Grading OSPE Questions with Decision Learning Trees: A First Step Towards an Intelligent Tutoring System for Anatomical Education

Jason Bernard[1,2], **Bruce Wainman**[3,5], **O'Lencia Walker**[3], **Courney Pitt**[3], **Ilana Bayer**[3,5], **Josh Mitchell**[3], **Alex Bak**[4], **Anthony Saraco**[3] **Ranil Sonnadara**[1,2]

[1] Department of Surgery, McMaster University, Hamilton, Ontario, Canada
[2] Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[3] Education Program in Anatomy, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada
[4] Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[5] Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada
bernac12@mcmaster.ca, wainmanb@mcmaster.ca, ranil@mcmaster.ca

Abstract

Intelligent tutoring systems (ITSs) have been used for decades as a means for improving the quality of education for learners primarily by providing guidance to students based on a student model, e.g., predicting their knowledge level on a subject. There have been few attempts to incorporate ITSs into anatomical education. Objective structured practical examinations (OSPEs) are an important, albeit challenging, means of evaluation in anatomical education. This research aims to create an ITS for anatomical OSPEs, and as a crucial first step looks to create a machine learning-based approach for grading OSPEs. To that end, decision tree learning was evaluated with, and without, spellchecking to produce a grading tool using the answer key developed by instructional assistants. Using answers from 428 learners, the tool obtained an average accuracy of 96.8% ($SD = 3.4\%$) across 60 questions.

Introduction

Intelligent tutoring systems (ITSs) in educational technology have been researched since at least the 1960s (Regian and Shute 1966). An ITS works by interacting with the learner and utilizing student modelling techniques to provide a customized experience based on their cognitive characteristics, such as affect, knowledge level, and interests (Regian and Shute 1966; Bakhshinategh et al. 2018; Joshi et al. 2019; Xu et al. 2019; Mousavinasab et al. 2021). This is often done using adaptive learning material; however, other pedagogical techniques can be used, e.g., gamification, prompting the learner to reflect on their answer, or proving an immediate review (Regian and Shute 1966; Bakhshinategh et al. 2018; Joshi et al. 2019; Xu et al. 2019; Mousavinasab et al. 2021). Overall, the experience with ITSs suggest that they have a positive effect on learning outcomes (Joshi et al. 2019; Xu et al. 2019).

Medical education has not been much of a focus for educational technology in general, and ITSs in particular. This research was undertaken to develop an ITS for anatomical sciences education. In anatomical education, the objective structured practical examination (OSPE) is considered an important part of the curriculum (Chan et al. 2019); however, it is an exam with which many learners struggle. OSPE questions are in the form of an image (or sample) with a pin indicating the anatomical structure to be considered by the student. The student is typically asked to either identify the structure or its function in the form of a short sentence (or sentence fragment). Therefore, an algorithm is needed that can grade short answer OSPE-style questions. While there has been much work on grading short answer questions (Leacock and Chodorow 2003; Shermis et al. 2015; Dumais 2004), these approaches use natural language processing (NLP) techniques that are intended to work with short paragraphs and to be more general to many topics (mainly in a K-12 context). Student answers to OSPE questions tend to be short sentence fragments that lack proper grammatical structure. A preliminary examination using NLP on the OSPE answers suggested that there was insufficient information for the algorithm to derive much meaning. Hence, due to the differences in the answer structure and the early NLP assessment, existing approaches were not evaluated.It was observed that the student answers, while short, generally used the unique, technical words of the anatomical sciences, although not often the same words used in the faculty derived answer key. Therefore, it was hypothesized that due to the technical nature of the anatomical sciences, particular and unique words should appear in correct answers, even if they are not the words expected by faculty. Furthermore, decision tree learning can use a series of derived simple true/false rules to determine the combinations of such words that infer a correct or incorrect answer. While the tool created is likely not generally useful for short answer questions, an evaluation showed it is able to grade OSPE questions using the students' lexicon with 96.8% accuracy.

Methodology

This section describes the approach used in this research to evaluate decision trees (DTs) to grade OSPE questions. It begins with an overview of the decision tree learning (DTL) algorithm. This is followed by a description of how decision trees are used to evaluate OSPE questions. Afterwards, the data and data gathering approach is discussed and then the metrics used to evaluate the OSPE grading tool.

Decision Tree Learning

Previously, it was shown that DTL could be useful for parsing grammatical structures and using the resulting tree to aide in grading short answer questions (Leacock and Chodorow 2003). While their approach differs as they use the structure as input to other NLP approaches, it suggests that the decision tree structure can be useful for this kind of problem. It is certainly possible that other algorithms may be effective for identifying correct answers to OSPE questions (e.g., potentially an unevaluated NLP algorithm or clustering algorithms). However, DTs seem particularly well suited to use in an ITS. Firstly, unlike most NLP algorithms that use neural networks, which are black box algorithms, DTs provide a transparent reasoning that can be expressed to students along with a confidence level of correctness. Secondly, other algorithms may struggle to define the relationship in an efficient way, e.g., clustering algorithms would likely create overlapping word spaces that would be difficult to evaluate. Hence, for this research study, DTs were used to produce a set of rules that describe a relationship between the words in an OSPE answer and correctness. The following description of DTs is summarized from Quinlan (1996) unless otherwise noted.

The aim of a DT is to produce a classification of a sample based on a sequence of true/false rules relating to a feature in the data. For example, to predict whether it will rain, a tree may consist of the rules “is it cloudy?”, and if so then “is the humidity above 60%?”. If the answer to both questions is “true” (or yes), then predict it will rain, and if the answer to either question is “false” (or no), then predict it will not rain. Each rule is represented structurally as a node consisting of: 1. the Boolean rule, 2. the certainty for each possible classification (described below), and 3. an optional connection to two other nodes (called child nodes). One node is designated for when the rule evaluates to “true” for a sample, and the other for “false”.

The collection of nodes is arranged in an (upside down) tree-like structure (a simplified sample from this research is shown in Figure 1), with a root node at the top and its children below it, and their children on the next level down, and so on. The bottommost nodes have no children and are referred to as leaf nodes. Each child is a subtree of its parent and referred to as S_T (subtree “true”) and S_F (subtree “false”). By convention, the S_T is to the left, and S_F on the right. Any movement to the right does not automatically mean an answer is incorrect as there may be many different combinations of words that are correct. Hence, for this tree, it is possible that an answer lacking the word “muscles” is still correct, and would be identified by the DT as such. For those cases, the word at the second level would likely be an alternative word that appears in correct answers. The effect of the DT is to create a serial of rules. If “subvavular apparatus” is a correct answer, then effectively the result of the tree traversal is to ask:

Does the answer not contain “muscles” and contains “subvalvular” and contains “apparatus”, and if so, then the answer is classified as correct.

The process of building a tree from data is called DTL. DTL is an iterative process that examines a data set to find the Boolean rule that has the greatest information gain measured by reducing entropy. In other words, all possible rules are considered, and the rule that allows for the most clarity to the prediction is selected. For example, it is difficult for it to rain if there are no clouds in the sky, so this is a reasonable first question to determine if it will rain. If there are no clouds, then it will clearly not rain.

The general case can then be described as follows. Let D_0 be the initial dataset. Starting with the first iteration, every possible rule that can be applied to D_0 is considered. The rule that is most effective at clarifying the prediction (called information gain or IG) is selected and D_0 is split into two datasets D_{T1} and D_{F1}. D_{T1} contains all samples for which the selected is true and D_{F1} all those for which the rule evaluates to false. The process iterates using D_{T1}, and then D_{F1}, which will each produce two additional datasets, and so on until no rule can split the dataset.

Mathematically, this is done by computing entropy (E) as shown in Equation 1 where T and F are the count of the samples in the dataset that would evaluate to true and false respectively if the rule were selected. For some iteration, let there be n possible rules. Information gain (IG) can then be computed by subtracting the entropy of a candidate rule n from the entropy using the current dataset (E_{curr}), e.g., in the first iteration D_0, in the second iteration, E_{curr} is computed using D_{T1}, and then separately using D_{F1}, and so on. This is shown in Equation 2. Finally, the rule with the best IG value is selected. The probability of each possible classification (correct and incorrect for this research) is computed and stored by count of the number of samples with each label divided by the total. The classification associated to the node is the one with the greatest probability. The certainty of the classification being right is the probability of that classification. For example, if 56% of all samples in a data set are labelled as “correct”, then the “correct” classification will be associated to the node with 56% certainty.

$$\begin{aligned} E(T,F) = &- T/(T+F) \cdot logT/(T+F) \\ &- F/(T+F) \cdot logF/(T+F) \end{aligned} \tag{1}$$

$$IG_(T_n, F_n) = E_{curr} - E(T_n, F_n) \tag{2}$$

Once the DT is built using this process, a new sample is classified by traversing the tree starting from the root node. At each node, the Boolean rule is applied to the sample and if it evaluates to “true”, then the true connection is followed; otherwise, the false connection is taken. If no traversal is possible (which can happen if not all samples have the same features), or if the node is a leaf node, then the process terminates, and the associated classification and certainty are returned.

Grading OSPE with a Decision Tree

The previous subsection described how a DT can mechanically grade a question based on some set of features. This subsection describes the features utilized and principles by which the tool functions. To begin, the feature set is simply

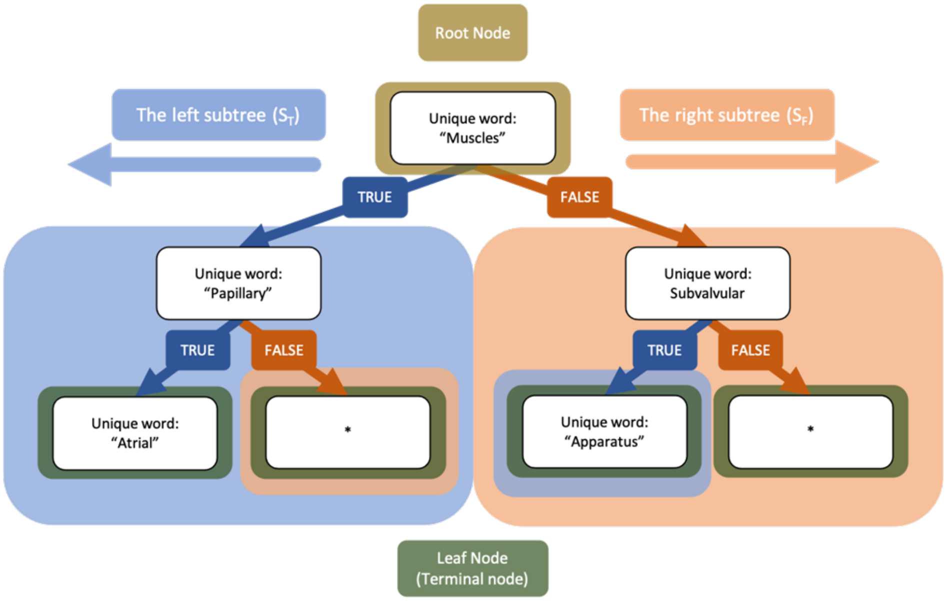

Figure 1: A decision tree (DT) is a series of nodes containing unique words that are connected by Boolean (True/False) decisions. The nodes are described as either a ”root node” (a node that has nodes stemming from it) or a “leaf/terminal node” (a node that does not have nodes stemming from another node). All nodes that result from the Boolean decision returning “True” are included in the left subtree (surrounded by blue) while all nodes that result from the Boolean decision returning “false” are included in the right subtree (surrounded by pink). The asterisks (*) are a “wild card”, representing any word that is not one of the unique words. In this case a correct OSPE answer was atrial papillary muscle(s) or subvavular apparatus and all other answers would have been found to be false by the DT.

all of the unique words that exists across all student answers for a question. It was hypothesized that students should have a particular shared lexicon of words for describing the correct answer, even if it is not exactly the same as the textbook answer. Such a lexicon can be used to train a decision tree as there will be a set of unique words that belong only to correct answers, and a set that belong to only incorrect answers. So while some will belong to both, if a student answer contains the words associated with a correct answer as determined by the decision tree, then this is positive evidence that the answer is correct, and vice versa.

Some preprocessing was done to the data set prior to training the algorithm. First, all blank answers were removed since they are trivially incorrect and uninteresting. Second, all answers were spellchecked using the Jazzy spellchecker v0.5.2 (Idzelis 2005). The dictionary included with the Jazzy library was used; however, since it does not contain many medical terms all words in the master answer key were included. In all cases where a misspelling was identified, the top word in the correction list was taken. This is one area where perhaps the algorithm could be improved by considering different possible corrections. Finally, the following common English words found in the student answers were removed as they do not provide any indication of correctness: “a”, “an”, “and”, “are”, “as”, “at”, “be”, “but”, “by”, “did”, “for”, “had”, “has”, “have”, “I”, “in”, “is”, “it”, “of”, “on”, “or”, “so”, “than”, “that”, “the”, “then”, “they”, “this”, “to”, “was”, “with”.

Data

The data for this research consisted of the answers from a 60 question OSPE in McMaster University’s Health Sciences Human Anatomy and Physiology (HTHSCI 2F03/2FF3/2L03/2LL3/1D06) undergraduate course. The exam consisted of 20 two-dimensional images from the Stereoscopic Atlas of Human Anatomy (Massachusetts General Hospital 2017). Digital markings (pins, asterisks, arrows, etc.) were added to the images to indicate the anatomical structure for students to consider. Each image had three associated questions. The exam was conducted online using “Desire to Learn”, McMaster University’s learning management system. The exam was completed by 428 students, who had 50 minutes to complete the exam. Virtual proctoring software Respondus® was used or virtual proctoring with a TA if Respondus® would not work on a student’s computer. The questions and the marking master for the OSPE were produced by the five senior faculty teaching the course. All student answers were graded by two teaching assistants (TAs), who then reviewed any differences and came to a consensus, referred to as the initial grade. All of the grades were then reviewed by the two instructional assistants and a final grade was produced. Overall, approximately 5% of initial grades were altered by the instructional assistants.

As DTL was used for this research, a training and test data set were required. The training set was used to produce the tree by examining the student answers as described in

Subsection 2.1. With the trained DT, a grade was produced for each student in the test set for each question. A 10-fold, cross-validation approach was used. Hence for each fold, 42 students were randomly selected as the test set, and the remainder was used as the training set. Students were selected such that no student appeared in more than one test set. The test set is treated as if their answer previously provided and evaluated, and is therefore referred to as the "student key".

Metrics

The performance of the OSPE grading tool was measured by comparing grade produced by the DT to the actual grade. Specifically, for each of the 42 students in each fold, and for each of the 60 questions, the grade produced by DT is compared to the actual grade. The fold accuracy for each question is then the average number of matches divided by 42. The accuracy for the question is the average across the 10 folds. While the average accuracy is useful to produce a general sense of the effectiveness of the tool for grading OSPE questions, the final grade, produced grade, and certainty were recorded to allow for a deeper analysis into the logic of the algorithm, especially when the DT does not agree with the final grade.

Results

The computed accuracy using the student answers for the OSPE grading tool is shown in Figure 2, along with average grade.

The key result is the accuracy when determining a final grade as this has the greatest effect on the students and is essential for building an ITS. It can be seen from the results that the accuracy when using the "Student Key" has an average accuracy of 96.8% ($SD = 3.4\%$), and lowest accuracy of 84.8% (Q27). These results suggest that students develop their own collective lexicon for answering anatomical questions, but are still considered correct. Pedagogically, while unexplored, it is possible that adapting learning material to the students' lexicon may be valuable in promoting a better learning outcome.

The result with spellchecking was not any better. In 11 questions, the average accuracy was slightly lower, while in 2 questions it was higher ($< 0.5\%$ higher or lower). For the most part, the DT algorithm seemed to learn the misspellings as frequently the mistakes were identical. Where it would make a difference, the spellchecker struggled with medical terms despite making some effort to adjust it to properly correct them. For this reason, the experiment will be done again with a spellchecking algorithm and dictionary designed specifically for medical use. For these results, no significant conclusion can be reached.

The relationship between accuracy and the average grade was determined using the Pearson correlation coefficients $r = 0.153$ and $p = 0.244$. While the R value suggests that there is no correlation, p value is greater than 0.05 so it is possible that the results are occurring by chance. Therefore, an algorithmic and practical evaluation of the reasoning was taken.

Algorithmically, if the AI has a bias then it should favour either guessing correct or incorrect. If the AI has a correct bias, then questions with a grade less than 50% should have consistently low accuracy. If the AI has an incorrect bias, then questions with a grade of 50% or higher should have a consistently low accuracy. It is evident from Figure 2 that this is not the case. Questions with grades between 40% to 60% have accuracy values throughout the distribution.

From a practical perspective, an anatomical expert was recruited to duplicate the first step taken by the tool. The expert was asked to examine the student answers for each question and pick from the lexicon which word would be most likely to indicate that the student had answered the question correctly. The expert made the selection without knowing what had been selected by the AI. Ideally, it would be better to have the expert completely duplicate the process of the tree; however, this would be quite time consuming for them, so this was taken as a rough approximation of logical agreement. The expert picked the same word as the tool in 45 of the 60 questions. For 4 of the 15 questions, the expert picked an acronym that was not frequently used by the students who preferred to write the answer out in full (which itself is potentially pedagogically interesting), indicating the expert was relying somewhat on their own expert knowledge beyond just the student answers. In the other 11 cases, the word selected by the expert was the OSPE grader's second choice. In particular it is notable, that when the answers were longest (10 or more words on average) the expert and the AI either agreed or it was the AI's second choice. This indicates that the AI is doing reasonably well at finding the important words by mirroring the human choice. Overall, in combination with the other observations discussed, this suggests that the reasoning used by the AI is valid, and that it is not simply guessing.

Limitations

While the result are promising, there are some noteworthy limitations to the findings. The long term goal of this research is to develop an ITS for anatomical education; however, this research has assumed that questions have been answered by a cohort of students from which the DT can be built. This is not necessarily ideal as it would be more practical to add a question to the hypothetical ITS and have the AI simply work. This work would suggest that over time the ITS would get better at grading the questions, which is welcome, but it does not address what happens early on. Of course, it is possible to simply take exam questions after they have been used and add them to the ITS; however, this requires the instructional assistants to constantly come up with a new pool of questions. Therefore, this tool needs to be evaluated without training using the student answers. To address these limitations, two steps have recently been taken. A group of third and fourth year university students of the anatomical sciences have been recruited to produce many OSPE questions. This will provide both the necessary questions for the ITS, and an initial seed of student answers to train the DT. Faculty will also add responses to the answer key. Additionally, an evaluation is being performed using only the faculty-derived answer key.

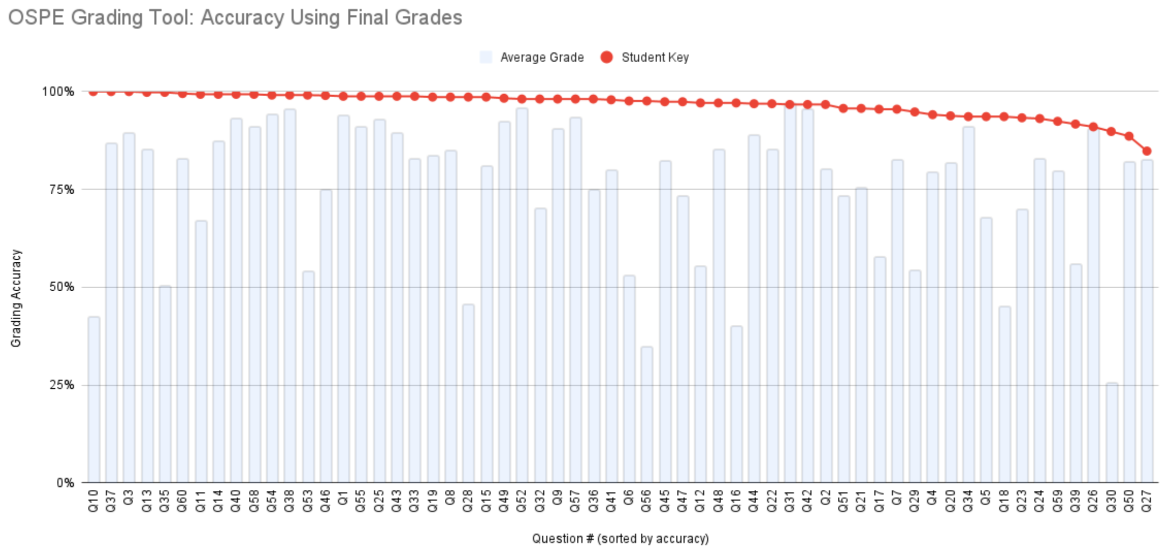

Figure 2: This figure shows the accuracy when using the student key (red line) ordered from highest to lowest. Additionally, the percentage of students who answered each question correctly is shown as determined by the faculty-generated mark master (background bar graph).

Conclusions

This paper has presented an early look at a machine learning-based tool for grading objective structured practical examinations (OSPEs), which are frequently used and viewed as an important aspect of anatomical sciences education (Chan et al. 2019). As OSPE questions are short answers, consisting of a few words or a short sentence, it is more difficult to grade than multiple choice (for example), where a student answer is definitively correct or incorrect. It was hypothesized that a decision tree could learn the lexicon used by learners to answer questions, and distinguish from that lexicon the words associated with correct and incorrect answers.

Using the answers obtained from 428 anatomical sciences students on a 60 question OSPE, the tool was trained using a 10-fold cross validation method. Overall, the algorithm obtained a 96.8% accuracy ($SD = 3.4\%$) for correctly grading the student answers. Based on a multifaceted analysis of the results, it was determined that the tool was not simply a guess. Firstly, the algorithm shows no bias towards guessing "correct" or "incorrect" based on an examination of questions with grades ranging from 40% to 60%. Secondly, an anatomical expert was recruited to examine the algorithms selected root words and the AI choices were found to be reasonable and matching 45 out of 60 times, and being the second choice for 11 questions. For the remaining four questions, the expert made a choice not possible for the AI by using an acronym not used by the students. Overall, the evidence suggests that the OSPE grading tool is using reasoning and not guessing.

While the average result was promising, three questions (Q27, 30 and 33) were notably lower than the mean with accuracies about 85%. The underlying causes for the errors were examined and some anatomical terminology was not recognized as words. For example, "C5", "C6", "CN11", and "CNXI" were not considered words, let alone unique words, by the DT and therefore, they were included in the solution space. For Q30, there were many different variations of words that appeared in correct answers, e.g., movement, motion, forward, extension, hyperextension; however, these words also appeared in incorrect answers as well. The potential solution is to blend the student key with the faculty derived answer key and have certain words marked as critical. Other solutions would be to use a more complex natural language processing solution that may be required to understand the words in context or have the tool learn the weighted importance of the different words.

The future of this research, beyond addressing the issues around accuracy, is to expand the grading tool into an ITS. The ITS can then be evaluated on a student cohort to see if it improves the learning outcomes. As part of building an ITS, an investigation will be conducted on learning outcomes when using the students' lexicon versus textbook answers. Recently, work has begun by developing an online OSPE practice tool for students using the AI-based grader.

References

Bakhshinategh, B.; Zaiane, O. R.; ElAtia, S.; and Ipperciel, D. 2018. Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies* 23(1): 537–553.

Chan, A. Y.-C. C.; Custers, E. J.; van Leeuwen, M. S.; Bleys, R. L.; and ten Cate, O. 2019. Does an Additional Online Anatomy Course Improve Performance of Medical Students on Gross Anatomy Examinations? *Medical Science Educator* 29(3): 697–707.

Dumais, S. T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* 38(1): 188–230.

Idzelis, M. 2005. Jazzy: The Java open source spell checker.

Joshi, A.; Allessio, D.; Magee, J.; Whitehill, J.; Arroyo, I.; Woolf, B.; Sclaroff, S.; and Betke, M. 2019. Affect-driven learning outcomes prediction in intelligent tutoring systems. In *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–5. IEEE.

Leacock, C.; and Chodorow, M. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities* 37(4): 389–405.

Massachusetts General Hospital. 2017. Bassett Collection: Stereoscopic Atlas of Human Anatomy.

Mousavinasab, E.; Zarifsanaiey, N.; R. Niakan Kalhori, S.; Rakhshan, M.; Keikha, L.; and Ghazi Saeedi, M. 2021. Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments* 29(1): 142–163.

Quinlan, J. R. 1996. Learning decision tree classifiers. *ACM Computing Surveys* 28(1): 71–72.

Regian, J.; and Shute, V. 1966. Arificial intelligence in training: The evolution of intelligent tutoring systems. In *Proceedings of the Conference on Technology and Training In Education*.

Shermis, M. D.; Burstein, J.; Brew, C.; Higgins, D.; and Zechner, K. 2015. Recent Innovations in Machine Scoring of Student and Test Taker Written and Spoken Responses. In *Handbook of Test Development*, 351–370. Routledge.

Xu, Z.; Wijekumar, K.; Ramirez, G.; Hu, X.; and Irey, R. 2019. The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology* 50(6): 3119–3137.