Two Modes are Better Than One: A Multimodal Assessment Framework Integrating Student Writing and Drawing

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

North Carolina State University, Raleigh, NC 27695 {spleeman, pmsmith4, bwmott, wiebe, lester}@ncsu.edu

Abstract. We are beginning to see the emergence of advanced automated assessment techniques that evaluate expressive student artifacts such as freeform written responses and sketches. These approaches have largely operated individually, each considering only a single mode. We hypothesize that there are synergies to be leveraged in multimodal assessments that can integrate multiple modalities of student responses to create a more complete and accurate picture of a student's knowledge. In this paper, we introduce a novel multimodal assessment framework that integrates two techniques for automatically analyzing student artifacts: a deep learning-based model for assessing student writing, and a topology-based model for assessing student drawing. An evaluation of the framework with elementary students' writing and drawing assessments demonstrate that 1) each of the framework's two modalities provides an independent and complementary measure of student science learning, and 2) together, the multimodal framework significantly outperforms either uni-modal approach individually, demonstrating the potential synergistic benefits of multimodal assessment.

Keywords: Formative Assessment; Multimodal Assessment; Student Writing Analysis; Student Sketch Analysis

1 Introduction

Recent years have seen a growing interest in real-time formative assessment. Recognizing 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 scaffolding and remediation for students, a broad base of research in science education has been investigating the role of formative assessment in instruction [1].

As a tool for formative assessment, short-text constructed response items reveal cognitive processes and states in students that are difficult to uncover in multiplechoice equivalents [2]. Even when it seems that items could be designed to address the same cognitive construct, success in devising multiple-choice and constructedresponse items that behave with psychometric equivalence has proven to be challenging [3]. Because standards-based STEM education in the US explicitly promotes the development of writing skills for which constructed response items are ideally suited, the prospect of designing text analytics techniques for automatically assessing students' textual responses has become even more appealing and has spawned a growing body of research in the area [4].

In a parallel development, drawing is becoming recognized as central activity in science education, particularly in earlier grades. Van Meter and Garner posit that the benefits of student-generated drawing arise from students engaging in three key 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 [5]. A wide range of studies have shown that learning strategies focusing on student-generated drawing can produce effective learning outcomes, such as improving science text comprehension and student engagement [6].

The intelligent tutoring systems community has begun to investigate each of these modalities for assessment. Automated assessment methods for short answer text has been the focus of many studies, with techniques ranging from Latent Semantic Analysis to Soft Cardinality achieving varying degrees of success [4]. Though less common, sketch understanding systems have been used to analyze undergraduate student drawings in a variety of contexts, including clustering via an analogical generalization [7], and drawing-based simulations[8]. However, this prior work has focused on single modality assessment frameworks.

To investigate the potential of multimodal assessment, we explore two research questions. First, we investigated how accurately an individual modality (student writing assessment and student drawing assessment) can automatically assess student artifacts in relation to a gold standard human coding. We found that a convolutional neural network approach for analyzing writing and a topology-based approach for analyzing drawing closely mirror the assessments performed by human graders.

Second, we investigated how accurately a multimodal assessment framework that considers student artifacts from multiple modalities can automatically assess student artifacts. 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 either uni-modal assessment individually.

2 Data Collection and Coding

For the past four years our laboratory has been developing a digital science notebook for elementary school science education, LEONARDO (Figure 1) [9]. Designed to run on both conventional and tablet computing platforms, 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. 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 sketches 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 drawing. While drawing, students choose from a variety of semantically grounded objects and can add, remove, rotate, and move the elements to produce the visual artifact.



Figure 1: LEONARDO Digital Science Notebook

For the data analyzed in this study, student writing and drawing samples were collected from a learning activity in which students investigate what happens to magnetic particles in the presence of a magnetic field. Two writing samples for each student were evaluated. The first 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?" The second written response was collected at the end of the exercise using the same prompt. During the exercise, two drawings were also collected. The first drawing prompt instructed students to draw what a paperclip and straw's particles look like when far from a magnet, and the second prompt asked what the particles would look like when close to the magnet.

To manually assess student learning, a rubric was designed to evaluate student responses in both written and graphic form. This rubric evaluated student responses against several criteria. Four of the criteria concerned the usage of core 'actors' from the magnetism investigation: paperclips, straws, magnifiers, and particles. 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). Written responses were also scored 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 (i.e., only using

words to represent concrete ideas, or symbolic, (i.e., using words representing abstract concepts).

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 3-classroom training set before coding the entire corpus. Coders initially coded a portion of the training set and discussed differences in order to refine the coding process and ambiguities in the rubrics. Coders then independently coded drawings for each question from the three training classrooms and achieved an acceptable level of agreement ($\kappa = .88$) 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.

3 Methods

To explore the hypothesis that multimodal assessment offers the potential to more accurately assess student learning than conventional uni-modal assessment, we created a multimodal assessment framework that considers 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.

3.1 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 conventional 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 [10]. 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. This process is illustrated in Figure 2. The first step, vectorization, consists of taking the input words and converting them into semantic vector representations. These representations are trained along with the model or via unsupervised techniques on large corpora. We used the word vectors available from GLoVe: 300 dimensions trained on 840 billion tokens in the Common Crawl corpus [11]. The second step, convolution, performs an affine transformation on, or *convolves*, sets of adjacent word vectors, defined by a window of a fixed size. For example, for the student answer "north and south poles," our system's convolution layer would take word vectors in groups of three, such as the vector representations of "north and south." In order to avoid words on the ends of the sentence being

underrepresented, we add empty padding values on either side. These convolutions go to the max pooling layer. Max pooling selects the three hundred highest values (based on the length of the word vector). The convolution and max-pooling layers can have multiple copies with different weights, each of which is known as a feature map. Each of these copies generates three hundred values. Finally, the values from the max-pooling layer are concatenated and used as the input layer to a shallow feed-forward neural network. This network outputs a real-valued grade. 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. [10].



Figure 2: A Convolutional Neural Network For Short Answer Analysis

We built the convolutional neural network for this task in Theano, a python-based deep learning library [12]. Because many students had written answers to the relevant questions but were missing data elsewhere, we 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 GLoVe vectors list we used are also represented with the same dummy value.

Hyperparameters were selected based on making a low-dimensional model with small root mean squared error (RMSE). Three is the smallest symmetric window size that still takes advantage of context. In a sweep over one through nine the best-performing number of feature maps was five. We use only one feed-forward hidden layer, and it is one hidden word in size, i.e., 300 values. Out of 50,100,150, and 200 as options for epochs, 50 performed best.

3.2 Topology-based Drawing Assessment

Building on previous work on automatic assessment of symbolic drawings [9], we endeavored to emulate human assessment of drawing evaluation 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. 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. Figure 3 shows an example student drawing and the corresponding topological network. For this question "far" was defined as closer than 100 pixels from the magnet, which is always present at a fixed location in the diagram. The "contains" relation is based on more than one 2D relation between objects.



Figure 3: Student Drawing Converted Into Semantic Network

Because particles could only be evaluated in relation to another object, the system assigns particles by first determining if it can assign a relationship between any magnifiers in the drawing and any straws, paper clips, or magnets. It does this by checking if each magnifier's magnification point intersects with any such objects. If a magnifier happens to intersect multiple objects, the particle is assigned to the object with the closest center. After assigning magnifiers, particles are assigned to the magnifier that they overlap, deciding shared overlaps based on closeness to the magnifier's center. 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. Particles assigned to the same parent object are split into two groups, inert and magnetic. The orientation of the magnetic particles 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. After the final network is completed, it can be queried to generate scores based on the expertdefined rubric. For example in the network shown in Figure 3, points would be credited for the presence of a "far" edge between the magnet and paperclip, a "contains" edge between aligned particles and magnet, as well as the other connections relevant to the rubric described in Section 2.

4 Evaluation and Discussion

To evaluate how well the uni-modal assessment models' performance levels compare to the gold standard human scoring, machine-generated scores were compared to human scores. To measure the accuracy of the continuous scores generated for writing, root mean squared error (RMSE) was used. Although the rubric described earlier allowed for scores between zero and sixteen, no student achieved a score above nine. The deep learning model produced a RMSE of 1.23, or 13% when normalized by the observable range across the 190 student answers analyzed (95 students, 2 answers per student). A Pearson correlation between the human and machine writing scores was conducted with an r of .53 (p < .001), in range with scores from previous systems on a similar task [4]. Further analysis of our model's RMSE shows that is greatly impacted by a small number of outliers in the human grades with 5 answers out of 190 accounting for 22.7% of the total squared error.

For the drawings, our system was able to produce scores for each of the seven rubric criteria. Cohen's κ was calculated to measure agreement between machine and human scorings for each criteria of the two drawings with an average $\kappa = .89$ for the first drawing and an average $\kappa = .85$ for the second drawing. This result suggests that the drawing assessment model is capable of replicating human scoring with a high level of agreement.

The encouraging results also suggest directions for 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.

The next question we investigated was whether machine-scored written and drawing artifacts are useful predictors of student conceptual knowledge. We first looked at the predictive power of drawing and writing separately and found them to be significant predictors, even when controlling for pre-test. We next built a model combing the predictors and found that, even when combined, both scores provide unique and complementary predictive value. These results are summarized in a series of multiple linear regression models shown in Table 1. For all models, the dependent variable predicted was the student performance on a summative multiple-choice posttest. The independent variable pre-test represents student performance on a 20-question multiple-choice assessment administered before students used LEONARDO. We include the pre-test score as a covariate in our analysis as a proxy for prior knowledge and to provide a more rigorous standard for our model to meet. Auto Writing Score and Auto Drawing Score represent the average of the machine-generated scores for the two exercises of each modality.

By themselves, writing score and drawing score are both significant predictors of post-test performance, even when controlling for pre-test performance. Both models explain similar amounts of variance, with Drawing providing slightly more predictive value. Further supporting the value of combining writing and drawing is the almost 8% increase in variance explained by the third model containing averages of both scores as well as the pre-test score. Writing and Drawing Score are both significant predictors in this model, with analysis of the semi-partial R^2 values showing that while there is some common variance captured by the different factors, writing and

drawing uniquely represent 8% and 11% respectively of the total variance captured by the model. These results suggest that the level of conceptual understanding in the student writings and drawings are complementary, and that there is additive value in assessment across multiple modalities.

A potential explanation for the cause of these encouraging results is the "cognitive complementarity" of the two modalities. Recognizing that writing and drawing exercises different cognitive processes, the science education community advocates the use of science notebooks in the elementary grades because they provide an effective tool for promoting learning through both writing and drawing [13]. Prior research has demonstrated that students' scientific knowledge is distributed across both of 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.

| Variable | В | t | sr ² | R ² | ΔR^2 |
|--------------------|------|---------|-----------------|----------------|--------------|
| Pre-Test Only | | | | .270 | .270 |
| Pre-Test | .52 | 5.87*** | | | |
| Pre-Test + Writing | | | | .380 | .110 |
| Pre-Test | .439 | 5.20*** | .182 | | |
| Auto Writing Score | .34 | 4.03*** | .110 | | |
| Pre-Test + Drawing | | | | .412 | .032 |
| Pre-Test | .37 | 4.30*** | .118 | | |
| Auto Drawing Score | .405 | 4.71*** | .141 | | |
| Full Model | | | | .491 | .079 |
| Pre-Test | .317 | 3.87*** | .084 | | |
| Auto Writing Score | .292 | 2.56*** | .079 | | |
| Auto Drawing Score | .362 | .426*** | .112 | | |

Table 1: Regression Models Using Machine Assigned Scores

Note. N=95; *p<.05, **p<.01, ***p<.001

Table 2: Regression Model of Human Scores

| Variable | | Sig | sr ² | \mathbf{R}^2 |
|----------------------|------|------|-----------------|----------------|
| Model | | | | .451 |
| Pre-test | .356 | .000 | .109 | |
| Human Writing Scores | .216 | .010 | .040 | |
| Human Drawing Scores | .332 | .000 | .090 | |

One particularly curious result is that both of our automatic systems outperform their human equivalents in predicting post-test score, as seen by the predictive power of the human scores in Table 2. As one possible explanation on the writing side, it might in fact be an advantage rather than a disadvantage that our system dampens some of the high variance found in the human scores.

7 Conclusions and Future Work

Formative assessment is a crucial part of the instructional process, enabling both teachers and students to evaluate conceptual understanding and misconceptions. Similarly, automated assessment methods are emerging that can evaluate students' understanding across an increasingly wide range of modalities. These modalities are typically studied in isolation, with research often stopping at measuring the reliability and validity of a given assessment. However, there is great potential in better understanding how the different modalities work in consort. For example, automated writing and drawing assessment each provide meaningful insights into student science understanding. Together, writing and drawing assessment have the potential to provide a much more nuanced picture of student science comprehension than either alone.

To investigate the potential of assessment with multiple modalities, we have introduced an integrated multimodal assessment framework. The multimodal assessment framework has been studied in the context of science education with a student writing assessment model that uses a convolutional neural network approach and a student drawing assessment model that uses a topology-based approach for drawing analysis. An evaluation shows that 1) both 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 more than a single mode.

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 be exhibit great complementarity. Future data collections will focus on more closely coupling the drawing and writing tasks and encouraging explicit references between artifacts. It will also be important to empirically investigate how multimodal assessment can be integrated into a real-time formative assessment system and used as the basis for generating personalized scaffolding.

Acknowledgments

The authors wish to thank our colleagues from the LEONARDO project for their contributions to the design, development, and classroom implementations of LEONARDO: Courtney Behrle, Mike Carter, Angela Shelton, 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 are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- 1. Abell, S., Lederman, N.: Handbook of Research on Science Education. Routledge, New York, NY (2007).
- 2. Nicol, D.: E-assessment by Design: Using Multiple-choice Tests to Good Effect. Journal of Further and Higher Education. 31, 53–64 (2007).
- Kuechler, W., Simkin, M.: Why is Performance on Multiple-Choice Tests and Constructed-response Tests not More Closely Related? Theory and an Empirical Test. Decision Sciences Journal of Innovative Education. 8, 55–73 (2010).
- Burrows, S., Gurevych, I., Stein, B.: The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education. 60– 117 (2014).
- Van Meter, P., Garner, J.: The Promise and Practice of Learner-Generated Drawing: Literature Review and Synthesis. Educational Psychology Review. 17, 285–325 (2005).
- Schmeck, A., Mayer, R.E., Opfermann, M., Pfeiffer, V., Leutner, D.: Drawing Pictures during Learning from Scientific Text: Testing the Generative Drawing Effect and the Prognostic Drawing Effect. Contemporary Educational Psychology. 39, 275–286 (2014).
- Chang, M., Forbus, K.: Clustering Hand-Drawn Sketches via Analogical Generalization. Proceedings of the Twenty-fifth Annual Conference on Innovative Applications of Artificial Intelligence. pp. 1507–1512. Bellevue, WA (2013).
- Van Joolingen, W., Bollen, L., Leenaars, F.: Using Drawings in Knowledge Modeling and Simulation for Science Teaching. Advances in Intelligent Tutoring Systems. pp. 249–264 (2010).
- Smith, A., Wiebe, E., Mott, B., Lester, J.: SketchMiner : Mining Learner-Generated Science Drawings with Topological Abstraction. Proceedings of the Seventh International Conference on Educational Data Mining. pp. 288–291. London, U.K. (2014).
- Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A Convolutional Neural Network for Modelling Sentences. Proceedings of the Fifty-Second Annual Meeting of the Association for Computational Linguistics. pp. 655–665. Baltimore, MD (2014).
- Pennington, J., Socher, R., Manning, C.D.: GloVe : Global Vectors for Word Representation. Proceedings of Empiricial Methods in Natural Language Processing. Doha, Quatar (2014).
- Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., Bengio, Y.: Theano: New Features and Speed Improvements. The Deep Learning and Unsupervised Feature Learning Workshop. pp. 1–10. Lake Tahoe, CA (2012).
- Campbell, B., Fulton, L.: Science Notebooks: Writing About Inquiry. Heinemann, Portsmouth, NH (2003).
- Minogue, J., Wiebe, E., Bedward, J., Carter, M.: The Intersection of Science Notebooks, Graphics, and Inquiry. Science and Children. 48, 52–55 (2010).
- 15. Schnotz, W., Bannert, M.: Construction and Interference in Learning from Multiple Representation. Learning and Instruction. 13, 141–156 (2003).