of written and drawing artifacts were accurate enough to assess learning of conceptual knowledge by the students. As with the regression analysis using the human scores, we used performance on a multiple-choice assessment as a proxy for conceptual knowledge, and looked at the predictive power of the drawing and writing assessments. Separately each assessment was found to be a significant predictor of post-test performance, even when including pre-test performance as a proxy for prior knowledge and to provide a more rigorous standard with which to evaluate performance. We then ran a regression analysis using both drawing and writing, and found both scores to provide both independent and complementary value. The results of each of these models are shown in Table 7. For all models, the dependent variable predicted was the student performance on a summative multiple-choice post-test.

TABLE 6
COHEN'S K FOR DRAWING SCORES BY FACET AND PROBLEM

| Rubric Facet | Drawing 1 | Drawing 2 |
|---|---|---|
| Clip Placement | .95 | .82 |
| Straw Placement | .93 | .71 |
| Particles | .93 | .93 |
| Magnet Particles | .92 | .94 |
| Clip Particles | .75 | .87 |
| Straw Particles | .87 | .87 |
| Magnet Particles | .90 | .80 |
| Average | .89 | .85 |

Both the writing only and drawing only models explain similar amounts of variance, with the drawing scores explaining about 3% more of the variance than the writing scores. Further supporting the complementary value of combining writing and drawing is the over 7% increase in variance explained by the third model containing averages of both scores. *Auto Writing Score* and *Auto Drawing Score* are both significant predictors in the full model, with analysis of the semi-partial $R^2$ values showing writing and drawing uniquely represent 7.2% and 10.4% respectively of the total variance captured by the model. These results suggest that the conceptual understanding expressed in the student writings and drawings are complementary, and that there is additive value in assessment across the modalities.

A potential explanation for the cause of these encouraging results is the "cognitive complementarity" of the two modalities. Writing and drawing utilize different cognitive processes, leading to members of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2799871, IEEE Transactions on Learning Technologies



science education community to advocate the use of science notebooks in the elementary grades because they provide an effective tool for engaging student learning through both modalities [13]. Previous research has shown that students demonstrate different aspects of their scientific knowledge across these sources [14], and not surprisingly, because both drawing and writing shape and reveal underlying student mental models [15], there is a growing recognition that science notebooks offer a potent source of data for formative assessment of students' scientific knowledge.

TABLE 7
REGRESSION MODELS USING AUTOMATED ASSESSMENT MODELS

| Variable | $\beta$ | t | $sr^2$ | Adj. $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| **Pre-Test Only** | | | | **.262** | |
| Pre-Test (PT) | .52 | 5.87* | | | |
| **PT + Writing** | | | | **.367** | **.105** |
| Pre-Test | .464 | 5.58* | .210 | | |
| Auto Writing Score | .337 | 4.04* | .110 | | |
| **PT + Drawing** | | | | **.399** | **.032** |
| Pre-Test | .370 | 4.30* | .118 | | |
| Auto Drawing Score | .405 | 4.71* | .141 | | |
| **Full Model** | | | | **.468** | **.069** |
| Pre-Test | .343 | 4.22* | .101 | | |
| Auto Writing Score | .277 | 3.58* | .072 | | |
| Auto Drawing Score | .353 | .429* | .104 | | |

Note. N=95; *p<=.001

Additionally, it is interesting to observe that both of our automated systems outperform their human equivalents in predicting post-test score, as seen by the performance of the human scores model shown in Table 4. One potential explanation is that for both writing and drawing, the auto-scoring tends to produce scores lower than the human-coded scores. That these scores better predict post-test performance implies there may be aspects of the conceptual knowledge captured in the rubric that are not reflected in the multiple choice post-test assessment.

## 7 CONCLUSIONS AND FUTURE WORK

To investigate the potential of assessment utilizing multiple modalities, we have introduced an integrated multimodal assessment framework. At the foundation of the framework is a shared rubric for evaluating scientific concepts regarding key "actors" in magnetism, as well as more complex interactions between magnetic particles and distance. Evaluation of student writings and drawings found that drawings are more predictive of student conceptual knowledge, and student writings offer a complementary source of diagnostic information.

The next step in integrating the framework was to create automated assessment models. To create the student writing assessment model, we used a convolutional neural network approach leveraging word embeddings to accurately score student short answer questions without requiring any expensive, hand-authored features. While not guaranteed to generate the optimal solution, this approach shows significant potential for rapidly developing accurate short answer scoring systems without extensive feature engineering. The results are particularly encouraging given the high level of misspellings and grammatical errors in the student writing. The student drawing assessment model uses a topology-based approach for drawing analysis. The system accurately produced drawing scores compared to human drawings.

An evaluation shows that 1) both automated methods are capable of assessing student work accurately compared to a human scoring, and that 2) the multimodal assessment framework utilizing both models is predictive of students' post-test performance, even when controlling for prior knowledge. These results suggest that multimodal assessment may be a valuable approach to utilizing the new generation of formative assessment approaches designed to evaluate students' responses formulated in multiple modalities.

There are several limitations to our approach that will need to be addressed. While designed to be open-ended and encourage longer responses, the writing prompts tended to produce short, often ambiguous explanations, as one would expect when working with students in this age range. The prompts should be revised to ensure student artifacts fully align with the aspects of student knowledge we are hoping to assess, such as probing deeper levels of scientific understanding. To aid in this process we will leverage learning activity design techniques such as Evidence-Centered Design[54], to identify what scientific understanding we expect students to show evidence of for in each exercise and to align scoring rubrics accordingly. Additionally, the scoring rubrics for both modalities should be expanded to attempt to better evaluate the thinking and reasoning behind the artifacts, reflective of current assessment frameworks [55], potentially leveraging the drawings to reason about ambiguities in the writings and vice-versa.

With regard to the automated assessment, training a different model for each facet of the writing rubric will not scale well to new questions, and future work should seek architectures capable of outputting multiple facet scores from one model. Additionally, while the writing models have the benefit of being data driven and requiring minimal feature engineering, it will be valuable to compare them directly to existing bag-of-words or linguistic feature-based approaches to potentially identify ways of leveraging the expertise of

the content authors or evaluate the potential of ensemble approaches. Different methods of incorporating known misconceptions or common errors into the model, or rubric, could benefit the system and make the output more meaningful instructors. Finally, in order to scale this approach to larger numbers of questions, it will be necessary to investigate ways of generalizing the model, so that a given model can at least accurately score writings from the same domain using similar vocabulary.

In future work, it will be important to identify the families of modalities that offer the greatest potential synergistic benefits. We anticipate that some combinations of modalities may have overlap in their diagnostic power, while others will exhibit great complementarity. To investigate this complementarity, more detailed rubrics mapping to a wider range of concepts need to be developed and combined with assessment methods that accurately extract this knowledge from the many forms of student artifacts. Regarding writing and drawing specifically, future studies should explore how loosely coupled drawing and writing tasks encourage explicit references between artifacts, and how confounding evidence between the modalities can be used by the system to potentially identify conceptual knowledge that was not expressed due to either poor writing or drawing ability. Moreover, the writing and drawing artifacts can be analyzed to both discover and detect patterns of common misconceptions that could be used by the teacher or automated system to appropriately modify future lessons. These features will be important for investigating the impact of multimodal assessments after they have been integrated into a real-time formative assessment system.

## ACKNOWLEDGMENT

The authors wish to thank our colleagues from the LEONARDO project for their contributions: Courtney Behrle, Mike Carter, and Robert Taylor. This material is based upon work supported by the National Science Foundation under Grant No. DRL-1020229. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

[1] R. E. Bennett, "Formative assessment: a critical review," *Assess. Educ. Princ. Policy Pract.*, vol. 18, no. 1, pp. 5–25, 2011.

[2] K. VanLehn, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," *Educ. Psychol.*, vol. 46, no. 2005, 2011.

[3] V. J. Shute and Y. J. Kim, "Formative and Stealth Assessment," in *Handbook of Research on Educational Communications and Technology*, New York, NY: Springer, 2014, pp. 311–321.

[4] D. T. Tempelaar, A. Heck, H. Cuypers, H. van der Kooij, and E. van de Vrie, "Formative Assessment and Learning Analytics," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 2013, pp. 205–209.

[5] B. Bell and B. Cowie, "The Characteristics of Formative Assessment in Science Education," *Sci. Educ.*, vol. 85, no. 5, pp. 536–553, 2001.

[6] M. A. Sao Pedro, R. S. J. D. Baker, and J. D. Gobert, "What Different Kinds of Stratification Can Reveal About the Generalizability of Data-Mined Skill Assessment Models," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*, 2013, pp. 190–194.

[7] K. E. Arnold and M. D. Pistilli, "Course Signals at Purdue," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 2012, pp. 267–270.

[8] D. Nicol, "E-assessment by Design: Using Multiple-choice Tests to Good Effect," *J. Furth. High. Educ.*, vol. 31, no. 1, pp. 53–64, 2007.

[9] W. Kuechler and M. Simkin, "Why is Performance on Multiple-Choice Tests and Constructed-response Tests not More Closely Related? Theory and an Empirical Test," *Decis. Sci. J. Innov. Educ.*, vol. 8, no. 1, pp. 55–73, 2010.

[10] A. Porter, J. McMaken, J. Hwang, and R. Yang, "Common Core Standards the New US Intended Curriculum," *Educ. Res.*, vol. 40, no. 3, pp. 103–116, 2011.

[11] S. B. Stage, E.K.; Asturias, H.; Cheuk, T.; Daro P.A.& Hampton, "Opportunities and challenges.," *Science (80-. ).*, vol. 340, no. 6130, pp. 276–277, 2013.

[12] S. Jordan and P. Butcher, "Does the Sun Orbit the Earth ? Challenges in using Short Free-Text Computer-Marked Questions .," in *Proceedings of HEA STEM Annual Learning and Teaching Conference 2013: Where Practice and Pedagogy Meet*, 2013.

[13] A. Schmeck, R. E. Mayer, M. Opfermann, V. Pfeiffer, and D. Leutner, "Drawing Pictures during Learning from Scientific Text: Testing the Generative Drawing Effect and the Prognostic Drawing Effect," *Contemp. Educ. Psychol.*, vol. 39, no. 4, pp. 275–286, Jul. 2014.

[14] S. Ainsworth, V. Prain, and R. Tytler, "Drawing to Learn in Science," *Science (80-. ).*, vol. 333, no. 6046, pp. 1096–1097, 2011.

[15] B. Moore and H. Caldwell, "Drama and drawing for narrative writing in primary grades," *J. Educ. Res.*, vol. 87, pp. 100–110, 1993.

[16] R. N. Carney and J. R. Levin, "Pictorial illustrations Still improve students' learning from text," *Educ. Psychol. Rev.*, vol. 14, no. 1, pp. 5–26, 2002.

[17] J. Minogue, E. Wiebe, J. Bedward, and M. Carter, "The Intersection of Science Notebooks, Graphics, and Inquiry," *Sci. Child.*, vol. 48, no. 3, pp. 52–55, 2010.

[18] M. M. Cooper, L. C. Williams, and S. M. Underwood, "Student Understanding of Intermolecular Forces: A Multimodal Study," *J. Chem. Educ.*, vol. 92, no. 8, pp. 1288–1298, 2015.

[19] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," *Int. J. Artif. Intell. Educ.*, vol. 25, no. 1, pp. 60–117, Oct. 2014.

[20] V. Sampson and D. B. Clark, "Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions," *Sci. Educ.*, vol. 92, no. 3, pp. 447–472, 2008.

[21] B. Campbell and L. Fulton, *Science Notebooks: Writing About Inquiry*. Portsmouth, NH: Heinemann, 2003.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. New York, New York, USA: Pearson Prentice Hall, 2008.

[23] A. Graesser, "Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor," *Interact. Learn. Environ.*, vol. 8, no. 2, pp. 1–33, 2000.

[24] S. Bailey and D. Meurers, "Diagnosing meaning errors in short answers to reading comprehension questions," *Proc. Third Work. Innov. Use NLP Build. Educ. Appl. - EANL '08*, pp. 107–115, 2008.

[25] J. Sukkarieh and J. Blackmore, "C-rater: Automatic content scoring for short constructed responses," *Proc. 22nd Int. FLAIRS Conf.*, pp. 290–295, 2009.

[26] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU : a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the Fortieth annual meeting on association for computational linguistics*, 2002, no. July, pp. 311–318.

[27] M. Heilman and N. Madnani, "ETS : Discriminative Edit Models for Paraphrase Scoring," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 2012, pp. 529–535.

[28] M. Heilman and N. Madnani, "ETS : Domain adaptation and stacking for short answer scoring," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, vol. 1, no. SemEval, pp. 96–102.

[29] M. Dzikovska, R. Nielsen, and C. Brew, "Towards effective tutorial feedback for explanation questions: A dataset and baselines," in

*Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 200–210.

[30] S. Leeman-munk, A. Smith, B. Mott, E. Wiebe, J. Lester, and N. Carolina, "Two Modes Are Better Than One : A Multimodal Assessment Framework Integrating Student Writing and Drawing," in *17th International Conference on Artificial Intelligence in Education, AIED 2015*, 2015, pp. 205–215.

[31] W. Yin and H. Schutze, "Convolutional Neural Network for Paraphrase Identification," pp. 901–911, 2015.

[32] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," in *Proceedings of the Fifty-Second Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 655–665.

[33] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-Aware Neural Language Models," 2015.

[34] P. Van Meter and J. Garner, "The Promise and Practice of Learner-Generated Drawing: Literature Review and Synthesis," *Educ. Psychol. Rev.*, vol. 17, no. 4, pp. 285–325, Dec. 2005.

[35] A. Schwamborn, R. E. Mayer, H. Thillmann, C. Leopold, and D. Leutner, "Drawing as a Generative Activity and Drawing as a Prognostic Activity.," *J. Educ. Psychol.*, vol. 102, no. 4, pp. 872–879, 2010.

[36] L. Fiorella and R. E. Mayer, "Eight Ways to Promote Generative Learning," *Educ. Psychol. Rev.*, 2015.

[37] L. Verhoeven, W. Schnotz, and F. Paas, "Cognitive Load in Interactive Knowledge Construction," *Learn. Instr.*, vol. 19, no. 5, pp. 369–375, Oct. 2009.

[38] H. Zhang and M. Linn, "Using Drawings to Support Learning from Dynamic Visualizations," in *Proceedings of the 8th International Conference of the Learning Sciences*, 2008, vol. 3, pp. 161–162.

[39] C. V. Schwarz *et al.*, "Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners," *J. Res. Sci. Teach.*, vol. 46, no. 6, pp. 632–654, Aug. 2009.

[40] J. Minogue, J. C. Bedward, E. N. Wiebe, L. P. Madden, M. Carter, and Z. King, "An Exploration of Upper Elementary Students' Storyboarded Conceptions of Magnetism," in *Paper presented at the NARST Annual Meeting*, 2011.

[41] T. Nelligan, M. Helms, S. Polsley, J. Linsey, J. Ray, and T. Hammond, "Mechanix : A Sketch-Based Educational Interface," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015, pp. 53–56.

[42] L. Bollen and W. van Joolingen, "SimSketch: Multi-Agent Simulations Based on Learner-Created Sketches for Early Science Education," *IEEE Trans. Learn. Technol.*, vol. 6, no. 3, pp. 208–216, 2013.

[43] K. Forbus, J. Usher, A. Lovett, K. Lockwood, and J. Wetzel, "CogSketch: Sketch Understanding for Cognitive Science Research and for Education," *Top. Cogn. Sci.*, vol. 3, no. 4, pp. 648–666, Oct. 2011.

[44] M. Chang and K. Forbus, "Clustering Hand-Drawn Sketches via Analogical Generalization," in *Proceedings of the Twenty-fifth Annual Conference on Innovative Applications of Artificial Intelligence*, 2013, pp. 1507–1512.

[45] B. D. Jee *et al.*, "Drawing on Experience: How Domain Knowledge Is Reflected in Sketches of Scientific Structures and Processes," *Res. Sci. Educ.*, vol. 44, no. 6, pp. 859–883, 2014.

[46] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter, "Intelligent Tutoring Systems with Conversational Dialogue," *AI Mag.*, vol. 22, no. 4, pp. 39–52, 2001.

[47] J. Wetzel and K. Forbus, "Design Buddy : Providing Feedback for Sketched Multi-Modal Causal Explanations," *Twenty Fourth Int. Work. Qual. Reason.*, no. 2, 2010.

[48] CAPSI, "Elementary Science Notebooks: Impact on classroom practice and student," 2006. [Online]. Available: http://www.capsi.caltech.edu/research/Projects.htm#Elementary. [Accessed: 20-Jan-2015].

[49] E. N. Wiebe, J. C. Bedward, and L. P. Madden, "Graphic Representations in Science Notebooks : A Vehicle for Understanding Science Inquiry in the Elementary Classroom," in *Presented at AERA*, 2009.

[50] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 151–161.

[51] J. Pennington, R. Socher, and C. D. Manning, "GloVe : Global Vectors for Word Representation," in *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

[52] F. Bastien *et al.*, "Theano: New Features and Speed Improvements," in *The Deep Learning and Unsupervised Feature Learning Workshop*, 2012, pp. 1–10.

[53] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: the RPROP Algorithm," in *IEEE International Conference on Neural Networks*, 1993, pp. 586–591.

[54] M. Benson, D. Fay, K. L. Kunze, R. J. Mislevy, and J. Behrens, "Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment," *J. Educ. Data Ming*, vol. 4, no. 1, pp. 49–110, 2012.

[55] NGSS Lead States, *Next Generation Science Standards: For States, By States*. Washington DC: National Academic Press, 2013.

**Andy Smith** received B.S. degrees in Electrical Engineering and Computer Science from Duke University in 2005, and a M.C.S. from North Carolina State University in 2010. He has published one journal article and eight conference papers. His research interests focus on leveraging machine learning techniques to understand student drawings.

**Samuel P. Leeman-Munk** received his BA degree in 2010 in computer science from Earlham College, his M.C.S in 2014, and his Ph.D. in 2016, both from North Carolina State University. He now works as part of the Cognitive Computing team at SAS institute. His primary research areas are deep neural networks and natural language processing.

**Angi Shelton** earned the B.S. degree from York College in Secondary Education in 2005, a Master's Degree in Educational Development and Strategies from Wilkes University, in 2007, and the doctorate in Science Education degree in 2012 from Temple University. She has 6 published articles and 28 conference papers. Her research has focused on the intersection of scientific inquiry, assessment, and strategic scaffolding as well as teaching perspectives.

**Bradford Mott** received his B.S., M.C.S., and Ph.D. in Computer Science from North Carolina State University, where he is currently a Senior Research Scientist at the Center for Educational Informatics at North Carolina State University. His research focuses on game-based learning environments, intelligent tutoring systems, computer games, and computational models of interactive narrative.

**Eric Wiebe** is a Professor of STEM Education at North Carolina State University and a senior research fellow at the Friday Institute for Educational Innovation. His research interests include STEM learning in technology-rich environments, multimodal communication of scientific and technical information, and research-based strategies for helping schools and teachers maximize the potential of new instructional technologies.

**James Lester** is a Distinguished Professor of Computer Science at North Carolina State University. His research focuses on technology-rich learning environments and ranges from game-based learning environments and intelligent tutoring systems to affective computing, computational models of narrative, and natural language tutorial dialogue. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).