Integrating Real-Time Drawing and Writing Diagnostic Models: An Evidence-Centered Design Framework for Multimodal Science Assessment

Andy Smith[1], Osman Aksit, Wookhee Min, Eric Wiebe, Bradford W. Mott, and James C. Lester

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

Abstract. Interactively modeling science phenomena enables students to develop rich conceptual understanding of science. While this understanding is often assessed through summative, multiple-choice instruments, science notebooks have been used extensively in elementary and secondary grades as a mechanism to promote and reveal reflection through both drawing and writing. Although each modality has been studied individually, obtaining a comprehensive view of a student's conceptual understanding requires analyses of knowledge represented across both modalities. Evidence-centered design (ECD) provides a framework for diagnostic measurement of data collected from student interactions with complex learning environments. This work utilizes ECD to analyze a corpus of elementary student writings and drawings collected with a digital science notebook. First, a competency model representing the core concepts of each exercise, as well as the curricular unit as a whole, was constructed. Then, evidence models were created to map between student written and drawn artifacts and the shared competency model. Finally, the scores obtained using the evidence models were used to train a deep-learning based model for automated writing assessment, as well as to develop an automated drawing assessment model using topological abstraction. The findings reveal that ECD provides an expressive unified framework for multimodal assessment of science learning with accurate predictions of student learning.

Keywords: Assessment**,** multimodalilty, evidence-centered design

1 Introduction

Formative assessment can play a central role in enabling intelligent tutoring systems (ITSs) to provide students with personalized, adaptive learning experiences [1]. Effective formative assessment can be used to infer students' underlying mental models as well as their movement through learning progressions [2, 3]. The models inferred from these assessments can then be used as the basis for real-time feedback and adaptive support [4]. Formative assessment can improve science learning, and because science learning often features both drawing and writing activities, intelligent

[1] Corresponding Author: Andy Smith, Department of Computer Science, North Carolina State University

tutoring systems for science education should support multimodal assessment of both student drawing and student writing [5].

Evidence-centered design (ECD) provides a systematic approach to designing and developing assessments [6]. ECD identifies multiple phases in the design process, each with its own explicit goals. These phases include the creation of a Competency Model, an Evidence Model, and a Task Model that operate in concert to recognize evidence of conceptual understanding from student work. For multimodal assessment, ECD can provide a systematic way of mapping between learning goals and student artifacts from various modalities that show evidence of student learning. Of particular interest is how ECD might provide a unified framework for assessing both written and drawn artifacts of student work for formative purposes.

This paper introduces a new ECD-based framework for multimodal science assessment. First, we use a multimodal approach to ECD to define a competency model and a multimodal evidence model for elementary science to understand how conceptual understanding about magnetism is revealed in both drawing and writing tasks. Specifically we aim to evaluate student writings and drawings using a common competency model that contributes to a deeper understanding of the relative contributions of the two modalities. Second, with the long-term goal of integrating multimodal assessments into an ITS, we present computational models for evaluating student writings and drawings in real-time and compare their predictive accuracy to expert human scorings. The findings reveal that ECD provides a unified framework for multimodal assessment of science learning with accurate predictions of student learning.

2 Related Work

Though much less investigated than short-answer writing assessment, there has been some work on assessment of learner-generated drawings. Mechanix [7] utilizes free-hand sketch recognition to convert student drawings in the domain of statics into free-body equations that the system can then analyze and provide corrective feedback. Van Joolingen et al.'s SimSketch system seeks to merge free-hand sketching with modeling science phenomena. The system first segments the free-hand drawing into distinct objects that can be annotated by the user with a variety of behaviors and attributes [8]. Students can then run a simulation based on those behaviors and attributes. SimSketch was used in a planetarium setting by elementary students for modeling and simulation, showing evidence for increasing student learning and engagement. CogSketch [9], which aims to support open-domain sketch understanding, has been employed to compare the drawings of expert and novice users to analyze differences in drawings, as well as differences in the ways the drawings are created.

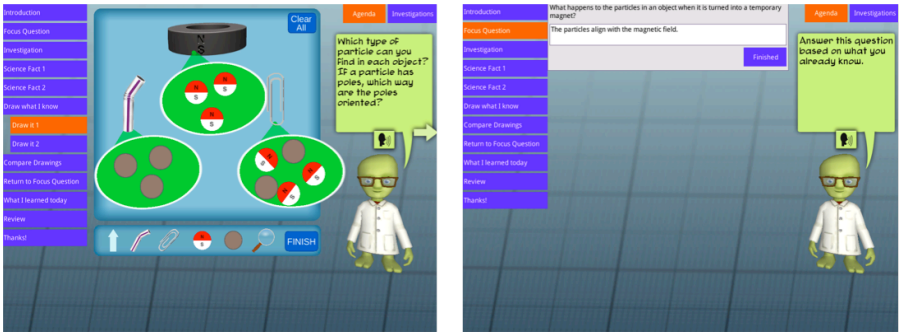
Automatic grading of written short answers has long been the focus of the ITS and natural language processing (NLP) communities, with short answers being defined as natural language responses varying in length from one sentence to one paragraph [10]. Many of these approaches, such as the widely used Latent Semantic Analysis [11], rely on "bag-of-words" approaches that focus primarily on the occurrence or

frequency of words that appear in text. Other approaches, such as the ones embodied in Educational Testing Service's C-Rater, use a variety of preprocessing techniques to generate syntactic relationships between words in a sentence [12]. The technique employed by Dzikovska et al. uses dependency parses in a facet-based approach to assessment, which provides more fine-grained information about assessments than a monolithic overall score [13]. Other approaches have used word embeddings and convolutional neural networks that incorporate information across sequences of words [14]. Our work proposes an approach combining word conversion techniques and feedforward neural networks to address noisy students' answers that contain various forms of misspellings to implement a reliable writing assessment solution.

Recent years have also seen a growing interest in evidence-centered design as a method for interpreting the complex data streams generated by virtual learning environments. Gobert et al. used ECD to create predictive models of student inquiry skills from action logs generated in a science microworld [15]. Rupp et al. utilized ECD in both the design of an interactive training application for employees of a networking company, as well as the design of the accompanying assessments [16]. Finally, ECD is used in conjunction with computational methods such as Bayesian networks and stacked autoencoder networks to construct "stealth" assessments for educational games [2, 17]. Our work builds on this line of investigation by introducing a unified framework for *multimodal assessment* of both drawing and writing based on ECD.

3 The LEONARDO Digital Science Notebook

Data for the work reported here was collected with LEONARDO, a cloud-based digital science notebook developed for elementary school science education [14]. LEONARDO was designed for use in the classroom and runs on both desktop computers and tablets. LEONARDO supports inquiry learning by providing adaptive support to students as they engage in both virtual and physical lab activities as well as providing them with tools to create their own visual and written representations (Figure 1). LEONARDO currently supports three science units: *Electricity*, *Magnetism,* and *Weather*. Each unit consists of several subunits driven by Focus Questions (FQs), which are organized around an open-ended driving question (e.g., What makes a magnet magnetic?). The activities and tasks employed in each FQ were designed to facilitate student learning of the underlying science concept to answer the driving question.



(a) (b)

Fig. 1 Examples of LEONARDO drawing (a) and writing (b) prompts.

In most FQs, students are required to construct written and visual explanations by completing a series of drawing and writing tasks to solidify and extend their understanding of the observed scientific phenomena. To facilitate meaningful writing composition, the writing tasks require students to compose short responses, and in some cases a starter prompt is given to help students build an argument sentence, (e.g., A magnet attracts a paperclip because...). In drawing tasks students manipulate built-in pictorial symbols representing key scientific concepts relevant to the current FQ. Pictorial symbol manipulation includes selecting appropriate symbols from a toolbar and organizing them in the drawing field by modifying their direction, alignment and relative placement. At the end of each FQ, students are presented with their original answers to the driving question and offered to construct a new response based on what they have learned so that they can recognize and monitor the change in their own understanding of the subject matter by comparing their old and new response.

4 ECD Coding Framework

Evidence-centered design is a holistic approach to designing, implementing, evaluating, and delivering educational assessments [18]. ECD recognizes assessment as an evidentiary reasoning process that entails making arguments on learning based on the limited evidence provided by the learner [6]. ECD has gained considerable popularity in a broad range of fields in recent years, and it has been used in conjunction with several forms of learning technologies, including game-based learning environments [2], and educational data mining [1,16]. The ECD framework formalizes the different phases in the assessment design process as "layers" and each layer has its own specific objectives and associated products. In this work, we employed ECD's Conceptual Assessment Framework layer to analyze our assessment models in LEONARDO's magnetism unit, and we generated a comprehensive rubric to score students' drawing and writing artifacts based on this analysis.

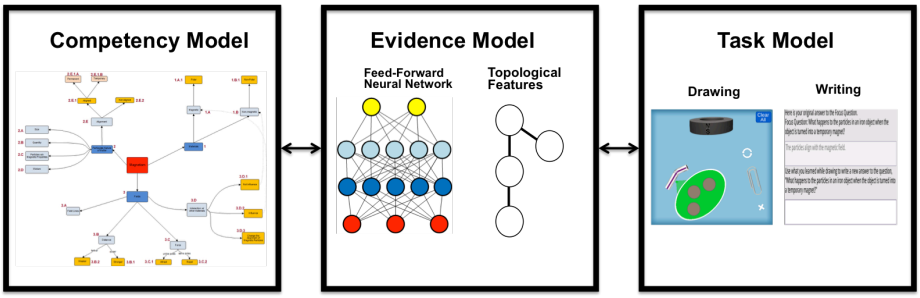


Fig. 2 Application of the ECD process to science notebook task data.

The Conceptual Assessment Framework layer consists of three components: the Competency Model, the Evidence Model, and the Task Model (Fig. 2). The first layer begins by identifying and determining what collection of knowledge, skills or practices on which the learner will be assessed. These concepts are then combined to

form the Competency Model, sometimes referred to as the Student Model. Once defined, values in the Competency Model can be inferred across multiple interactions using a variety of techniques including Bayesian knowledge tracing and dynamic belief networks [4]. The second step is determining what types of observations of student work or artifacts will provide measurable evidence for the target competencies, including defining specific evidence for each of the modalities to be evaluated. The Evidence Model is the product of this layer. The final step focuses on designing tasks—the Task Model—that will give the learner relevant opportunities to provide the expected evidence. When the possible evidence that students may exhibit have been identified, the tasks can then be designed that will require students to generate those evidence. Mislevy and Haertel note that ECD is a sequential process but can include iterations and refinements within and across the layers during the design cycle [6].

To develop our models we used a subset of a larger sample of fourth grade students from the 42 schools who implemented LEONARDO's magnetism module during the 2013-2014 and 2014-2015 school years. The participating schools are located across the United States. A total of 98 students from 19 different classrooms were selected based on the requirement that they completed all eight of the drawing and writing tasks in FQ-3 and FQ-5 in the Magnetism unit. Although there are six instructional units (FQs) in the magnetism module, we chose to analyze FQ-3 (What happens to the particles in an iron object when the object is turned into a temporary magnet?) and FQ-5 (Can magnets work through materials like paper, cardboard, and metal foil?) because they provide the richest set of drawing and writing tasks in terms of the number and variety of the scientific concepts that they address.

An initial coding was completed by two human raters individually grading students' drawing and writing artifacts. Initial practice trials were completed using data from students not included in the final sample to train raters, formalize the rubric, and align their interpretations. Cohen's kappa (κ) was run to determine the inter-rater reliability based on a randomly selected subset of 20% of responses coded by both raters. A high level of agreement was found between the two raters' drawing scores, $\kappa = .838$ (95% CI, .806 to .869, $p < .001$) and a substantial level of agreement between the two raters' writing scores, $\kappa = .754$ (95% CI, .669 to .838, $p < .001$).

Table 1. Means* and Standard Deviations for Total Scores (N = 98)

Questions	Min	Max	Mean	SD
FQ-3 Drawings	0	100	62.6	32.9
FQ-5 Drawings	4	100	60.8	26.2
FQ-3 Writings	0	88	28.5	23.6
FQ-5 Writings	10	100	63.3	20.9
Post-Test	20	95	68.0	20.4

*Scores are converted to a 0-100 scale for ease of interpretation.

Table 1 shows the students' drawing and writing scores and post-test performance. Although the mean scores of FQ-3 and FQ-5 drawings and FQ-5 writings are close to each other, the mean score of FQ-3 writings is much lower than the others. This might be explained by the fact that one of the FQ-3 writing tasks asks students to compare their two drawings, and thus has a higher number of potential concepts to be observed

than the other writings. However, most students' responses compared only one or two aspects of their drawings resulting in the lower scores. A hierarchical multiple regression test was conducted to analyze how student knowledge revealed by multiple drawing and writing artifacts predict their post-test performance. The first model, which uses only FQ-3 and FQ-5 drawing scores, significantly predicted approximately 36% variance in the post-test scores $F(2, 95) = 27.17$, $p < .001$, $R^2 = .364$, while the second model containing FQ-3 and FQ-5 both drawing and writing scores significantly predicted about 48% variance in total in the post-test scores $F(4, 93) = 27.75$, $p < .001$, $R^2 = .483$, producing an R^2 change of .119.

5 Automated Assessment Systems

With the goal of integrating these new assessments into LEONARDO, we used the human scorings to devise computational assessment models to assess both student drawing and writing. We next introduce the drawing scoring, using a rule-based system based on topological features, as well as the writing scoring, using word conversion techniques combined with feedforward neural networks.

5.1 Automated Assessment of Symbolic Drawings

Building on techniques developed in our previous work [14], drawings are represented as a set of objects and their associated x, y coordinates and rotation. For example, the set of possible objects in the drawing space include a paper clip, a plastic straw, a magnifier bubble to indicate microscopic properties, inert particles, magnetic particles, and an arrow. For this work we decompose the drawing into a set of topological relations between these objects. Topological features allow us to discretize a wide range of continuous features in a way that facilitates symbolic manipulation. In this case we use these topological relations to generate a labeled graph representation of the drawing. The first step in the translation from drawings to topological graphs is encoding the primary elements for the domain. Initially, this consists of defined elements drawn by the student. In the later steps these elements can be combined into new elements, such as converting a group of similarly rotated magnetic particles into a single "aligned particles" element.

After creating the nodes of the graph, edges are generated based on topographical relationships between elements. Many potential 2D relationships are encoded, with the goal of generating a sufficiently large number of relationships to capture the relevant information expressed by the drawing, while excluding irrelevant relationships that will unnecessarily complicate the computation. For example, one solution could be to generate a complete set of all possible relations for every pairwise combination of elements in the drawing, though, this approach would quickly produce a large amount of features, many of which are unnecessary for the intended analysis. To simplify this task, each object is assigned a type. For each type, we specify a set of related types for which topological relationships will be generated. The set of qualitative 2D relations used in this work are *near, far, intersects*, and *aligns-with*. Finally, more complex relationships are defined based on combinations

of atomic spatial relations. For example, the point of the magnifier object intersecting with a paperclip generates a set of *contains* relationships between the paperclip and the elements that have been drawn within the larger magnification bubble.

Finally, to convert the symbolic representation into a rubric score, we assign a set of rules for each rubric component. For example, for the component associated with the concept of a straw containing only non-magnetic particles, a *contains(straw, inert)* relationship must exist, as well as *contains(straw, aligned)* and *contains(straw, unaligned)* not existing. These rules can also be defined to compare relationships between drawings, as is required by some components of the competency model.

5.2 A Feedforward Neural Network for Short Answer Analysis

Building automated writing assessments entails devising computational models that take as input students' text-based responses and predict as output their grades according to the pre-specified rubric discussed in Section 4. A key challenge posed by the automated assessment of elementary-grade students' writing is effectively dealing with many forms of misspelled words, including cognitive misconceptions (e.g., *magnetism* misspelled by *magnetizm*) and typographical errors (e.g., *paperclip* misspelled by *paperrclip*). Misspellings caused by cognitive misconceptions tend to persistently appear in the student's writing, whereas typographical errors, such as injecting an extra character or mistakenly typing a neighboring character, occur in other places less frequently. To address this challenge, we implement a two-step writing assessment system, in which the system first creates a dictionary to convert similar words to the same representative word using Levenshtein edit distance and then trains classifiers based on a bag-of-words representation based on the induced dictionary.

For computational writing assessment models, we utilize feedforward neural networks. Deep neural networks, often called *deep learning* [19], have demonstrated considerable success for a wide range of computational challenges, such as computer vision, natural language processing, and speech recognition. A model is trained per short-answer question. Since every writing question has multiple labels (i.e., competencies) to predict, this task is cast as multi-label classification. The hyperparameters for neural networks are often empirically determined using grid search [14]. In this work, we explore the number of hidden units using 256 and 512, and the number of hidden layers from 1 to 4. We fix the following parameters: setting all the activation functions to sigmoid, adopting the dropout regularization technique [19] with the dropout rate of 0.5, and using binary cross entropy and stochastic optimization for the loss function and optimizer, respectively.

6 Evaluation

To evaluate the assessment models, we conducted validation studies with the corpus of fourth grade writings and drawings collected with the LEONARDO system. The drawing models were assessed using 4 drawings each from the 98 students scored by human coders. For each drawing, rules mapping between topological features and

competency scores were authored based on notes from the rubrics used by human scorers and from tuning on a scored set of drawings not used in the evaluation sample. As the drawing models used authored rules and were not machine-learned, cross validation was not used. The baseline accuracy rate is calculated by computing the most common class rate per competency, and then averaging across all the competencies within each question.

Table 2. Automated drawing assessment results (N = 98)

Question	Concepts	Accuracy	Baseline
FQ3 – Drawing 1	11	90.4%	66.8%
FQ3 – Drawing 2	13	87.9%	62.8%
FQ5 – Drawing 1	12	86.3%	61.8%
FQ5 – Drawing 2	13	90.8%	61.6%

As shown in the table, the models performed well compared to the baseline. Analysis of the classification errors showed a small number of cases where the automated model incorporated elements that were occluded from the drawing presented to human coders. The majority of the error cases were the result of the system not giving credit for a concept for which the human coders gave credit. These types of errors could be potentially corrected by creating more scoring rules, though many would be difficult to author without incurring an unacceptable level of false positives.

For the writings, four feedforward networks were trained for each of the four questions (2 for FQ-3 and 2 for FQ-5), adjusting the number of hidden layers from one to four. Each model was evaluated using a 10-fold student-level cross validation. The accuracy levels shown in Table 3 represent the average accuracy rates across all competencies for the question. The baseline accuracy rate was calculated using the same process as for the drawings. The accuracy rate of neural networks that achieve the highest predictive performance in the 10-fold cross validation is reported along with the number of hidden layers the models leverage.

Table 3. Automated writing assessment results (N = 98)

Question	Hidden Layers	Concepts	Accuracy	Baseline
FQ3 –Writing 1	1	13	78%	71.4%
FQ3 –Writing 2	1	5	73%	62.6%
FQ5 –Writing 1	1	3	90%	82%
FQ5 –Writing 2	1	7	76%	65.3%

Overall the writing assessment system performed very well with accuracies ranging from 73% to 90% for the 4 questions. While the shallow networks exhibited the best overall performance for each question, the accuracies of the other models performed very similarly, often less than 1% different. This result is perhaps not surprising given that deep networks are more likely to suffer from overfitting when trained with small datasets [19]. The high baseline accuracy for the first writing sample in FQ-5 suggests that that question in particular may have been over-scaffolded and should be revised in future implementations. Further analysis of the errors reveals the majority are likely due to the high level of misspellings and

grammatical errors in the text, indicating that while the steps taken to cope with noisy text were effective, there is room for improvement.

7 Conclusions and Future Work

Multimodal assessments that operate on both student drawing and student writing hold great potential for expanding the diagnostic power of ITSs. ECD provides a unifying framework for multimodal assessment by defining targeted learning concepts of a given exercise, and for identifying evidence of those concepts in student work that includes both drawing and writing. We hypothesized that a unified ECD-based multimodal assessment framework would support the design of computational models of assessment that could operate on both drawing and writing.

In this paper, we introduced a framework for applying ECD to multimodal student learning. First, a competency model is defined, identifying scientific concepts of interest. Next, rubrics are created to define which features of student writings and drawings constitute evidence of the previously defined competencies. Using the rubrics we then found that the evidence measured from both drawing and writing were significantly predictive of performance on a multiple-choice summative post-test. We also found that students were generally able to express more concepts through drawing than writing, although this could be related to the inherent scaffolding afforded by the symbolic drawing tasks. Finally, with the long-term goal of incorporating automated multimodal assessments into interactive learning environments such as the LEONARDO digital science notebook, we developed computational methods for the real-time automated assessment of student drawing and writing artifacts. An evaluation of the resulting multimodal assessment framework found that the models outperformed baseline models in accurately assessing student work across multi-faceted rubrics for both modalities.

In future work it will be important to further refine the automated assessment techniques to increase their accuracy. A second promising line of investigation is to use ECD to better understand how student knowledge of low-level concepts relates to higher-order concepts. Finally, it will be important to investigate how to best incorporate multimodal assessment into an ITS and utilize real-time assessment results to drive personalized feedback and scaffolding.

Acknowledgements

This work is supported in part by the National Science Foundation through Grant No. DRL-1020229 and the Social Sciences and Humanities Research Council of Canada. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation or the Social Sciences and Humanities Research Council of Canada.

References

1. VanLehn, K.: The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. Educational Psychologist. 46, 197–221 (2011).
2. Shute, V.J., Kim, Y.J.: Formative and Stealth Assessment. In: Handbook of Research on Educational Communications and Technology. pp. 311–321. Springer (2014).
3. Bennett, R.E.: Formative assessment: a critical review. Assessment in Education: Principles, Policy & Practice. 18, 5–25 (2011).
4. Desmarais, M.C., Baker, R.S.J.D.: A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction. 22, 9–38 (2011).
5. Minogue, J., Wiebe, E., Bedward, J., Carter, M.: The Intersection of Science Notebooks, Graphics, and Inquiry. Science and Children. 48, 52–55 (2010).
6. Mislevy, R.J., Haertel, G.D.: Implications of Evidence-Centered Design for Educational Testing. Educational Measurement: Issues and Practice. 25, 6–20 (2006).
7. Nelligan, T., Helms, M., Polsley, S., Linsey, J., Ray, J., Hammond, T.: Mechanix : A Sketch-Based Educational Interface. In: Proceedings of the 20th International Conference on Intelligent User Interfaces. pp. 53–56. ACM (2015).
8. Bollen, L., Joolingen, W. van: SimSketch: Multi-Agent Simulations Based on Learner-Created Sketches for Early Science Education. IEEE Transactions on Learning Technologies. 6, 208–216 (2013).
9. Jee, B.D., Gentner, D., Uttal, D.H., Sageman, B., Forbus, K., Manduca, C.A., Ormand, C.J., Shipley, T.F., Tikoff, B.: Drawing on Experience: How Domain Knowledge Is Reflected in Sketches of Scientific Structures and Processes. Research in Science Education. 44, 859–883 (2014).
10. Burrows, S., Gurevych, I., Stein, B.: The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education. 25, 60–117 (2014).
11. Graesser, A.: Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor. Interactive Learning Environments. 8, 1–33 (2000).
12. Sukkarieh, J., Blackmore, J.: C-rater: Automatic content scoring for short constructed responses. Proceedings of the 22nd International FLAIRS Conference. 290–295 (2009).
13. Dzikovska, M., Nielsen, R., Brew, C.: 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. pp. 200–210. , Montreal, Canada (2012).
14. Leeman-munk, S., Smith, A., Mott, B., Wiebe, E., Lester, J.: Two Modes Are Better Than One : A Multimodal Assessment Framework Integrating Student Writing and Drawing. 17th International Conference on Artificial Intelligence in Education. 205–215 (2015).
15. Gobert, J.D., Pedro, M. a S. a O., Baker, R.S.J.D., Toto, E., Montalvo, O.: Leveraging Educational Data Mining for Real-time Performance Assessment of Scientific Inquiry Skills within Microworlds. Journal of Educational Data Mining. 4, 111–143 (2012).
16. Rupp, A., Levy, R., Dicerbo, K.E., Sweet, S.J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K.L., Mislevy, R.J., Behrens, J.: Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment. Journal of Educational Data Mining. 4, 49–110 (2012).
17. Min, W., Frankosky, M.H., Mott, B.W., Rowe, J.P., Wiebe, E., Boyer, K.E., Lester, J.C.: DeepStealth: Leveraging deep learning models for stealth assessment in game-based learning environments. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence in Education. pp. 277–286. , Madrid, Spain (2015).
18. Mislevy, R.J., Almond, R.G., Lukas, J.F.: A brief introduction to evidence-centered design. ETS Research Report Series. 16, (2003).
19. Lecun, Y., Bengio, Y., Hinton, G.: Deep Learning. Nature. 521, 436–444 (2015).