An AI toolkit to support teacher reflection

Tanner M. Phillips, Corresponding Author

Department of Instructional Systems Technology, Indiana University, Bloomington, Indiana, USA. Orcid ID: 0000-0002-1769-0584

tanphill@iu.edu

Asmalina Saleh

Center for Research on Learning and Technology, Indiana University, Bloomington, Indiana, USA

Gamze Ozogul

Department of Instructional Systems Technology, Indiana University, Bloomington, Indiana, USA. Orcid ID: 000-0002-8044-9021

Abstract

Encouraging teachers to reflect on their instructional practices and course design has been shown to be an effective means of improving instruction and student learning. However, the process of encouraging reflection is difficult; reflection requires quality data, thoughtful analysis, and contextualized interpretation. Because of this, research on and the practice of reflection is often limited to pre-service training or short professional development cycles. This study explores how natural language processing, deep-learning methods can be used to support continuous teacher reflection by facilitating data collection and analysis in any instructional setting that includes ample linguistic and assessment material. Data was collected from an existing introductory undergraduate biology course. A Bidirectional Long-Short Term Memory network was trained to predict assessment item difficulty and tasked with assigning difficulty to recorded lectures. Comparison with the instructor's perceptions of lecture material difficulty suggested the model was highly reliable at predicting difficult lecture material. We discuss how this model could be expanded into an AI toolkit meant to aid in teacher reflection on their practices and curriculum.

Keywords: Natural Language Processing, Deep Learning, Reflective Practices, Blended Learning, Higher Education

An AI toolkit to support teacher design and reflection

In recent years, there has been an interest in reframing teaching as an act of design and continuous reflection, especially with the advent of technologies that can be used to support pedagogy (Warr & Mishra, 2021). Research on the use of artificial intelligence tools has often focused on teacher orchestration or understanding how AI tools such as dashboards that can support teachers in coordinating complex classroom interactions across multiple social learning planes (Dillenbourg et al., 2018). Central to the effective use of AI tools in the classroom is reflective practice, which requires three dimensions: (1) the collection of data about what is happening in the classroom, (2) analysis and evaluation of this data, and (3) an exploration of how this data relates to and can inform one's practice and beliefs as a teacher (Liaqat, 2017). Despite substantial empirical evidence that reflective practices substantially improve teaching quality and, by extension, student outcomes, support for reflective practice for in-service teachers is much less substantial (Peercy et al., 2019). This is further exacerbated when one considers the variations in how teachers can use these AI-tools in their design and classroom instruction (van Leeuwen et al., 2021). This presents an opportunity to expand how AI-tools can be used to understand the nature of information that can be provided to teachers to support reflective practice.

This paper presents the creation and validation of an AI toolkit that leverages natural language methods to support teacher reflection. In designing the AI toolkit, we considered (a) how the method leverages existing data streams and eases the burden of data collection, (b) how the method allows for automated analysis, (c) how the output of the analysis could aid in teacher reflection. We designed a toolkit that can be easily adapted to measure the relationship between student learning and curriculum content in a variety of in person and blended learning contexts,

instead of tying the AI to a specific curriculum or instructional tool. In the first section of the paper, we first consider our objective, which is supporting teacher reflection by focusing on the data collection and analysis dimensions. We highlight how NLP methods can be used to augment data collection for teachers and offer interpretable data that can support teacher analysis of student performances. We then introduce and test the models that are embedded in the toolkit before discussing how student data could be summarized for teachers in a dashboard meant to prompt the types of reflections necessary for continuous improvement of teacher practices. We address the question: how can an AI toolkit that leverages NLP methods be used to collect, analyze, and present classroom data to aid in teacher reflection?

Teacher reflection

Teacher reflection is the process of teachers critically evaluating their instructional experiences to improve their practices and students' learning. Teacher reflection is an important aspect of many teacher-education programs, whereby pre-service teachers are mentored and guided in reflective practices by both teacher educators and in-service teachers (Hatton & Smith, 1995; Ozogul et al., 2018). The goal of reflective teaching during pre-service instruction is to improve the educational experience of pre-service teachers, and to help develop long-term critical reflection skills. In addition, pre-service teachers should apply skills into their professional practice, becoming lifelong learners who continuously evaluate their teaching and adapt their practices to meet changes in curriculum and the varied needs of their current students (Liaqat, 2017). Ideally, reflective practices are a continuous process by which teachers are always considering their and their students' experience in the classroom and attempting to modify their practices accordingly. Over the last several decades, research on teacher reflection has grown

into a separate and substantial area of research (Hatton & Smith, 1995; Loughran, 2002; Mor et al., 2015). Additionally, research on teacher reflection has also expanded beyond pre-service education; there is now substantial research on the reflective practices of in-service teachers (Khoshsima & Nosratinia, 2019).

Although the reflective practices of pre-services teachers are easily structured by teacher educators and teacher mentors, the reflective practices of in-service teachers are much more difficult to support. Professional development interventions carried out by both researchers and professional organizations often aim to develop the metacognitive skills necessary for teacher reflection with the hope that these skills will persist long-term. When effective teacher reflection programs are implemented for in-service teachers, they most frequently involve teachers working together with other teacher, administrators, or researchers who help scaffold this process and encourage them to continuously challenge their preconceptions (Escamilla & Meier, 2018; Liaqat, 2017).

Additionally, reflection is dependent on the teacher's ability to capture and analyze high-quality data about classroom practices, student experiences, and outcomes. Often, the data collected for teacher reflection is subjective; for example, teacher journal entries are often the primary source of data for reflection (Cornford, 2002). Although the process of journaling about teaching experience has value, the process of reflection often needs an external evaluator who can provide additional assessments about teaching practices (Ozogul et al., 2018). Assessment data is another frequent, the less common, source of data for reflection (Leitner et al., 2017). Although assessments may offer teachers another source of data, quality formative and summative assessments are often difficult to create and the interpretation of these assessments often requires a significant time investment (National Research Council, 2001; Messick, 1995;

Mislevy, 2005). A growing interest in the integration of learning analytics into teacher reflection aims to solve this issue through the collection and analysis of robust, multimodal data (Persico & Pozzi, 2015). However, these professional development activities are still ad-hoc and resource intensive. Thus, there needs to be more support for teacher reflection among in-service teachers, particularly in terms of sustainable, continuous collection, analysis, and interpretation of rich, usable data.

Research on intelligent tutoring systems (ITS) offers some insights into how we might support teacher reflection. Some popular ITS, such as ASSISTment, are popular in part because of the extensive and easily interpretable reports they give to teachers (M. Feng & Heffernan, 2005). Reports generated by ITS are primarily framed as a tool for teacher to individualize instruction to different students (M. Feng & Heffernan, 2005). However, these reports could easily be modified to focus more on promoting teacher reflection. Moreover, the design principles developed for ITS, especially the focus on simplicity and readability offer a starting point for the design of reports and dashboards for teacher reflection (Baker, 2016). Although ITS offer us a window into the type of analyses and reports that may be useful for teacher reflection, they do not solve the problem of data collection. Most teachers do not use ITS, and, while they are growing in popularity, some forms of instruction cannot be performed with ITS.

What is needed is a more general form of data collection that can supplement existing teacher reflection practices and can be reconciled with a wide variety of teaching methods and contexts. One form of data that is shared between most instructional contexts is language. The main issue with language is that it is complex and idiosyncratic, and for this reason does not easily lend itself to analysis. However, recent and continuing advancements in natural language processing (NLP) may allow language as a more feasible source of data (Chau et al., 2021;

Mikolov et al., 2013). In ITS, language data, such as the content of curriculum material and assessment questions, is coded manually, most frequently by topic. If NLP methods could replicate this topical coding, it would be possible to apply the same types of models and reports in ITSs to other contexts. To accomplish this, two distinct tasks must be accomplished: capturing linguistic data and extracting topics for linguistic data.

Much of the data from classroom teaching often exists in a linguistic, written form: textbooks, exams and homework assignment, and other curricular material are readily available in written form, often digitally. However, much of teachers practices during the act of teaching are not written down ahead of time. Both lectures and just-in-time soft scaffolding (Saye & Brush, 2002) are most frequently verbal. These data also may most closely reflect the practices and beliefs of teachers that are important to teacher reflection. Until recently, extracting any sort of information about verbal communication in the classroom required substantial manual work. However, free, or cheap automatic transcription services and APIs are now readily available. These tools continue to improve in accuracy, and, while imperfect, research suggests they are now "good enough" to be used as part of complex analyses (Bokhove & Downey, 2018).

Once verbal information is captured in written form, it must also be analyzed to extract topics. Prior popular methods for extracting topical information from text, such as latent Dirichlet analysis, relied on a "bag of words" assumption, whereby all words in a document are treated as equally related to one another (Wallach, 2006). This meant that some documents, for example, an hour-long transcription of a teacher talking in a class, were not reconcilable with these types of analysis. More recently, the advent of more sophisticated models for representing topical information such as continuous bag-of-words (CBOW) and skip-gram (Mikolov et al., 2013) have allowed for efficient analysis of documents of arbitrary size. Both the skip-gram and

CBOW algorithm look at all the words within a small, predefined window (generally between 5-20 words) and attempt to use some combination of the words in the window to predict the other words in the window. After performing this task on many samples of words in the corpus, the model outputs a *word embedding*, a representation of all the words in the corpus in a high-dimensional space. Words closer to each other in the high dimensional space are more closely related. These word embeddings greatly simplify the problem space of linguistic analysis. Where a corpus may contain thousands or tens of thousands of unique words, word embeddings generally contain at most several hundred dimensions. Word embeddings also easily interface with different neural networks such as long short-term memory (LSTM) and other recurrent neural networks (e.g. Geden et al., 2020).

This analytical method: using word embeddings as the input for neural networks, has been shown to be an effective way to extract topics from educational material (Chau et al., 2021). Word embeddings do differ somewhat in format from the types of topical information usually created by manual coding in ITSs: manually codes are generally discrete, where word embeddings are continuous. Additionally, word embeddings are not immediately interpretable. To use word embeddings as a primary topical source for modeling student learning, some changes would need to be made to existing methods. Additionally, more work would need to be put in to making sure output was interpretable. A key concern in our design of the AI toolkit is to ensure the interpretability of the data so that we can support teacher reflection. To that end, the goal of the paper is to address these challenges by exploring methods for an AI-toolkit that (a) supports contextualized interpretation of student performance, and (b) can be used to support teacher analysis and evaluation of the data.

Methods

Context

Data for this study were collected from an introductory biology course at a large private university in the western United States. The student body of the university is predominantly white, and the inter-quartile range of ACT scores for the incoming freshmen class in 2019 was 26-31 (SAT equivalent 1250-1400). This introductory biology course had 27 students and used a flipped inquiry-based learning curriculum (Jensen et al., 2018), where students spent time before each class session learning content, and bi-weekly class time was spent on active learning activities. Before each class, students viewed several brief videos assigned by the instructor which delivered the content. The average video length was approximately 10 minutes. During class, students engaged in guided scientific inquiry activities with the help of the instructors and teaching assistants. For example, in the week on human evolution students watched three videos that introduced different concepts related to human evolution. In class, they made observations about the features of different pre- Homo Sapien skulls and attempted to create a phylogenetic tree of the hominin group of species. The class met twice a week for a total of 2.5 hours.

The course assessment material consisted of four different types: (a) homework (n=871), (b) exams (n=155), (c) a practice exam (n=131), and (d) "other" assessment material (n=71) as defined below. Homework consisted of a combination of open-ended free-response questions (n=422) and multiple-choice questions (n=449) and was administered through a learning management system. Homework assignments were open-note and open-book. Students took a total of seven exams throughout the semester, six formative mid-term exams, and one final, comprehensive exam at the end of the semester. Before the final, students completed an open-book practice exam. Several other assessments were present in the course and are combined in a broad "other" category in this study. This included a pre and post-semester scientific reasoning

test (Jensen & Lawson, 2011), as well as video quizzes, which asked students to self-report whether they had completed the assigned video assignment. All questions in the final exam, practice exams, and "other" assessment items were multiple-choice items. The course covered a broad range of topics including (but not limited to) the nature of science & data literacy, evolution and natural selection, ecology, genetics, reproduction, physiology, cell structure, and DNA transcription/replication.

This course was chosen in part because the instructor was a discipline-based biology education researcher, and the course curriculum and assessment instruments had been developed and studied for nearly a decade (Jensen et al., 2013; Jensen et al., 2014, 2018; Jensen & Lawson, 2011; Kummer et al., 2016). From an AI toolkit development perspective, this allowed us to explore the extent to which instructors can make sense of the data that is derived from their own classrooms, a key facet of teacher reflection. This also meant that the course material was of high quality and provided a uniquely expert instructor with a deep understanding of the material. The unique expertise presented the opportunity to use the instructor's insights as a means of validating the accuracy of the model due to our increased trust in her insights. Although the goal of the AI toolkit presented in this study is to generalize beyond this highly structured context, having high trust in the reliability and quality of the instruction and assessment tools allowed us to use the expert instructor as a means of validating and measuring the accuracy of our AI toolkit.

Data Collection

Data were obtained from three different sources: (1) automatically generated transcripts of recorded video lectures, (2) class assessment texts, and (3) student scores on assessments. Because the class followed a flipped classroom format, all lectures were pre-recorded and uploaded to YouTube. A total of 71 different videos were analyzed, totalling approximately 40,000 words.

YouTube's automatic transcription was then downloaded for analysis. We manually reviewed 3 randomly selected sections of text totalling 1255 words. The error rate in transcription was 1.5%. This is more accurate than previously reported error rates for automated transcriptions. (Lee & Cha, 2020), and is also free. This may be due to the scripted format of the videos, as well as the accent of the instructor who recorded the videos. This low of an error rate may not be replicable under all circumstances. Because one of the goals of this study was to develop a tool that would be widely affordable, we did not pay for transcription services, thereby situating the study closer to what would be feasible if the AI toolkit was deployed at scale. All text from class assessment material (homework, exams, practice exam, and "other") was also collected and used for modeling. There were a total of 1228 unique graded items in the class, the text of which was approximately 20,000 words. Finally, student (N=27) scores on all graded material were aggregated.

Data Processing

All analysis was completed on a computer with 8GB of RAM and no dedicated graphical processing unit. This limitation was imposed to restrict model architecture and training to those that would be accessible at scale. Because of the limited size of the data set, it was necessary to augment the training data. Although there are many different forms of text data augmentation, there is limited research directly comparing methods (Feng et al., 2021). Many complex methods, such as graph-structured augmentation have been recently proposed. However, the properties of these methods are still not well understood and are actively being researched (Shorten et al., 2021). One of the simplest and most well understood form of text data augmentation is *rule-based augmentation,* where words are deleted, added, or swapped at random in copies of the training data (Feng et al., 2021). Of these three, adding and swapping have been criticized because they can imply spurious relationships between words. For example, "I went for a run" could become "I

went for purple run" with both addition and swapping augmentation. Deletion does not create such spurious relationships. In the previous example, with deletion the phrase "I went for a run" becomes "I went for run." This method does, however, creates grammatical inaccuracies. One remedy for this issue is to insert a reserved token in place of the deleted word (Xie et al., 2017). Here, "I went for a run" becomes "I went for [MaskedToken] run." This method, known as *blank-noising,* explicitly informs any neural network analysing the data that a token has been removed and allows it to adjust accordingly.

Five-fold cross validation was used to train the model. Blank-noising was used to create 10 variants of the text of each assessment item used for training (n=983), with each word in an observation blanked with probability $\gamma = .1$. The blank-noising algorithm was not applied to data withheld for validation (n=246) and testing to ensure the model was validated only on real data. The output of the model is a prediction of the difficulty of the assessment item. The difficulty of each assessment item was measured as the proportion of students who correctly answered each item (often referred to as the difficulty index or D-index).

Word Embeddings

The first step in analysis was to determine the best representation of the text. To reduce the computational cost of creating and working with word embeddings, the text was modified so that only the top 2,000 most frequent words in the corpus were retained, with all other words assigned a reserved token signifying a rare word. Approximately 2% of the corpus was rare words. Three different word embeddings were explored in this study (a) a GloVe embedding, (b) a free, untrained embedding and (c) a custom skip-grams embedding (Mikolov et al., 2013), trained on the corpus that included all YouTube transcripts and assessment items. All three embeddings had the same dimensions: 2000 words by 128 embedding attributes. The skip-grams

algorithm functions by training a shallow neural network that includes an intermediate dot-product layer. The network takes as input three words from the corpus (in a one-hot encoded format): a randomly selected *target* word, a *context* word selected from within a given window of the target word, and a *negative samples* word, chosen at random from the rest of the corpus. The goal of the model is to discriminate between the context and negative sample word given a target word. After training, the dot-product layer is extracted from the network and used as an embedding space that represents the relationships between words within he corpus (Mikolov et al., 2013). In this study we trained the skip-grams network on a total of 100,000 samples from the corpus. The sampling window was five words in either direction of the target word, with one negative sample per correct observation.

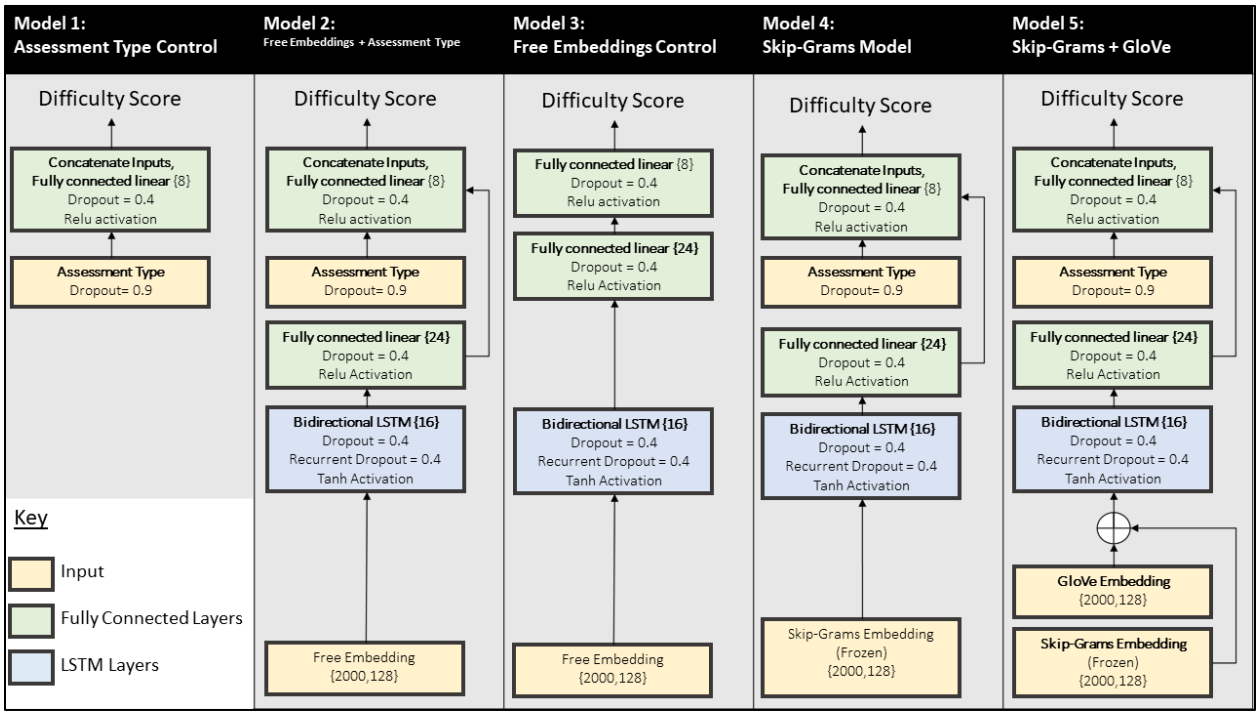
Model Structures

Five different models were initially evaluated. The goal of each model was to reconcile the various texts within the class into a single lingual model that could be used to uncover the relationship between assessment material and class content. For training, each model took as input the text of the assessment item and attempted to predict the difficulty of the assessment question, with a score of 1 representing a question all students got correct, and a score of 0 representing a question that all students got wrong. Assessment texts were truncated at a maximum of 250 words. 5 assessment items were truncated. In each model, an embedding representation of the text of assessment items was fed into a bidirectional LSTM layer. Most models also utilized a one-hot-coded variable that represented the type of assessment question (e.g., exam, assignment, practice exam, “other”). This additional input allowed the models to adjust to the different grading standards used for different types of assessment material; while tests were automatically graded, the instructor manually graded assignments and may have been

lenient in grading. The last time step of the bidirectional LSTM layers, as well as assessment type layers, were then fed into a fully connected layer and a final output node. Figure 1 gives the architecture of all five models.

Figure 1

Model architectures



The skip-grams word embedding was the only feature explicitly connecting the assessment text to the lecture text, as the embedding space was trained using the corpus that contained both sets of text. Because we aimed to bridge the linguistic divide between assessment and content language, the skip-gram embedding was included in all cases where an embedding was used. Without this custom embedding, we could not justify generalizing our model beyond the training and validation data (assessment text) to the test data (class lecture content). As a result, a model that utilized only the GloVe embedding was not tested. The common method of

fine-tuning the GloVe embedding was also not attempted, as it was deemed too computationally costly to be performed at scale. Instead, one model that included both the GloVe and Skip-grams embedding was created. Both layers were fed into a dense layer, where in theory the information from both embeddings could be reconciled, leading to similar information as a fine-tuned GloVe embedding. Note that model 1 utilizes only the one-hot-coded assessment information, and model 3 utilizes only an untrained 128d embedding space. These models were not hypothesized to perform well, but instead treated as baselines models to estimate if the success of more complex models was most likely due to mainly linguistic features, assessment-type features, or if some synergistic interaction existed between these two feature sets.

We note here specific concerns related to *overparameterization,* when the number of training parameters exceeds the size of the training data, as it does in the case of models 2 – 4. Unlike deterministic statistical methods where a model always converges to a global minima, neural networks have been shown to be robust at times to overparameterization. While an exact theoretical or empirical has not been found, several rules-of-thumb have been suggested. First, overfitting of a model is more likely in the case of wide models, where the number of trainable parameters in a single layer nears the size of the training set (Oymak & Soltanolkotabi, 2020). Second, regularization via dropout, data augmentation, and other means can discourage overfitting and increase the robustness of the model (Zhang et al., 2021). Finally, because overparameterization is an open question in deep learning, cross validation allows for the empirical measurement of overfitting within a model. This rules-of-thumb suggests models with the free embedding may be most likely to overfit, as there is a single wide layer within these models. With other models, the widest layer is the LSTM layer, with 16 units. It should be noted that each LSTM node includes several trainable parameters, though these parameters are not

independent. With current knowledge of deep networks, it is not possible to determine if a model will overfit before training. Fortunately, it can be measured through cross validation.

Figure 2

Tested models

	Assessment Type	Free Embedding	Frozen Embedding	GLoVe Embedding
Model 1	**YES**	No	No	No
Model 2	**YES**	**YES**	No	No
Model 3	No	**YES**	No	No
Model 4	**YES**	No	**YES**	No
Model 5	**YES**	No	**YES**	**YES**

For regularization, a dropout of .4 was used after both the LSTM and dense layers. Because models tended to train to a local minimum where scores were predicted as the mean of the assessment type to which the item belonged, a dropout of .9 was applied to the assessment type data before concatenating to the dense layers, ensuring the model did not over-rely on assessment type features.

Both the densely connected and output layers used a rectified linear unit (Relu) activation function. Although a sigmoid activation function could also be used on a data set where $y \in [0,1]$, initial exploration of the use of a sigmoid activation function showed that it was slow to converge to a solution. Initial models had difficulty modeling questions where y = 1 due to the exponential features of the sigmoid function. Although the Relu activation could potentially return values outside the range of the data, this disadvantage was thought to be outweighed by increased training speed and accuracy.

All models were trained in R using the Keras functional API (Chollet & Allaire, 2017). Models used an RMSprop optimizer and mean squared error loss function, and training was terminated with patience = 5 or a maximum of 50 epochs. Because of the limited data size, a 5-fold (as opposed to 10-fold) cross-validation was used for hyperparameter tuning, with four folds shuffled between during tuning as needed. Each fold included 245 original assessment items as well as ten blank-noised copies of each assessment item in that fold, for a total of 2695 training examples per fold. A fifth fold was withheld for testing (n=246) and not blank-noised. The bidirectional LSTM layer was evaluated at sizes 16, 32, and 64 for each model individually. The dense layer was evaluated at sizes 8 and 24 for each model individually.

Quantitative model evaluation

To select a final model to explore in detail, all models were evaluated based on their r-squared value on the validation data, and their over-fit on the validation compared to the training data. Additionally, pairwise F-statistics were calculated for all model pairs to determine if the difference in r-squared was significant. Because of the limited data size, the same data was used for model-stopping as well as for validation metrics. This process was used uniformly across all five models to ensure equitable comparisons when selecting a final model.

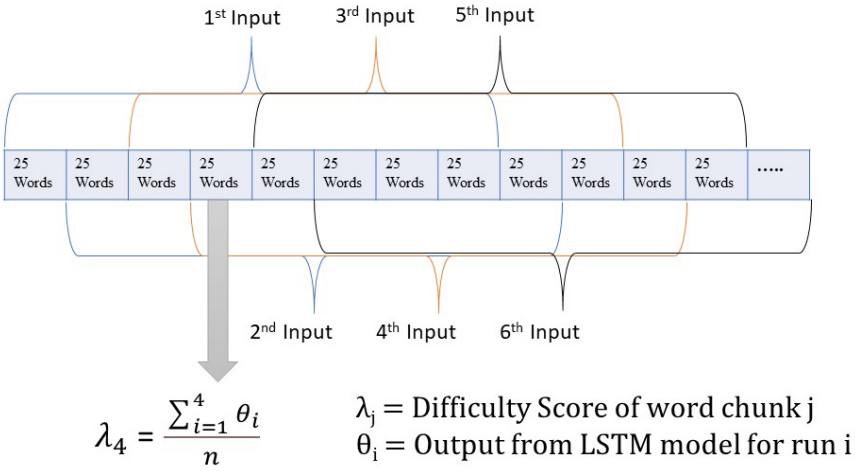
The final model was also evaluated by assessment type, with an r-squared value calculated for each assessment type separately. This type of evaluation allowed for the separation of model fit due to linguistic features and model fit due to simple linear shifts based on assessment type information. Because the selected model included the skip-grams generated word embedding, some additional analysis was undertaken to better understand and visualize the nature of the embedding.

Qualitative model evaluation

After selection of a final model, the text of the lecture material was fed through the neural network in chunks of 200 words. This meant that chunks of text could start and end in the middle of sentences. To mitigate this issue, a rolling window was used to input the text into the model, with the 200-word window shifting 25 words each iteration. This meant that multiple predictions were made for each 25-word chunk in the model. The final difficulty score for each 25-word chunk was taken as the mean of all iterations of which that chunk was a member (see Figure 3).

Figure 3

Mapping difficulty to lecture text.



The final goal of the model is to map assessment difficulty onto lecture material so that the AI toolkit can support teacher analysis of student learning. This is accomplished by connecting shared language across diverse course material, a task that is otherwise difficult and time-intensive for teachers utilizing a flipped-inquiry-based instructional model. However, this is not an entirely objective goal, and no objective metric exists for testing the accuracy and utility of the model. Instead, a novel qualitative method was developed for exploring the accuracy and utility of the model. A subset of key lecture passages (as suggested by the model) were compared to instructor-generated codes that identified easy and difficult topics within the course. The

instructor's codes were used as a competitive benchmark against which to evaluate the model. This had several advantages. First, because the instructor was not directly involved in this research, they were not predisposed to the same biases as those more closely involved in the process. Second, because of the unique expertise of the instructor and their decade of research related to this specific course (Jensen et al., 2013; Jensen et al., 2014, 2018; Jensen & Lawson, 2011; Kummer et al., 2016) they offered a high level of expertise not common among instructors. This expertise helps to increase the validity of the qualitative model evaluation. Finally, the involvement of the instructor also allows us to determine the interpretability of the results from the AI toolkit – a key concern in our design. Thus, a structured interview was performed where the expert instructor provided researchers with what, in their opinion, were the three to four easiest and most difficult topics covered in the class. Then, the three lecture sections with the highest and lowest predicted difficulties were compared to this list to estimate both the accuracy and utility of the model.

Results

Results indicated that the model was effective at identifying the most difficult topics within the course material. Based on this model, it is feasible to develop AI tools that explain these findings to instructors. Properly implemented, this AI toolkit could greatly decrease the time needed to evaluate how students' scores on assessment material are connected to instruction. Table 1 displays the results of the five evaluated models. The assignment-type-only baseline model captured 24% of the variance in the data and did not over-fit the validation data. Both models that utilized a free, appear to have reached a near-perfect fit on the data, achieving an r-squared of 90%. It is noteworthy that blank-noising was employed at a rate of 10%; it may be only this regularization that kept models 2 and 3 from reaching a perfect minima on the

training data. Both models that utilized the skip-grams-generated embeddings performed similarly with a validation r-squared of .31 and .27. Pairwise F-statistics comparing the r-squared of all models (on validation data) were all significant with p <.0001. While the model that augmented this data with the GloVe embedding did perform better on the training data, it caused the model to overfit the validation data more severely. The addition of the GloVe embedding also increased the complexity and training time. Because of both the overfitting effect of the GloVe embedding and for parsimony, the model that did not use the GloVe embedding (model 4) was selected as the final model.

Table 1

Model comparison

Model	Training Data R-Squared	Validation R-Squared	Overfit Percent
Model 1: Assignment Type Baseline	.23	.24	**- 0.01**
Model 2: Free Embedding + Assignment Type	.90	.10	.80
Model 3: Free Embedding Baseline	**.91**	.10	.81
Model 4: Class Transcript Frozen Embedding +Assignment Type	.46	**.31**	.15
Model 5: Class Transcript Frozen Embedding + GloVe Embedding + Assignment Type	.55	.27	.23

Quantitative model evaluation

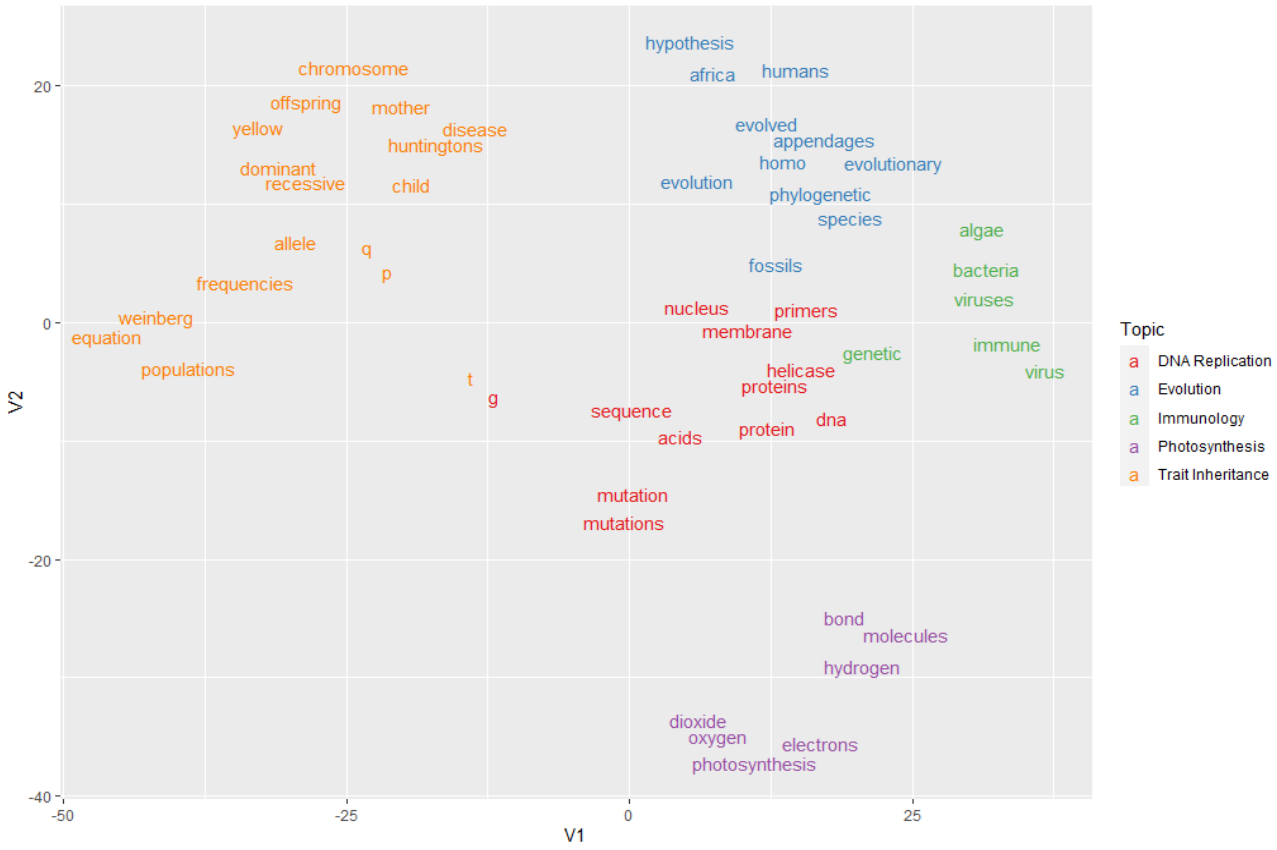
Skip-grams embedding

An important step in building the AI toolkit was validating the assumption that the skip-gram algorithm properly represented the linguistic space in a way that could generalize across both the assessment and content in a flipped inquiry-based classroom. To understand the nature and quality of the skip-grams word embedding, a t-SNE plot was created to visualize key content words (Figure 4). Additionally, several individual words were evaluated to determine how closely related key content words were. The t-SNE plot revealed a clear separation of content words into topical clusters. Plurals (e.g., mutation, mutations; protein, proteins) are clustered

together. Additionally, words related to examples used in the class (e.g., Huntington's disease for genetic inheritance) are closely related to the content words they describe.

Figure 4

t-SNE plot of skip-grams embedding



The intermediate layer of the skip-grams network that is extracted as a word embedding is a dot-product layer, and the cosine of two-word vectors x and y is proportional to the dot product of the same vectors. Because of this, calculating the cosine similarity of the vectors for two words in a skip-grams-generated embedding accurately quantifies the relationships of words in the embedding space. The cosine similarity is a value between -1 and 1, where closely related words have a value near 1, and less related words have values near -1. Cosine similarity, by

definition, is a non-centered Pearson's correlation, and while it does not have the same strict interpretation, it can be conceptualized similarly. Table 2 gives select examples of the most similar words. Both content words and words used for examples words are effectively modeled, suggesting the embedding captured example-driven linguistic relationships inherent to the topics in this inquiry-based classroom. For example, *yellow* is most related to *white* and *dominant*. This relationship comes from the example of Gregor Mendel's early experiments on gene inheritance that used pea color as a key trait. These example/content word relationships may help explain why the GloVe embedding underperformed, as it relies on a more general definition of words and would underrepresent disciplinary-specific examples.

Table 2

Select cosine similarities

"Hardy"		"Experiment"		"Yellow"		"Evolution"	
Related Words	Cosine Similarity	Related Words	Cosine Similarity	Word: Yellow	Cosine Similarity	Related Words	Cosine Similarity
weinberg	.95	test	.61	white	.60	convergent	.74
equilibrium	.69	independent	.58	dominant	.57	homology	.60
predicted	.55	design	.56	round	.55	selection	.55
assuming	.54	variables	.45	green	.55	natural	.52
pq	.51	hypothesis	.53	yy	.54	analogy	.49

Predictive power

Findings indicated that the type of assessment (e.g., homework, exams) plays a central role in supporting learning. Table 3 displays information about the accuracy of the model when data is separated by assessment type. By calculating r-square by category, we remove assessment type information that may lead to superficially high results not related to linguistic features (see Table 1, Model 1). When splitting out data by assessment type, linguistic features account for 17% of the variability in scores for both homework and exams. The model has virtually no predictive power for practice exam questions. The "other" category is a combination of "video quizzes," which asked participants if they watched the video assigned for that week, and the

Lawson test of scientific reasoning (LCTSR) which was administered to students at the beginning and end of the semester. The model performed best on this category. However, this is most likely because the LCTSR test and video quizzes were one-hot-coded as separate categories, meaning the calculations in Table 3 have not fully removed assessment-type information from the "other" category. They are combined in Table 3 due to low sample size across these subcategories and included for transparency. However, this r-squared value is not strictly interpretable as having removed all assignment-type information and should not be interpreted the same as the rest of the categories as it is inflated substantially.

An exact interpretation of the r-squared value is not necessarily possible. For example, we can imagine a hypothetical AI that perfectly understood all the language in the content and assessment material. Due to random guessing by students, variance in the student sample, and other ways students learned that are not included in the model, we would not expect the R-squared value to approach one. In other words, the proportion of linguistic information is not equal to the R-squared value, but instead $\frac{R^2}{l}$, where l is the total proportion of learning that is generated by linguistic features within the corpus.

Table 3

Model fit by assessment type

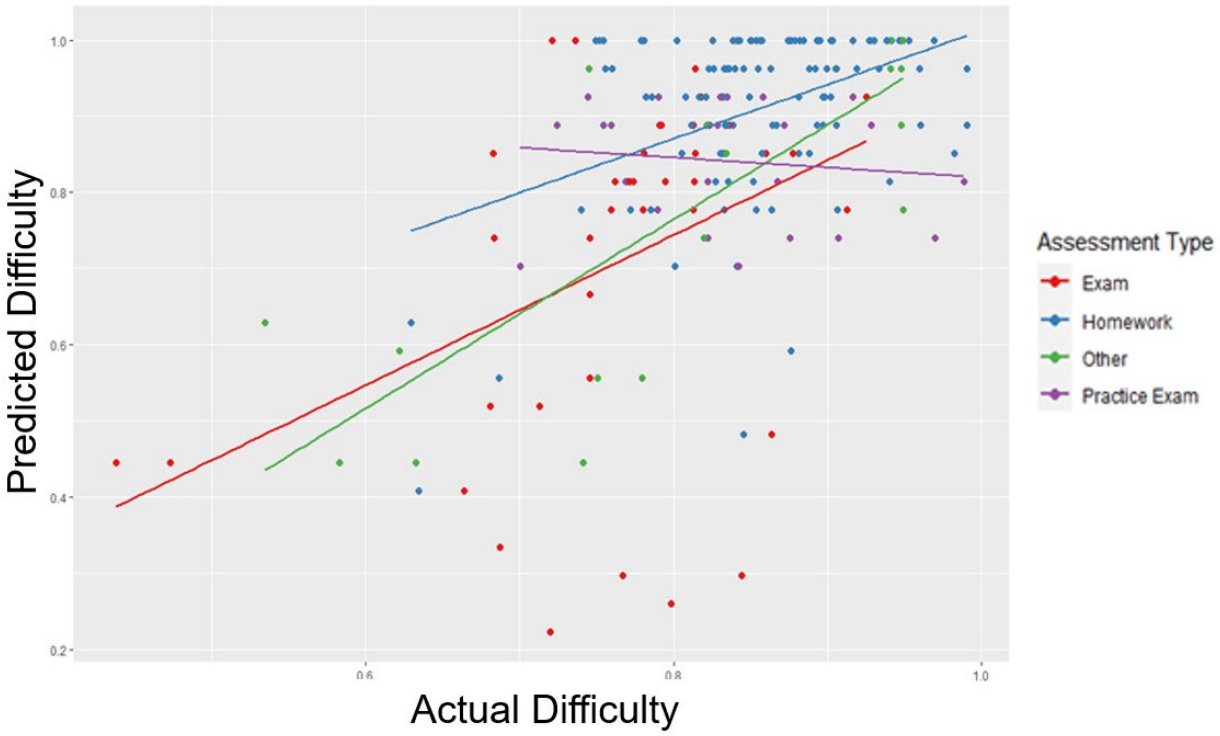
Assessment Type	N	Data Standard Deviation	Residual Mean-Squared Error	R-Squared
Homework	106	0.11	0.06	0.17
Exams	37	0.23	0.09	0.17
Practice Exams	23	0.07	0.07	0.01
Other	20	0.21	0.08	0.66
All	186	0.17	0.14	0.31

Figure 5 shows the predicted model difficulty on the y-axis, and the actual difficulty on the x-axis. Points are colored by assessment type, and an ordinary least squared line of y|x is

included to help visualize the relationship between the predicted and actual difficulty values. This figure gives some intuition as to the reason for the accuracy of the model across different categories. Because practice exams were open-book and voluntary (i.e., although a score was calculated, it did not affect student's final grade) most question scores fell in a smaller range compared to exam and homework questions (all purple points have an x-value > .7. Because the practice exam is open-book, other factors besides linguistic features may have accounted for the variation in difficulty. For example, a practice exam question asking student to list 10 organelles would be very easy to look up. Memorizing and reproducing this on a closed-book exam would be more difficult. Both exam and "other" questions appeared to be treated similarly by the model. When dealing with exam questions, the model appeared to have higher discrimination at the tails of the data but struggled to properly assign difficulty to moderately difficult questions (between a difficulty of approximately 0.6 and 0.85).

Figure 5

Model fit by assessment type

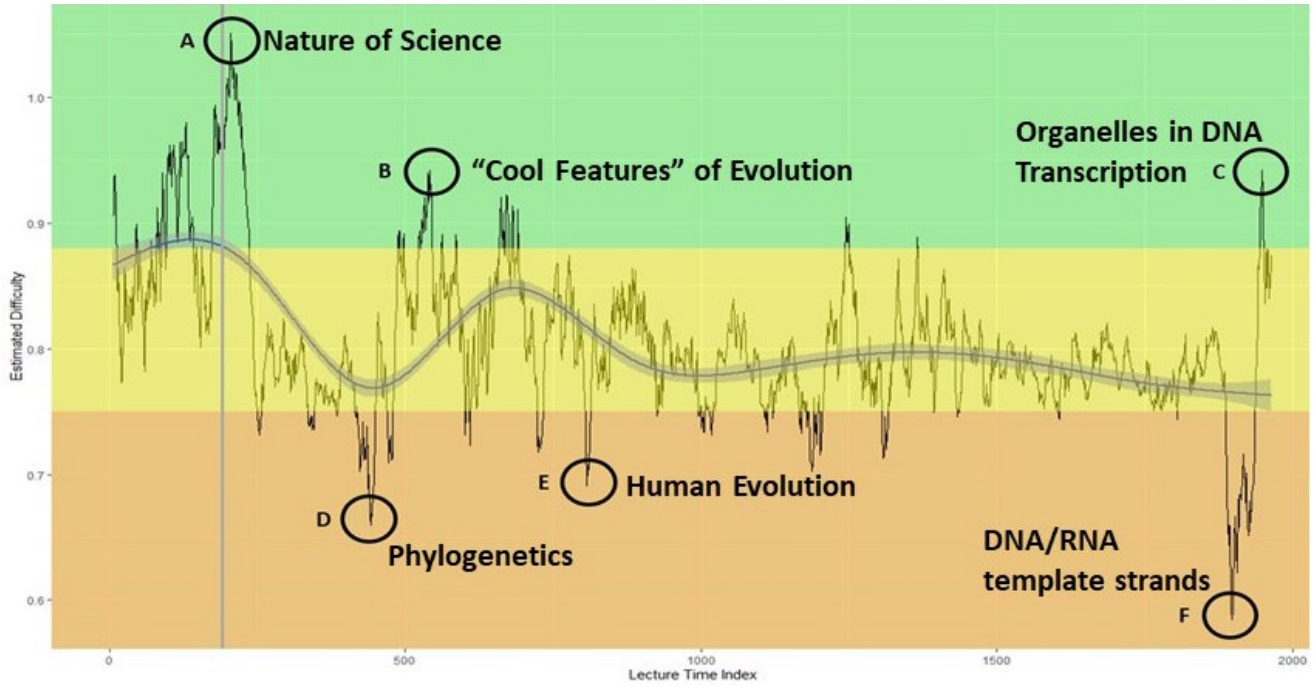


Qualitative model evaluation

Figure 6 displays the results of feeding the semester's lecture material through the model. The x-axis displays all videos from the semester in the order they were assigned, smoothing between videos. Each time index represents one block of 25 words. Before selecting the three easiest and most difficult lecture portions for further evaluation, the first video (all material to the left of the vertical line on Figure 6 at time index = 220) was removed. This video covered the syllabus and class structure, not content. It is noteworthy that when faced with content outside the problem space it was trained, the model tended to predict that material was easy. The three easiest and most difficult portions of lecture material as predicted by the model are highlighted in Figure 6 at points A-F and are explored in further detail later. At point A, the model predicted a difficulty above 1 due to the use of a Relu activation.

Figure 6

Predicted lecture difficulty over semester



To determine if the model could be generalized beyond assessment data to lecture material, the instructor provided what they believed were the easiest and most difficult topics in the course based on their expertise (see Table 4).

Table 4

Instructor-selected topics

Expert-Identified Topic	Expert-Assigned Difficulty
Nature of science	Easy
Convergent evolution	Easy
Homology versus analogy	Easy
Macro Evolution (e.g., phylogenetics, speciation events)	Hard
Coding and template DNA/RNA strands	Hard
Hardy-Weinberg equilibrium	Hard
Photosynthesis and cellular respiration processes	Hard

After soliciting topics from the instructor, the six selected lecture portions were extracted from the transcripts and compared to the instructors' topics. Table 5 compares the model assigned difficulty, the instructor assigned difficulty and includes an excerpt from each selected portion of the lecture transcript. All three of the model-selected *difficult* lecture portions matched the instructor codes. Of the three model-selected *easy* portions, (1) one matched instructor coding, (2) one contradicted instructor coding, and (3) one was not a topic selected by the instructor as either easy or difficult.

Table 5. *Key lecture periods*

Topic of excerpt	Model-Assigned Difficulty	Matches Expert-Assigned Difficulty?	Sample Excerpt from Transcript (Punctuation introduced for clarity)
Nature of science (A)	Easy	Yes	*"[the fruit flies] grew in normal conditions. So, that was kind of their control to see what would happen and then they weighed them. So, what results would support the hypothesis? Well, we should see bigger bugs with more oxygen. This graph is showing males, this graph is showing females and then our independent variable… is actually displayed in two different colors of graphs on the x-axis…"*
Casual example of "cool features" animals have evolved (B)	Easy	Not in Expert's Codes	*"[When mammals are pregnant] they're carrying this giant baby with them not to mention the fact that you're constantly feeding [the baby] from your own food supply. … But we can see based on how much mammals have taken over that … the advantages outweigh the disadvantages. So those are just my thoughts on some of these really cool features that that we've evolved as animals have progressed."*
Role of organelles in DNA transcription (C)	Easy	No	*"The last step is translation at this point our processed MRNA that we have here gets attached to the ribosome out in the cytoplasm, and it gets read. So, we're going to have transfer RNAs, bring in amino acids depending on what their anticodon is, so if this transfer RNA is going to attach to AUG which is the codon it would read UAC and if we look at the code we have to find AUG. So first letter "A," second letter "U,"…"*
Phylogenetic trees (D)	Hard	Yes	*"We know that there is one trait that every single species had, and this one was trait 21. So, we're going to bring that one over here because it even included our out group. Now we'll just go through one at a time with the remaining trait...."*
Human evolutionary history (E)	Hard	Yes	*"Paranthropus is another side group. Paranthropus didn't survive beyond to where we see homo species coming about. …. It has sexual dimorphism, so very much like a gorilla where the males are really large and the females are much smaller. One of the things you should notice about these species is that sagittal crest on the top of their head…"*
Coding and template DNA/RNA strands (F)	Hard	Yes	*"We start with our DNA which is the original recipe in the nucleus and we make a photocopy into RNA which is a close cousin of DNA that we'll talk about in a moment. … We'll talk about each step of this process…you'll notice the name of the sugar, (ribonucleic acid versus deoxyribonucleic acid) so there's a slight difference in the sugar…"*

Discussion

In designing an AI toolkit for teaching reflection, we considered (a) how the toolkit leverages existing data streams and eases the burden of data collection, (b) how the toolkit allows for automated analysis, (c) how the output of the analysis could aid in teacher reflection. Here, we discuss these three factors in a more integrated manner and explore the implications of the models embedded in the toolkit as it relates to supporting teacher reflection.

Teacher reflection by necessity often relies on qualitative data such as teacher journals to support reflection on practices and instructional design decisions (Cornford, 2002). When quantitative data is collected, it is frequently through temporary, researcher-lead means that are not sustainable beyond the initial research cycle (Persico & Pozzi, 2015). The AI toolkit suggested here offers an additional source of information, data from student's grades, and leverages this data to infer the difficulty of course material. The use of these two existing and readily available data streams, student grades and the class corpus of linguistic material, means this model could be employed without the need for extensive manual coding. Although the contexts of and exact means of data collection differ based on context, this method could be employed in any class for which there exists a body of graded material and a collectible corpus of linguistic information from textbooks, lectures, or other materials. Further replication would be needed to see how generalizable the results of this case study are, and what course features are necessary to reproduce the accuracy seen here.

The use of existing data and automatable analysis eases the burden of creating such an AI toolkit; however, it is equally important that the AI toolkit is useful in aiding in teacher reflection. This question is not directly addressed in the analysis of this paper, as the main goal of this study was to show the ability to create an accurate model based on available student

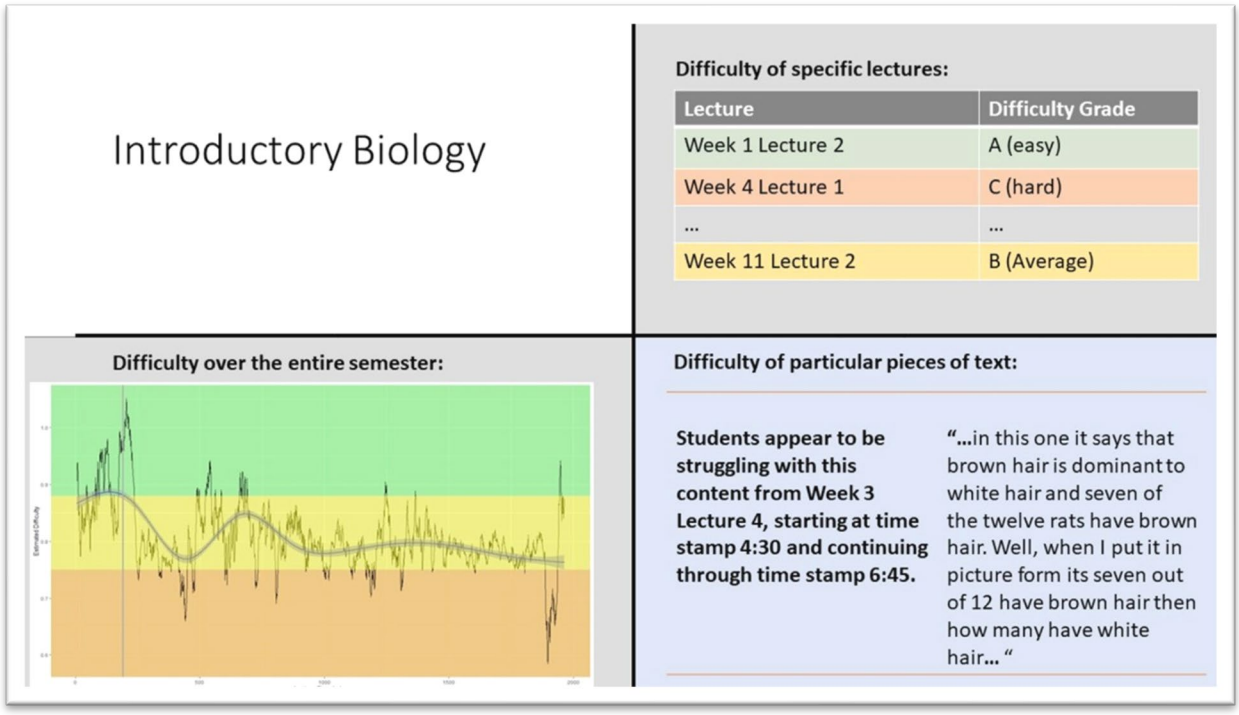
grades and a class corpus. However, we discuss below the ways in which we believe the model explored in this study could properly be embodied in an AI toolkit that aids in teacher reflection. Additionally, we give to brief hypothetical examples of the types of insight the data in this study would give to both a novice and more expert instructor.

As previously discussed, teacher reflection often relies on an outside actor or support to foster discussions and interpretations of practice. The AI toolkit explore in this study could act as an additional support. By directly connecting the language used by the teacher in instructing their students to the student's outcomes, the toolkit can inform teachers of what may and may not be working in the classroom. One difference between what we are suggesting and what is common in the reflection literature is a stronger focus on *instructional design*; most literature on reflection focuses largely on classroom management, instructional strategies, scaffolds, and other real-time teaching decisions (Knight et al., 2006). However, teachers are often equal parts instructors and instructional designers (Warr & Mishra, 2021). This AI toolkit could allow teachers to reflect more on their role as instructional designers and on the effectiveness of the instructional materials they create. Additionally, this tool does not preclude or diminish the qualitative insights teachers have into their role as instructors. In our design, we included the instructors' interpretation of content difficulty, which allows for comparisons between student experiences and instructor expectations. The AI toolkit is agnostic to interpretation; it points out what content appears to be most difficult to students without offering a reason why this is the case. It is the role of the teacher to decide *why* students might be struggling with material. We explore specific examples of this type of interpretation in the specific context of the data collected in the next section. Figure 7 shows one potential form the "home page" of an AI toolkit for teacher

reflection could take. Additional information, such as revisualizations of the information in Figures 4 and 5 could also be useful inclusions.

Figure 7

A mock-up of a hypothetical content summary of an introductory biology course



Expert Instructor

An expert instructor may have many of the same characteristics as the instructor of the course used in this study; the expert instructor may have taught and continually modified the same course for many years, acting as both instructor and instructional designer. Because of this, they have a high level of knowledge about the course. Additionally, marginal changes in curriculum and assessment over the course of many class iterations may mean that the original learning objectives are not aligned with the actual material and assessment material. In the specific case

of the course considered in this study, the expert instructor may glean several insights from the AI toolkit.

First, although the instructor appears to have a high level of understanding of the most difficult topics in their course, the AI toolkit further narrows the focus from the four topics suggested as difficult by the instructor (see Table 3); two of the three most difficult passages deal with evolution, more specifically with phylogenetic trees and traits of related species (Table 4, items D & E). If the instructor has only a few days between semesters to improve one unit, then that time may be best spent on phylogenetic trees. However, the instructor may also believe that phylogenetics is inherently difficult, and little can be done to easily improve this section. In this case, the expert instructor may select a different section—such as DNA transcription and translation—for revision. With any of the topics discussed here, it is the teacher's responsibility to interpret the output of the model and decide the best course of action. Second, based on the results, the expert instructor may also experiment with the use of practice exams in supporting student learning. In this study, the model was able to accurately predict students' scores on exams, but not on the practice exam. This may suggest that the practice exam is not a good reflection of the exams, and, by extension, not a good means of preparing students for the final exam. In both two examples, the teacher can make revisions, repeat the course, and then quantitatively measure if the changes were successful.

Novice Instructor

A novice instructor who is still developing instruction and learning about their students may use the AI toolkit for a much different purpose. Whereas the expert instructor may use the AI in a more targeted way, the novice instructor may be trying to get a general grasp of what concepts students struggle with the most or is focused on ensuring that the assessments provided

map to the content being taught (i.e., valid). When the AI toolkit flags information as difficult, the novice teacher may reflect and realize an assessment item needs to be modified It is likely that novice teachers are more inflexible in changing their instructional practices and may not recognize meaningful patterns in the assessment data (Berliner, 2001). In such a scenario, the novice instructor may benefit from summary information on student performance and lecture difficulty, such as conceptualized in Figure 7. This use would be more in line with the more traditional role of reflection in focusing on instructor (not instructional design) skills. The AI toolkit offers a holistic view of when things are easy and when things are hard, aiding teachers in reflecting on what skills they need to develop during their formative years as an instructor. Ideally, the dashboard could be used across multiple semesters to view how modifications to curriculum effect student performance.

Limitations and future research

Future research could further develop the methods explored in this study in several ways to increase the accuracy and utility of these methods. Some simple modifications could lead to marginal gains in accuracy. Some of the inaccuracies in prediction are likely due to linguistic differences between lecture and assessment differences. While there is no perfect solution to this problem, the inclusion of more material in the skip-grams embedding, such as slides, textbook material, and other content-related material may lead to higher accuracy. This text may act as a better medium for projecting difficulty as it is not prone to the same transcription errors as lecture material. Unlike the lecture material, it also includes punctuation, so arbitrary truncations would not be needed to feed the text through the model. This projection could also act as a method for triangulating the accuracy of projections onto lecture material.

Future research may also address increasing the trustworthiness and interpretability of the AI toolkit. Because the LSTM network is inscrutable, it is difficult to know why the model makes predictions in the manner it does. For this reason, the model is vulnerable to failing non-gracefully and without warning. Additionally, the correlations between predicted and actual difficulty are difficult to interpret, (see Table 3) as it is unclear what the upper bound of the model accuracy is. One method for combating both of these issues would be to analyze which training inputs (i.e. assessment items) are most influential or each portion of lecture material (Charpiat et al., 2019). This would effectively lead to a model that explained its decisions; the model would return both the most difficult lecture material and the assessment material that led to this prediction. Errors due to superficial linguistic relationships or other modeling issues would be easily detected when manually reviewing the model findings.

Conclusion

This study explored methods for developing a flexible AI toolkit for teacher reflection. Initial results show that skip-gram word embeddings and LSTM networks can be used to predict the difficulty of assessment material with moderate difficulty, and initial qualitative analysis shows that this same model retains utility when extended to lecture material. We suggest future research may work to extend this model to multiple contexts, additional curricular material, and focus on the addition of model-generated explanations to increase the trustworthiness and utility of the AI toolkit.

Declarations

Funding

This study was support by the Indiana University Instructional Systems Technology Kemp Research Grant.

Conflicts

All authors declare that they have no competing or conflicting interest in regards to any research presented within this publication.

Availability of data and material

Data from this project is available by request to the corresponding author.

Code availability

All R code used in this study is available upon request to the corresponding author.

Ethics approval

This research was conducted under supervision by and with approval from the Institutional Review Board of Brigham Young University.

References

Baker, R. S. (2016). Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education*, *26*, 600–614. https://doi.org/10.1007/s40593-016-0105-0

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, *35*(5), 463–482. https://doi.org/https://doi.org/10.1016/S0883-0355(02)00004-6

Bokhove, C., & Downey, C. (2018). Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, *11*(2), 205979911879074. https://doi.org/10.1177/2059799118790743

Charpiat, G., Girard, N., Felardos, L., & Tarabalka, Y. (2019). Input similarity from the neural network perspective. *Advances in Neural Information Processing Systems*, *32*(NeurIPS), 1–10.

Chau, H., Labutov, I., Thaker, K., He, D., & Brusilovsky, P. (2021). Automatic Concept Extraction for Domain and Student Modeling in Adaptive Textbooks. *International Journal of Artificial Intelligence in Education*, *31*(4). https://doi.org/10.1007/s40593-020-00207-1

Chollet, F., & Allaire, J. (2017). *R Interface to Keras*. GitHub. https://github.com/rstudio/keras

Cornford, I. R. (2002). Reflective teaching: Empirical research findings and some implications for teacher education. *Journal of Vocational Education and Training*, *54*(2), 235. https://doi.org/10.1080/13636820200200196

Council, N. R. (2001). *Knowing What Students Know* (J. W. Pellegrino, N. Chudowsky, & R. Glaser (eds.); 1st ed.). The National Academies Press. https://doi.org/https://doi.org/10.17226/10019

Dillenbourg, P., Prieto, L. P., & Olsen, J. K. (2008). *Classroom Orchestration*. 180–190.

Escamilla, I. M., & Meier, D. (2018). The Promise of Teacher Inquiry and Reflection: Early Childhood Teachers as Change Agents. *Studying Teacher Education*, *14*(1), 3–21. https://doi.org/10.1080/17425964.2017.1408463

Feng, M., & Heffernan, N. T. (2005). Informing Teachers Live about Student Learning : Reporting in the Assistment System. *Tech., Inst., Cognition and Learning*, *3*(508), 1–14.

Feng, S., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). *A Survey of Data Augmentation Approaches for NLP*. 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84

Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., & Lester, J. (2020). Predictive Student Modeling in Game-Based Learning Environments with Word Embedding Representations of Reflection. *International Journal of Artificial Intelligence in Education*, 1–23. https://doi.org/10.1007/s40593-020-00220-4

Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education*, *11*(1), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U

Jensen, J., Holt, E. A., Sowards, J. B., Heath Ogden, T., & West, R. E. (2018). Investigating Strategies for Pre-Class Content Learning in a Flipped Classroom. *Journal of Science Education and Technology*, *27*(6), 523–535. https://doi.org/10.1007/s10956-018-9740-6

Jensen, J., Kummer, T., & Banjoko, A. (2013). Assessing the Effects of Prior Conceptions on Learning Gene Expression. *Journal of College Science Teaching*, *42*(4), 82–91.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the Test...or Testing to Teach: Exams Requiring Higher Order Thinking Skills Encourage

Greater Conceptual Understanding. *Educational Psychology Review*, *26*(2), 307–329. https://doi.org/10.1007/s10648-013-9248-9

Jensen, J., & Lawson, A. (2011). Effects of collaborative group composition and Inquiry instruction on reasoning gains and Achievement in undergraduate biology. *CBE Life Sciences Education*, *10*(1), 64–73. https://doi.org/10.1187/cbe.10-07-0089

Khoshsima, K., & Nosratinia, M. (2019). Inspecting the Prospect of Augmenting Classroom Management by Reflective Teaching and Use of Motivational Strategies. *International Journal of Applied Linguistics and English Literature*, *8*(1), 93–103. https://www.journals.aiac.org.au/index.php/IJALEL/article/view/5250

Knight, P., Tait, J., & Yorke, M. (2006). The professional learning of teachers in higher education. *Studies in Higher Education*, *31*(3), 319–339. https://doi.org/10.1080/03075070600680786

Kummer, T. A., Whipple, C. J., & Jensen, J. L. (2016). Prevalence and Persistence of Misconceptions in Tree Thinking †. *Journal of Microbiology & Biology Education*, *17*(3), 389–398. https://doi.org/10.1128/jmbe.v17i3.1156

Lee, J.-H., & Cha, K.-W. (2020). An Analysis of the Errors in the Auto-Generated Captions of University Commencement Speeches on YouTube Jeong-Hwa. *Journal of Asia TEFL*, *17*(1), 143–159. https://doi.org/10.18823/asiatefl.2020.17.2.10.463

Leitner, P., Khalil, M., & Ebner, M. (2017). Teaching and Learning Analytics to support Teacher Inquiry: A Systematic Literature Review. In *Learning Analytics: Fundaments, Applications, and Trends, Studies in Systems, Decision and Control* (Vol. 94, Issue January). https://doi.org/10.1007/978-3-319-52977-6

Liaqat, N. (2017). Reflective Practices: a Means To Teacher Development. *Asia Pacific Journal*

of Contemporary Education and Communication Technology, *ISSN*(3), 2205–6181.

Loughran, J. J. (2002). Effective reflective practice in search of meaning in learning about teaching. *Journal of Teacher Education*, *53*(1), 33–43. https://doi.org/10.1177/0022487102053001004

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, *50*(9), 741–749. http://psycnet.apa.org/journals/amp/50/9/741.pdf&uid=1996-10004-001&db=PA

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mislevy, R. (2005). Evidence-Centered Assessment Design: Layers, Structures, and Terminology. *Principled Assessment Designs for Inquiry Technical Report 9*, *9*(July), 46. http://padi.sri.com/downloads/TR9_ECD.pdf

Mor, Y., Ferguson, R., & Wasson, B. (2015). Editorial: Learning design, teacher inquiry into student learning and learning analytics: A call for action. *British Journal of Educational Technology*, *46*(2), 221–229. https://doi.org/10.1111/bjet.12273

Oymak, S., & Soltanolkotabi, M. (2020). Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks. *IEEE Journal on Selected Areas in Information Theory*, *1*(1), 84–105. https://doi.org/10.1109/jsait.2020.2991332

Ozogul, G., Karlin, M., & Ottenbreit-Leftwich, A. (2018). Preservice Teacher Computer Science Preparation: A Case Study of an Undergraduate Computer Education Licensure Program. *Jl. of Technology and Teacher Education*, *26*(3), 375–409.

Peercy, M. M., Sharkey, J., Baecher, L., Motha, S., & Varghese, M. (2019). Exploring TESOL teacher educators as learners and reflective scholars: A shared narrative inquiry. *TESOL*

Journal, *10*(4), 1–16. https://doi.org/10.1002/tesj.482

Persico, D., & Pozzi, F. (2015). Informing learning design with learning analytics to improve teacher inquiry. *British Journal of Educational Technology*, *46*(2), 230–248. https://doi.org/10.1111/bjet.12207

Saye, J., & Brush, T. (2002). The use of embedded scaffolds with hypermedia-supported student-centered learning. *Journal of Educational Multimedia and Hypermedia*, *1*(2), 1–12. http://www.editlib.org/p/8439

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00492-0

van Leeuwen, A., Knoop-Van Campen, C. A. N., Molenaar, I., & Rummel, N. (2021). How teacher characteristics relate to how teachers use dashboards: Results from two case studies in k–12. *Journal of Learning Analytics*, *8*(2), 6–21. https://doi.org/10.18608/JLA.2021.7325

Wallach, H. M. (2006). Topic Modeling : Beyond Bag-of-Words. *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, 977–984. https://doi.org/10.1145/1143844.1143967

Warr, M., & Mishra, P. (2021). Integrating the discourse on teachers and design: An analysis of ten years of scholarship. *Teaching and Teacher Education*, *99*, 103274. https://doi.org/10.1016/j.tate.2020.103274

Xie, Z., Wang, S. I., Li, J., Daniel, L., Nie, A., Jurafsky, D., & Ng, A. Y. (2017). Data noising as smoothing in neural network language models. *ICLR 2017*, 1–12.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

https://doi.org/10.1145/3446776