Latent Skill Mining and Labeling from Courseware Content

Noboru Matsuda
NCSU*
Noboru.Matsuda@ncsu.edu

Jesse Wood
NCSU*
jwood96706@gmail.com

Raj Shrivastava
NCSU*
rksmail20@gmail.com

Machi Shimmei
NCSU*
mshimme@ncsu.edu

Norman Bier
CMU†
nbier@cmu.edu

A model that maps the requisite skills, or knowledge components, to the contents of an online course is necessary to implement many adaptive learning technologies. However, developing a skill model and tagging courseware contents with individual skills can be expensive and error prone. We propose a technology to automatically identify latent skills from instructional text on existing online courseware called SMART (**S**kill **M**odel mining with **A**utomated detection of **R**esemblance among **T**exts). SMART is capable of mining, labeling, and mapping skills without using an existing skill model or student learning (aka response) data. The goal of our proposed approach is to mine latent skills from assessment items included in existing courseware, provide discovered skills with human-friendly labels, and map didactic paragraph texts with skills. This way, mapping between assessment items and paragraph texts is formed. In doing so, automated skill models produced by SMART will reduce the workload of courseware developers while enabling adaptive online content at the launch of the course. In our evaluation study, we applied SMART to two existing authentic online courses. We then compared machine-generated skill models and human-crafted skill models in terms of the accuracy of predicting students' learning. We also evaluated the similarity between machine-generated and human-crafted skill models. The results show that student models based on SMART-generated skill models were equally predictive of students' learning as those based on human-crafted skill models— as validated on two OLI (Open Learning Initiative) courses. Also, SMART can generate skill models that are highly similar to human-crafted models as evidenced by the normalized mutual information (NMI) values.

Keywords: skill model discovery, learning engineering, massive open online course, text mining, natural language processing

* North Carolina State University (NCSU), Raleigh, North Carolina
† Carnegie Mellon University (CMU), Pittsburgh, Pennsylvania

1. INTRODUCTION

The demand for high-quality online education has been growing rapidly. Accordingly, the need to efficiently build online courseware becomes more and more evident (Tyton Partners, 2020). The current literature suggests that adaptive learning (Yang et al., 2014; Jovanovic and Jovanavic, 2015; Liu et al., 2017; Chen et al., 2018; Imhof et al., 2020) and personalization (Walkington, 2013; Paquette et al., 2015; Dai et al., 2016) are essential components for broad dissemination of practical online courses.

In this paper, we refer to the mapping of course contents to a set of skills as a *skill model*, aka a knowledge component model (Koedinger, Corbett, and Perfetti, 2012). A "skill", by definition, represents a piece of knowledge that students are supposed to learn. Typically, the skill model consists of a set of skills that are supposed to be obtained by students.

Many of the adaptive learning technologies assume that the online course is equipped with a skill model. It is therefore extremely important to create the skill model that accurately reflects the target skills to be learned. The skill model, for example, often serves as a basis for a student model (Desmarais and Baker, 2012; Pelanek, 2017). In turn, information from the student model can be used for adaptive instruction—e.g., selecting and sequencing instructional materials, providing recommendations to facilitate learning, presenting student's progress and knowledge state, aka, the open learner model (Paquette et al., 2015; Dai et al., 2016; Pelanek, 2017; Chen et al., 2018; Gavrilovic et al., 2018; Imhof et al., 2020).

In addition to identifying latent skills, we would also argue that labeling the skills also provides instructors with important insights into instruction that will in turn improve effectiveness of the courseware. The presence of labeled skills apparently facilitates the refinement of instructional materials by developers and other experts (Martin, Mitrovic, Mathan, and Koedinger, 2005; Martin, Mitrovic, Koedinger, and Mathan, 2011). It also facilitates the analysis of student learning, which will further allow instructors to identify areas in which students need assistance (Bier et al., 2014).

Despite the importance of creating a high-quality skill model, identifying the skills required for an authentic course (say, with a semester's worth of instructional material) and mapping course contents to the skills is extremely costly. As far as the authors are aware of, there are very few existing online courseware equipped with skill models—e.g., Open Learning Initiative (OLI) at Carnegie Mellon University (Bier and Rinderle, 2011; Bier et al., 2019). As a result, *developing a pragmatic method for automatic skill discovery is a critical component of learning engineering for successful dissemination of an adaptive online course*.

The goal of this paper is to introduce a technology to discover latent skills from existing online courseware and fully annotate the contents with the discovered skills. The proposed technology for skill mining is called SMART—**S**kill **M**odel mining with **A**utomated detection of **R**esemblance among **T**exts. In the current work, the term "instructional texts" is used as a collective term referring to written didactic paragraphs and assessment items in the given online courseware (including question stems, answers, multiple-choice items, and hint messages).

SMART analyzes instructional texts on the given online courseware and clusters those texts based on their linguistic similarity. Each cluster of text is then labeled with an extracted keyword(s) that represents the latent skill that the corresponding cluster of the text is describing. Thus, *among the significant features of SMART are the ability to automatically tag individual clusters of texts with a human-readable label*.

The proposed SMART method was developed as a part of our on-going effort to study a suite of evidence-based learning engineering methods, called PASTEL (**P**ragmatic methods to develop **A**daptive and **S**calable **T**echnologies for next generation **E**-**L**earning). PASTEL is a suite of advanced technologies for courseware developers to iteratively build adaptive online courseware (Matsuda et al., in press). The goal of PASTEL is to provide courseware developers with data-driven scaffolding and feedback while they are building the courseware. Specifically, PASTEL aims to assist with the implementation of adaptive instructions that include the optimal sequencing of formative assessments and proactive scaffolding on answering assessment questions. SMART discovers a latent skill model by analyzing the course contents and tags individual instructional elements with a skill in the discovered skill model to realize those adaptive instructions.

The current paper addresses two practical challenges: (1) The discovered skills need to be annotated with human-readable labels for courseware engineers and instructors to make sense of what those skills are, and (2) all instructional texts used in the online courseware, i.e., didactic paragraphs and assessment items, need to be tagged with the discovered skills. The former challenge is to ensure the interpretability of the resulted courseware. When analyzing students' learning, for example, the human-readable skill model will provide courseware developers with meaningful insights into how to improve the courseware. The latter challenge is practically important for implementing adaptive courseware because the system will then be able to utilize the association among courseware contents to provide students with the adaptive scaffolding.

A major contribution of the current paper is to provide the current educational data mining literature with a proof of concept for an innovative application of a text mining method, SMART, for automatic skill discovery from online courseware content. SMART is the first in the current literature that can automatically discover *labeled* skills that are latent in the existing courseware without labor intensive analysis of courseware contents. SMART consists of existing, well-established technologies, including k-means clustering and keyword extraction. Yet, we also demonstrated that a well-designed combination of existing technologies was able to be applied to genuine online courseware and the generated skill models were equally valid as human-crafted skill models in terms of fit to students' learning log data. The lessons learned from the current study provide insights into the next generation of online courseware engineering and adaptive pedagogy driven by a skill model.

2. RELATED WORK

When a skill model is manually developed by subject matter experts, they often use cognitive task analysis, CTA (Clark et al., 2008), to identify skills by breaking down learning objectives of a course into more specific goals (Bier et al., 2014; Koedinger et al., 2010). Likewise, other evidence-based approaches, e.g., Evidence-centered assessment design, ECD (Mislevy, R. J., Almond, R. G., and Lukas, J. F. 2003), have been proposed. However, a proper application of CTA and ECD requires intensive training. Also, developing a skill model with CTA is a time-consuming, iterative process. These and other factors can combine to make CTA and ECD not an optimal solution for a scalable skill mining technique for large sets of curricula (Crandall, Klein, and Hoffman, 2006; Shute, Torreano, and Willis, 2000).

To overcome the issue of expensive human labor, researchers apply data mining techniques to automatically identify latent skills. The current literature shows three types of approaches for skill mining: (1) *Response Analysis*: This type of skill mining techniques analyzes students' response on assessment items to identify latent skills required to correctly answer them. (2) *Content*

Analysis: This type of techniques analyzes courseware contents (e.g., assessment items and didactic texts) to identify latent skills embraced in them. (3) *Model Refinement*: This type of techniques analyzes an existing skill model(s) to improve the "validity" of the model, usually measured as the accuracy of a model prediction. SMART is an example of the Content Analysis technique.

For the first approach, *Response Analysis*, the student response data must be collected from authentic students using existing online courseware. The response data are usually represented as a matrix, called a response matrix, that shows the correctness of individual students' answers on each individual assessment item. The response matrix, **R**, is then decomposed into two matrices, one showing each student's understanding of latent skills, **U**, and another showing the mapping between latent skills and assessment items, **Q**, such as $\mathbf{R} = \mathbf{U} \times \mathbf{Q}$. The latter matrix is called the Q-matrix (Tatsuoka, 1983). The Q-matrix is a binary matrix in which columns represent assessment items and rows represent latent skills. Ones ('1') in cells indicate that answering a corresponding assessment item requires a corresponding skill to be applied.

The task of mining latent skills based on the response matrix is therefore often reduced to finding a Q-matrix, which has been intensively studied. For example, U- and Q-matrices can be found by applying the matrix factorization technique that algebraically factorizes the response matrix (**R**). Desmarais (2012) used the Non-negative Matrix Factorization technique. Winters et al. (2005) conducted an exploratory study comparing various factorization techniques including Non-negative Matrix Factorization, Sigmoidal Factorization, and Common-Factor Analysis. Barnes (2010) applied the Q-Matrix Method that uses a hill-climbing technique to incrementally refine the model fit of a Q-matrix to the actual response data. Bayesian Networks have also been applied to compute a Q-matrix (Gonzalez-Brenes et al., 2012). In another approach, conjunctive and disjunctive models (DINA and DINO) were used to generate a Q-matrix based on an item response matrix and a matrix of the relationships between skills that was partially defined by experts (Wang et al., 2020).

Although methods dependent on student response data have been effective in the proper circumstances, *one of the weaknesses of Response Analysis is a lack of interpretability of the resulted skill model*. That is, the latent skills in a Q-matrix are implicit in its columns. It is hard to understand what each of the columns represents, which limits their practical application since the skills identified require input from an expert to be labeled. The reliance on student response data for the methods also implies that the online courseware needs to be used to collect the response data even without a skill model (or with a temporal skill model).

For the second approach, *Content Analysis*, some researchers apply machine learning techniques to classify course contents and identify associated skills. One such approach involved training a neural network to compute an answer to the assessment items for a course and consider a hidden layer just prior to the output layer of the model as the Q-matrix after binarizing the values (Chaplot et al., 2018). However, since the output is a Q-matrix, the lack of interpretability related to the meaning of the skills is a concern.

Other researchers aimed to generate labeled skills using the supervised-machine learning techniques. Haris and Omar (2012) extracted rules in the form of regular expressions from sets of assessment items labeled with the skill names and applied the rules to classify unseen assessment items. Supraja et al. (2017) implemented Support Vector Machine and Extreme Learning Machine techniques to automatically label assessment items. More recently, a classifier based on a Bidirectional Encoder Representations from Transformers (BERT) was used to assign skill labels to course content (Shen et al., 2021). *Although, those models yielded adequate results, providing*

training data for supervised learning requires solving the problem with initial labeling of skills. The methods are also likely to be domain-dependent since their performance is related to the domain of the training data. SMART, on the other hand, utilized unsupervised learning hence no labor-intensive analysis of courseware contents is required. SMART, in theory, is also domain independent.

For the third approach, *Model Refinement*, researchers have studied ways to refine existing skill models. For example, Difficulty Factors Assessment (Koedinger and Nathan, 2004) uses observation and analysis of the actions of students by experts to refine the required skills for tasks. Stamper and Koedinger (2011) then developed a hybrid human-machine discovery approach using learning curve analysis where experts identify improvements by observing characteristics of a model generated from student response data. As another example, Learning Factor Analysis (LFA) is a semi-automatic approach to improve an initially given skill model when additional knowledge is provided about the features (called P-Matrix) that differentiate assessment items that are otherwise associated to the same skill (Cen et al., 2006).

To overcome time-consuming feature engineering done by subject matter experts, Koedinger, McLaughlin, and Stamper (2012) further proposed a modification to LFA by taking previously developed skill models as a P-Matrix. This proposal works quite well when a rich collection of skill models is available such as those on DataShop (Koedinger et al., 2010) and its successor Learnsphere. Those open data sharing platforms allow researchers to run data analytics using built-in functionalities, given that adequate skill models are provided. Yet, since the process of creating the initial skill models for refinement is primarily driven by human experts, it becomes impractical to apply it to MOOCs due to scalability issues. It is therefore desired to develop automated skill mining techniques that provides human-readable labels.

The current work builds on our previous work on eEPIPHANY that is a combination of Response Analysis, Content Analysis, and Model Refinement (Matsuda et al., 2015). The goal of eEPIPHANY is to discover a Q-matrix either from a given response matrix or a set of assessment item text. In either case, eEPIPHANY first computes embeddings for assessment items, called an F-matrix. The F-matrix may be computed either by matrix factorization with the response matrix (Response Analysis) or by applying the Bag of Words technique to the assessment items (Content Analysis). The rows of F-matrix (that represent latent features) are then clustered where each cluster is considered as a skill, which becomes a *default skill model*. The default skill model will be refined (Model Refinement) by merging and splitting the skills so that the resulting clusters yield a better fit to the response matrix. eEPIPHANY requires human experts to interpret the discovered skill model by identifying the instance of refinement which received the most improvement.

To our knowledge, *SMART is the only domain-agnostic technique capable of mining, labeling, and mapping skills without the use of an existing skill model or student response data*. Not only does SMART eliminate an intensive requirement for expert input, but it can also produce an interpretable skill model prior to actual use of the courseware. Another unique feature of the SMART method is its capability to make associations between assessment items and instructions (i.e., didactic paragraphs). This association between skills and instructions has a significant utility for adaptive online courseware. For example, when a student fails to correctly answer an assessment item, the system can automatically show a link to a corresponding didactic text.

These two features —(1) developing a skill model prior to the actual courseware dissemination and (2) making the association between skills and instructions— are the most important advantages

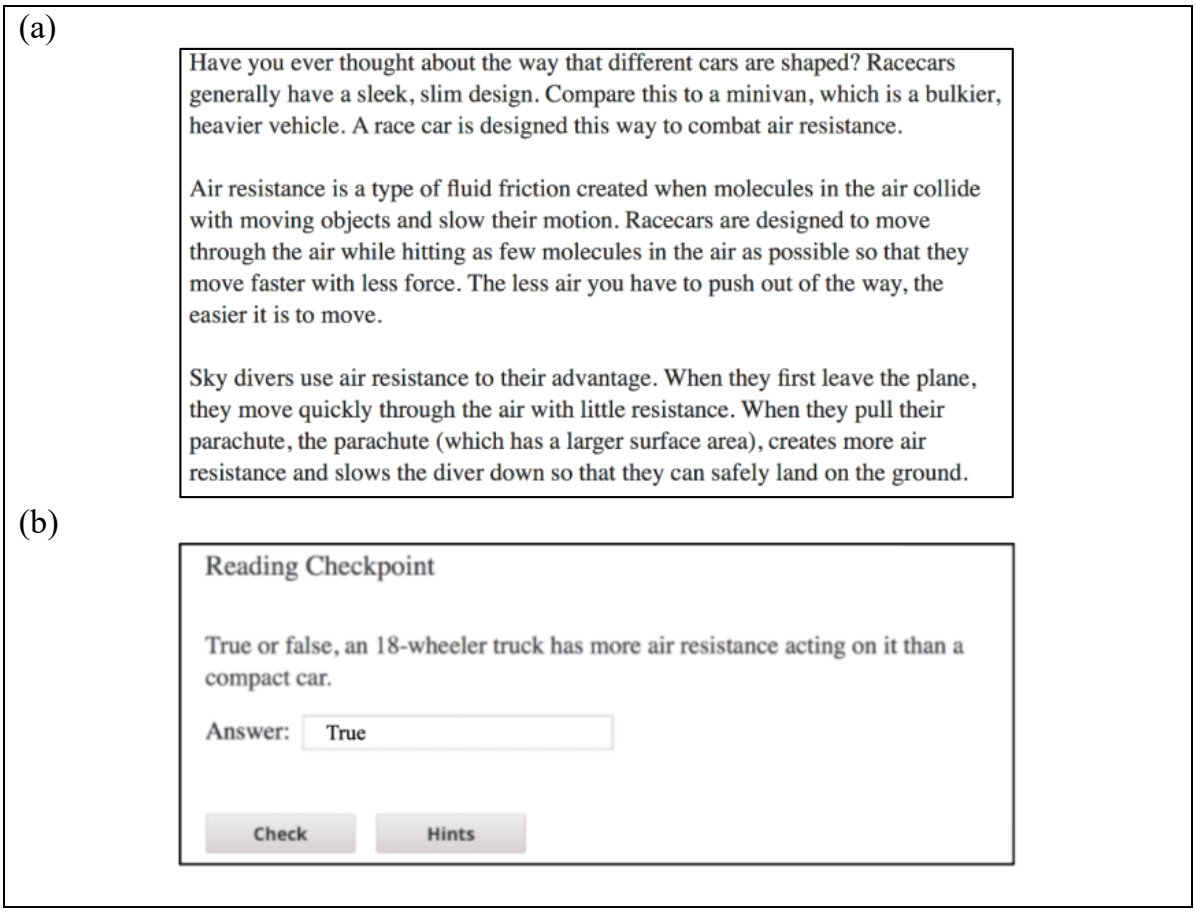

Figure 1. An example of (a) the paragraphs of instruction and (b) the assessment item extracted from existing courseware for middle school physics.

of the SMART method for our primary purpose, which is to develop pragmatic and scalable learning engineering methods to build adaptive online courseware.

3. TECHNICAL OVERVIEW OF SMART

Our primary research focus is on an automation of skill discovery from instructional text. The "*instructional text*" in the current study includes (1) written didactic paragraphs and (2) question sentences for assessment items as shown in Figure 1. The assessment items may include fill-in-the-blank questions (as shown in Figure 1), short answer questions, multiple-choice questions, and check list questions, etc. The assessment items might also include the correct choice item(s), feedback, and hint messages, wherever available.

Our central hypothesis is that when instructional texts are clustered based on their linguistic similarity, each cluster of text represents a unique latent knowledge component. We further assume that the most influential keyword extracted from the cluster of text will represent the latent knowledge component. Therefore, we implemented SMART as shown in Figure 2. We have integrated text-clustering and keyword-extraction techniques to identify latent skills from a collection of instructional texts and tag them with meaningful labels. The given instructional texts are clustered based on their linguistic similarity and each cluster is given a representative label.

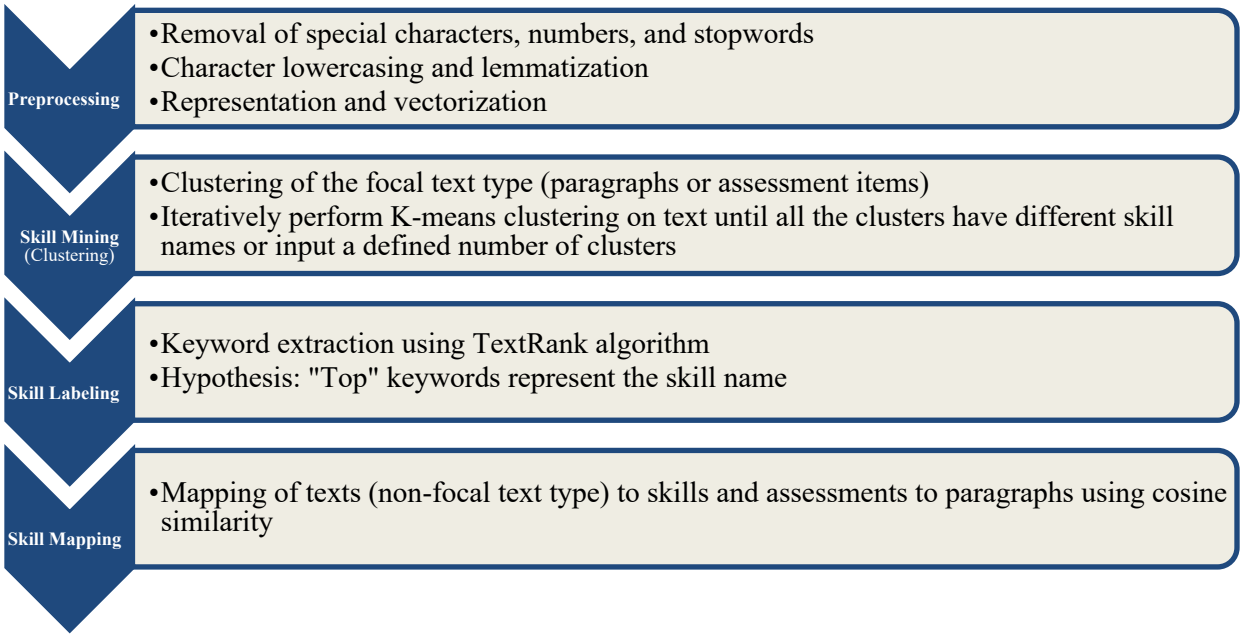

Figure 2. Multi-step process depicting the SMART method.

The rest of this section provides details of the SMART technology. With a lack of theoretical and empirical implications, we conducted an exploratory study to compare three hyperparameters to determine the optimal performance of SMART: (1) The *representation level* determines whether the embeddings of texts to be clustered are computed based on their linguistical characteristics or relationship among them (section 3.1). (2) The *number of clusters* determines a desired number of skills to be mined (section 3.2). (3) The *focal strategy* determines whether the latent skills should be extracted from assessment items or didactic paragraphs (section 3.2). A part of our research questions considers different combinations of these hyperparameters as discussed in section 4.1.

3.1. PREPROCESSING

The following preprocessing procedure is applied to individual sentences in the given instructional text (paragraph or assessment). A sentence text is first distilled by lowercasing and removing the special characters and numbers. Stop words are then removed from the distilled text prior to lemmatization.

The resulting text is then encoded into a vector representation based on the token frequency within the document using the Term Frequency-Inverse Document Frequency (TF-IDF) method (Salton and Buckley, 1988). The unit of analysis for text clustering is a paragraph or an assessment item. Consequently, TF-IDF is applied to didactic paragraphs and assessment items separately.

The *representation level* is the first hyperparameter for SMART. The TF-IDF vector representation of individual texts to be clustered are termed the first-level representation. The first level representation may be used for clustering the texts (as described below). However, the current literature indicates that the clustering based on the second-level representation sometimes results in a better classification accuracy (Rihák and Pelánek, 2017). The second-level text representation uses vectorizations that represent relative distance among the target text embeddings. Assume a set of first-level representations, $V_1 \in \mathbb{R}^{m \times d}$ where $|V_1| = m$ and d is the dimension of the TF-IDF

embeddings. The second-level representation, $V_2 \in \mathbb{R}^{m \times m}$, is computed such that the j-th element of the i-th vector in V_2 represents the cosine similarity distances (Salton and McGill, 1983) between the i-th and j-th ($0 < i, j < m$) embeddings in V_1.

3.2. SKILL MINING

Following the preprocessing step mentioned above, the resulting text embeddings (paragraphs or assessment items) are then clustered using the k-means technique (Hartigan and Wong, 1979). The k-means technique requires the number of clusters to be given a priori. Thus, the second hyperparameter for SMART is the *number of clusters* which specifies the variable k, which represents the number of skills.

Initially, we implemented an *iterative approach* to determine a sense-making k value, instead of providing the value of k in an ad-hoc fashion. It begins with a value of k specified by the number of unique instruction texts to be clustered divided by two (aiming an average of two items per cluster). SMART then performs the k-means clustering followed by skill labeling (described in the next subsection). Once all clusters are labeled, if two or more clusters have the same skill labels, they are assumed to be representing the same latent skill and, hence, get merged into one cluster. If any clusters are merged, then another iteration of k-means clustering takes place with the value of k equals to the number of unique clusters after merging (hence k is reduced at least by one). The iteration terminates when no duplicate cluster labels are found. Consequently, each of the final clusters represents a unique skill.

For further evaluation of determining the appropriate value of k, we applied a *user-specified approach* for identifying the number of clusters for SMART. The user-specified approach begins with a user specified value for k, representing the anticipated number of skills in the course. In contrast to the iterative approach, the user-specified approach applies k-means clustering only once before applying skill labeling and merging any clusters with the same skill label. Since clusters may be merged (at most once), the final number of clusters (unique skills) may be less than the value that the user specified originally.

The third hyperparameter for SMART, the *focal strategy* specifies the type of text on which the skill clusters are computed. There are two types of focus: *paragraph-based* and *assessment-based*. As mentioned above, the instructional text in the current study includes paragraphs and assessment items. The paragraph text is didactic whereas the assessment text is interrogational. With a lack of theoretical implications, there was no design justification on how this mixed type of text should be used. We, therefore, empirically compared two different focal strategies for skill mining. We presumed that clustering only one type of text (paragraph or assessment) to identify a set of skills first, and then mapping individual texts in another type to one of those identified skills (i.e., clusters of text) will result in a better skill model than clustering mixed text at once. We refer to the set of instructional texts used for initial clustering as the *focal text*. Accordingly, the set of instructional texts that are later mapped to the clusters is called the *non-focal text*.

In sum, once the focal text is clustered, each cluster is given a label that provides human-friendly interpretation of the latent skill among the clustered text. The individual items in the non-focal text are then mapped to one of the skill clusters using the cosine similarity measure. The hyperparameter focal strategy specifies whether the focal text is paragraph or assessment.

3.3. SKILL LABELING

A label for a cluster of text is computed using TextRank, a keyword extraction algorithm (Mihalcea and Tarau, 2004). We hypothesize that the keyword of the cluster represents the latent skill that the cluster of texts aims to convey.

The TextRank algorithm is an unsupervised graph-based ranking algorithm that takes words in a given document (the cluster of text in our case) as input and produces ranking of representative lemmatized words based on how closely they appear in the given document.

The "closeness" of a pair of words is computed as the frequency of co-occurrence within a pre-defined window. The "closeness" of the words is then reflected in the graph with nodes representing words and edges representing the frequency of the co-occurrence for the linked words. In the current study, the size of the window was set to be two based on empirical observations of the TextRank performance (Mihalcea and Tarau, 2004).

Once the graph of co-occurring words is formed, the TextRank algorithm computes a score of each node in the graph. In essence, the higher the number of links coming into a node, the higher the importance of the node, and hence the higher the score. The scores are then sorted in descending order to determine a rank associated with the nodes that represents the importance of the words within the graph. A list of the top n important words is retrieved and referred to as the keywords.

All the keywords that occur adjacent to each other with a particular order in the original text are collapsed into a single multi-word keyword, and their scores are averaged. For example, if the keywords "Skill", "Model", and "Mining" occur in a particular consecutive sequence (e.g., ".... skill model mining ...") in the original text (i.e., the text before removing stop words), then the three individual keywords are combined in the same order (e.g., "Skill Model Mining").

3.4. SKILL MAPPING

The following mapping procedure applies both for assessment-based and paragraph-based focal strategies. Let the *map base* be the focal text used for skill mining, i.e., a group of either assessment items or instructional paragraphs. Let the *mapping text* be the non-focal text. For example, if assessment items are the focal text, a cluster of assessment items is the map base whereas individual paragraphs are the mapping texts.

Skill mapping is done by computing the cosine similarity (Salton and McGill, 1983) between a mapping text and a map base. Individual clusters of text in the map base and individual sentences in the mapping text are first encoded using the TF-IDF representation. The cosine similarity between each sentence S_i in the mapping text and each cluster of text C_j in the map base is then computed using the Euclidean dot product. The cluster C_j in the map base with the highest cosine similarity is considered as a provider of a skill name for the sentence S_i in the mapping text. This way, didactic instructional paragraphs and assessment items are mapped to each other using the skill name as a key.

4. RESEARCH QUESTIONS

Our primary research questions are centered around the effectiveness of SMART, the proposed technique for latent skill discovery from courseware content. We group the research questions into two studies: (1) Hyperparameter Selection and (2) Comparison of SMART and Human Skill Models. The Hyperparameter Selection study explore the implementation and hyperparameter

choices associated with the approach. The Comparison of SMART and Human Skill Models study addresses the effectiveness of the approach for skill model generation by comparing the results with human skill models.

4.1. HYPERPARAMETER SELECTION

As a reminder, we consider three hyperparameters: representation level, number of clusters, and focal strategy. The first research question (RQ1), therefore, will analyze the different combinations of the hyperparameters for SMART by comparing the accuracy of resulting skill models predicting the student response on the existing student learning data.

(RQ1) Which set of hyperparameters for SMART are optimal for skill model generation?

As mentioned in section 3.2, while identifying skill clusters, SMART merges clusters with the same label into a single cluster. In the process of exploring the hyperparameter selection, we noticed that SMART often merged an unexpectedly large number of clusters—i.e., there were often too many clusters labeled with the same labels. Since, linguistically speaking, the merged clusters may represent different skills, we speculated that merging clusters with the same skill label degrades the validity of the resulting skill model. We therefore aim to address the following research question (RQ2) by evaluating the effect of merging clusters on the distribution of the number of assessment items per skill and the similarity of the resulting skill model with one made by human experts.

(RQ2) How does merging clusters impact the performance of SMART?

Poor clustering performance related to the loss of relevance for pairwise distance measures among high dimensional, sparse data is a common concern (Bansal and Sharma, 2021; Smieja et al., 2019; Zamora, 2017). Additionally, word embeddings that include contextual information have been shown to improve performance for certain natural language processing tasks compared to frequency-based representations such as TF-IDF. Therefore, we explore an additional research question (RQ3). RQ3 compares the model fit performance of SMART with the current sparse, frequency-based representation (TF-IDF) versus dense, contextual-based embeddings using Bidirectional Encoder Representations from Transformers (BERT) and Sentence-BERT.

(RQ3) What is the effect of using dense, context-based embeddings on the performance of SMART?

4.2. COMPARISON OF SMART AND HUMAN SKILL MODELS

We hypothesize that if a skill model discovered by SMART is valid, then the model should yield adequate prediction of students' responses on the existing student learning data that is equivalent to or better than human-crafted skill models. We also hypothesize that valid machine-generated skill models will have high similarity to human-crafted skill-models. The following research question (RQ4) is therefore explored:

(RQ4) Can SMART yield a skill model that is comparable to human-crafted skill models?

5. DATA AND METHODOLOGY

To address the research questions mentioned above, SMART was applied to two instances of existing online courseware. The machine-generated skill models were used to predict students' performance using the learning data collected by previous studies where students participated in the corresponding online courses. The learning data were obtained from DataShop, the open educational data repository (Koedinger et al., 2010). After selecting the optimal hyperparameters for SMART, we compared the best performing SMART-generated skill model with the best human-crafted skill model on each course (also obtained from DataShop) to measure the similarity among them.

5.1. DATA

Courseware content data and students learning data used in the current study were obtained from two existing online courses offered by the Open Learning Initiative (OLI) at Carnegie Mellon University (CMU), entitled "Introduction to Biology" and "General Chemistry I".

For the Biology course, a total of 1,095 assessment items and 1,608 paragraphs were parsed from the courseware. Students' learning data for this online course are available as "ALMAP spring 2014 DS 960 (Problem View fixed and Custom Field fixed)" dataset on DataShop. The learning data contain 268,822 observations, each of which shows a student's attempt on an assessment item including the correctness of the student's answer.

For the Chemistry course, a total of 1,505 assessment items and 2,838 paragraphs were parsed from the courseware. The student learning data are available as "AHA Chemistry 1 v2.3 Fall 2020" dataset on DataShop, which includes 200,327 observations.

5.2. METHODS

To address the research questions mentioned above, two evaluation studies have been conducted: (1) Hyperparameter selection, and (2) Comparison of SMART and human skill models.

5.2.1. Hyperparameter Selection Study

The goal of the first study, the Hyperparameter Selection Study (results shown in section 6.1), was to answer the first three research questions, RQ1-3. The first research question, RQ1, was investigated by computing the model fit for SMART-generated skill models across a range of values for the hyperparameters —representation level, focal strategy, and number of clusters. As a reminder of the hyperparameters, the representation level can be either first- or second-level. The focal strategy uses either paragraphs or assessment items as the focal text. For the number of clusters, the iterative approach can be applied, or we set the initial value of k to 10, 50, 100, 150, and 200 for the user-specified approach.

Some readers might question why the values of k were taken rather ad-hoc. While a more exhaustive exploration of the specified values of k is more rigorous, we chose to limit our study for two primary reasons: (1) Our primary motivation is to examine the trend in the quality of the skill model as the number of skills in a course (i.e., k) progresses into the lower range of pedagogically meaningful values. (2) Computing the AIC (Akaike Information Criterion) value (that was used to evaluate the quality of the model as described below) was very expensive hence the computational cost of the experiment needed to be reduced by limiting the scope of exploration.

To determine a best combination of hyperparameters, we compare the model fit of a student model based on the SMART-generated skill models to student learning data. We adopted the Additive Factor Model (AFM) implemented in DataShop as the basis for evaluating the model fit. The AFM is a logistic regression model with the probability of performing a problem step correctly on the first attempt as the dependent variable whereas the skill involved and the number of previous opportunities for that skill are independent variables. The individual student is also entered in the model as a random effect.

The validity of AFM is often measured using Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). AIC is commonly used for determining a model that best predicts observations (i.e., *predictive* model selection) hence lenient with the model complexity. BIC, on the other hand, is better suited for theorizing observations as a regression model (i.e., *explanatory* model selection) hence prefers simple models (Shmueli, 2010). Since we are most interested in generating a skill model that results in an accurate prediction of student learning, we used the AIC measure of the fit between the predictive model and actual student performance to evaluate the validity of the skill model.

Throughout the currently study, for a given skill model ***M***, the evaluation of the model fit to student learning data was performed according to the following steps:

(1) Associate a skill in ***M*** to each of the assessment items if ***M*** is computed by the paragraph-based focal strategy (see 3.2). This mapping is naturally done for the assessment-based focal strategy.
(2) Calculate the sequential opportunity count for each of the skill applications in the given student response data. Notice that for each opportunity of skill application, there might be multiple attempts with one and only one correct attempt at the end (unless students can answer the same assessment question correctly multiple times, which usually does not make sense).
(3) Fit the Additive Factor Model (AFM) using the correctness of the first attempt at each skill application opportunity as ground truth.
(4) Compute AIC that shows the fit between the predicted student responses according to the AFM and actual student responses.

Since the k-means clustering involves a stochastic component, SMART generates different skill models each time it runs even with a fixed set of hyperparameters. Therefore, for each combination of the hyperparameters, AIC values were averaged over 10 runs where each run consists of skill-model production and model fit. Notice that since AIC takes the model complexity into account, it is a common practice in the current literature to measure the model fit without cross validation. We fit the AFM model using all data.

An ANOVA test was then conducted with AIC as the dependent variable and the three hyperparameters mentioned above as independent variables to determine the optimal settings for representation level, focal strategy, and number of clusters. Since the evaluation of the model is expensive in terms of computation time, we chose to reduce the student learning data to the top 100 students according to number of responses while training the model. The sampling was done this way (instead of, say, random sampling) to maximize the coverage of skills in the response data sample used for the experiment.

To investigate the second research question, RQ2, the impact of merging clusters on the performance of SMART was examined in two ways. First, we examined the distribution of cluster

sizes for SMART-generated skill models with and without merging clusters. We assume that SMART-generated skill models should hold the skill distributions comparative with human-crafted ones. As far as the quality of the skill model is concerned, the one that is more similar to human-crafted skill model should be valued. We therefore computed the number of assessment items per skill cluster for each skill model to create an *empirical probability distribution of skill cluster size* for SMART-generated skill models (both with and without merging) and the human-crafted skill models. The similarity of the skill cluster size probability distributions for SMART-generated skill models (with and without merging) with the skill cluster size probability distribution for the human-crafted skill model was then computed using Kullback-Leibler divergence (KLD). The KLD value, $D_{KL}(P||Q)$, shows how the probability distribution Q differed from P (Kullback and Leibler, 1951). In our case, KLD tells us how the cluster size probability distribution for SMART-generated skill models differed from the cluster size probability distribution for human-crafted skill models.

Second, the Normalized Mutual Information (NMI) between SMART and the human-crafted model was measured. The NMI measures with and without the merging of clusters were then compared. The NMI is an information theoretic measurement commonly used to quantify the similarity of two clustering results on a scale between 0 and 1, where the similarity increases as the measure approaches 1 (Vinh et al., 2009). An ANOVA test was conducted with NMI as the dependent variable and the following variables as independent variables: (1) merge operation—on vs. off, (2) number of clusters (discrete values of 10, 50, and increments by 50 from 100 to 1000). Based on the results from the AIC comparison mentioned above, representation level and focal strategy were fixed at first-level and assessment focus, respectively. For each combination of independent variables, NMI values were averaged over 25 runs where each run consists of skill-model production and computation of the NMI between the SMART skill-model and the human-generated skill model.

The third research question, RQ3, was investigated by comparing the model fit performance of SMART using TF-IDF representation versus embeddings from Bidirectional Encoder Representations from Transformers (BERT) and Sentence-BERT. BERT embeddings are based on a pre-trained encoder on the masked language model and next sentence prediction tasks using a large corpus of text (Devlin, et al., 2019). Sentence-BERT embeddings are the result of fine-tuning BERT on a large collection of sentence pair data with a label showing their relationship, e.g., contradiction, entailment, or neutral (Reimers and Gurevych, 2019).

To avoid the requirement for course contents being labelled with a skill, we chose not to fine-tune the pretrained BERT and Sentence-BERT models for our task. For BERT, we used the *bert-base-uncased* pretrained model from the transformers python library, which has been pre-trained using a collection of unpublished books (BookCorpus) and Wikipedia. For Sentence-BERT, we used the *paraphrase-distilroberta-base-v2* pretrained model from the sentence-transformers python library, which was pretrained using sentence pair data from a variety of sources including StackExchange, Yahoo Answers, and Quora.

5.2.2. Comparison of SMART and Human Skill Models

The goal of the second study—comparison of SMART and human skill models (results shown in section 6.2) was to investigate the fourth research question RQ4. To achieve this goal, we compared the model fit of the best SMART-generated skill model against the best human-crafted skill model taken from DataShop. The best SMART-generated model was the one identified by the Hyperparameter Selection Study mentioned above. The AIC over AFM was used to measure the

model fit performance as described in section 5.2.1. To evaluate how the best SMART-generated model is similar to the human-crafted model, we applied the NMI measure as mentioned above.

Note that since SMART relies on the k-means clustering technique, an appropriate value for the number of clusters needs to be identified. In the current study, we applied the elbow method for determining the number of clusters. The elbow method uses measures intrinsic to the unsupervised clustering method to identify the value for the number of clusters at which the metric's improvement diminishes (Thorndike, R.L., 1953; Ketchen and Shook, 1996). We used the silhouette score that is a commonly used intrinsic measure for the elbow method. The measure considers the within-cluster dispersion and between-cluster dispersion to evaluate the quality of a clustering (Rousseeuw, 1987). The silhouette score improves as the within-cluster distances decrease and between-cluster distances increase, representing dense clusters that are relatively separated from other clusters. To enhance the rigor for this final comparison, we used the entire set of student learning data.

6. RESULTS

6.1. HYPERPARAMETER SELECTION

Regarding RQ1, Figure 3 and Figure 4 show the results of the model fit to students' learning data for the Additive Factor Model based on skill models identified using various combinations of the hyperparameters for SMART for the OLI Introduction to Biology and General Chemistry I courses, respectively. In the figures, each series is a combination of focal strategy and representation level, with the number of clusters as the x-axis and the average AIC from the model fit evaluation as the y-axis.

Regarding the number of clusters, the iterative approach clearly performed poorly compared to the user-specified number of clusters. Upon further review of the output for the iterative strategy, we noticed that our proposed skill labeling approach via TextRank frequently resulted in duplicate keywords for clusters. Consequently, the iterative strategy did not terminate until the number of clusters became very small, often less than 10, that resulted in poor performing skill models. Thus, we only include user-specified number of clusters in the following analysis.

To understand the impact of hyperparameters on the model fit, an ANOVA analysis was conducted with AIC as the dependent variable, representation level (first vs. second), focal strategy (assessment vs. paragraph), and number of clusters (10, 50, 100, 150, 200) as the independent variables, and course (Biology vs. Chemistry) as a random effect. The result did not show that representation level was a main effect; $F(1, 395) = 0.34$, $p = 0.56$. Therefore, we retain the simplest representation, first-level, for the following analysis.

The ANOVA found that focal strategy was a main effect; $F(1, 395) = 28.19$, $p < 0.001$. The assessment focal strategy yielded better skill models in terms of the model fit. The preference for assessment focus while clustering may be related to the use of the assessment item to skill mapping in the model fit evaluation. As a reminder, in the assessment strategy, the assessment items are directly clustered. On the other hand, in the paragraph strategy, the paragraphs are clustered and then the assessment items are mapped to the closest cluster of paragraphs. The indirect mapping of assessment items in the paragraph strategy is a potential source of error that may carry onto the skill model.

Finally, the ANOVA revealed that number of clusters was a main effect for the model fit; $F(1, 395) = 1610.46$, $p < 0.001$. The larger the number of clusters, the better the model fit. Specifically,

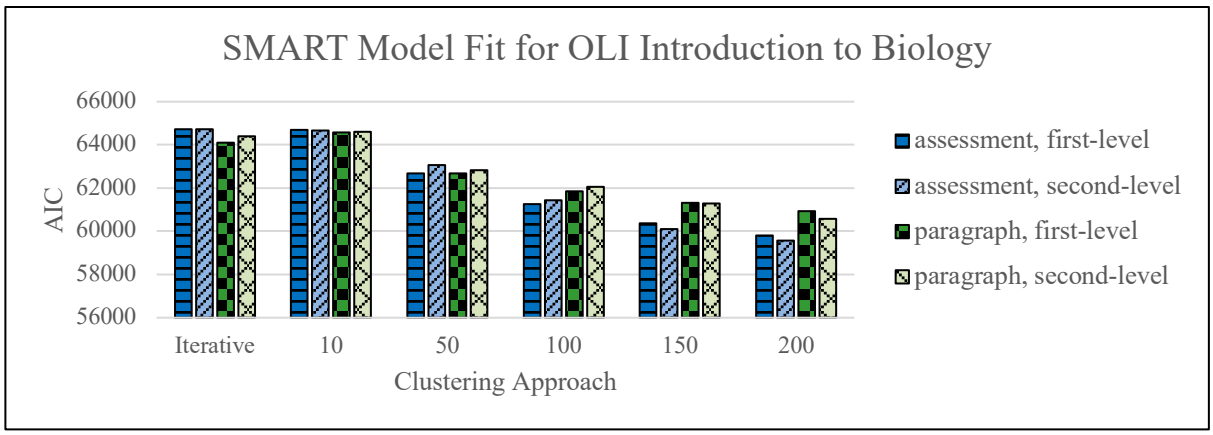

Figure 3. Results of the model fit for SMART on OLI Introduction to Biology across the range of hyperparameter values.

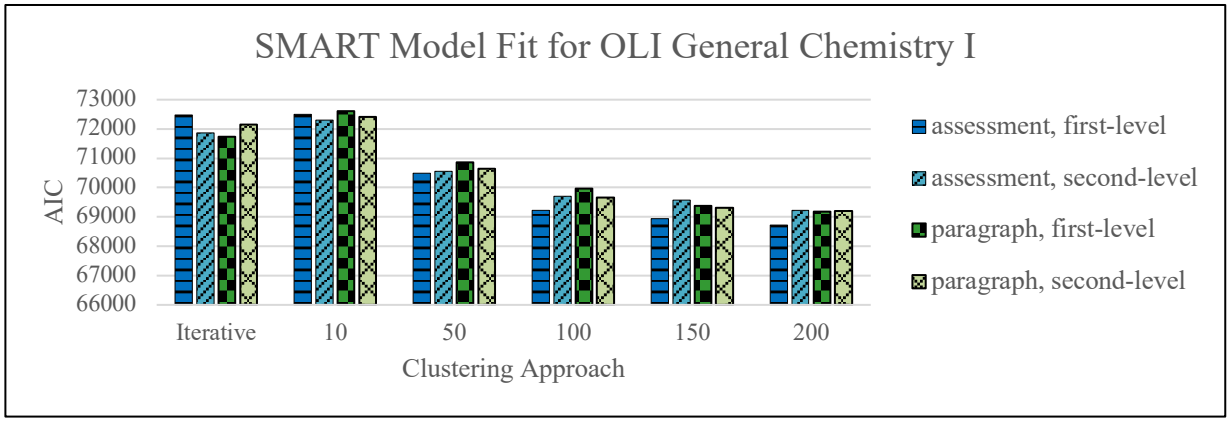

Figure 4. Results of the model fit for SMART on OLI General Chemistry I across the range of hyperparameter values.

as the number of clusters increased by one, the AIC value was improved by -19.49 (notice that the smaller the AIC value, the better the model fit).

Regarding RQ2, based on relatively poor performance of the iterative clustering strategy, we hypothesize that the duplicate keywords depict distinct skills under a common topic rather than replicated skills. For example, two clusters with the skill label of 'atom' may be related to unique skills related to the atom, such as identifying the structure of an atom or explaining their properties. To test this hypothesis, we developed a version of SMART that did not carry the merge operation hence created distinctive labels for clusters with identical skill labels from TextRank keyword extraction. For example, multiple clusters with TextRank keyword 'electron' become 'electron_1', 'electon_2', 'electron_3', and so on.

We then conducted an evaluation of the similarity between SMART-generated skill models with and without the merging operation and the human-crafted skill model to determine whether the merging of clusters had a statistically significant effect on the performance of SMART. We first compared the distribution of cluster sizes for the SMART-generated skill model with and without the merge operation and the human-crafted skill model using the KL Divergency measure as described in section 5.2.1. The results are shown in Table 1.

Table 1. Comparison of the Cluster Size Distribution between SMART with the merging of clusters, SMART without merging clusters, and the human-generated skill model.

Version of SMART	KL Divergence from Human Skill Model	
	OLI Introduction to Biology	*OLI General Chemistry 1*
Merge	0.1444	0.1121
No Merge	**0.1015**	**0.1021**

As shown in the table, SMART without merging clusters resulted in a cluster size distribution more similar to the human-crafted skill model than the one with the merging operation. *This result implies that the TextRank algorithm generates the same labels for clusters of assessments that human experts might consider as different skills*. Further research is needed to produce more distinctive labels than the current TextRank algorithm. We will elaborate this discussion later in section 7.2.

Regarding RQ2, we also compared the average Normalized Mutual Information (NMI) between the skill models generated by the versions of SMART and human experts (results in Figure 5). An ANOVA with NMI as the dependent variable, the status of merging clusters for SMART as the independent variable, and course and number of clusters as random effects revealed that the merging operation was a main effect for the normalized mutual information between SMART-generated and human-crafted models; $F(1, 2077)=2543.4$, $p < 0.001$. The skill models generated by SMART without the merging of clusters resulted in a 4% increase of NMI. Therefore, we selected the version of SMART without merging clusters for further comparison.

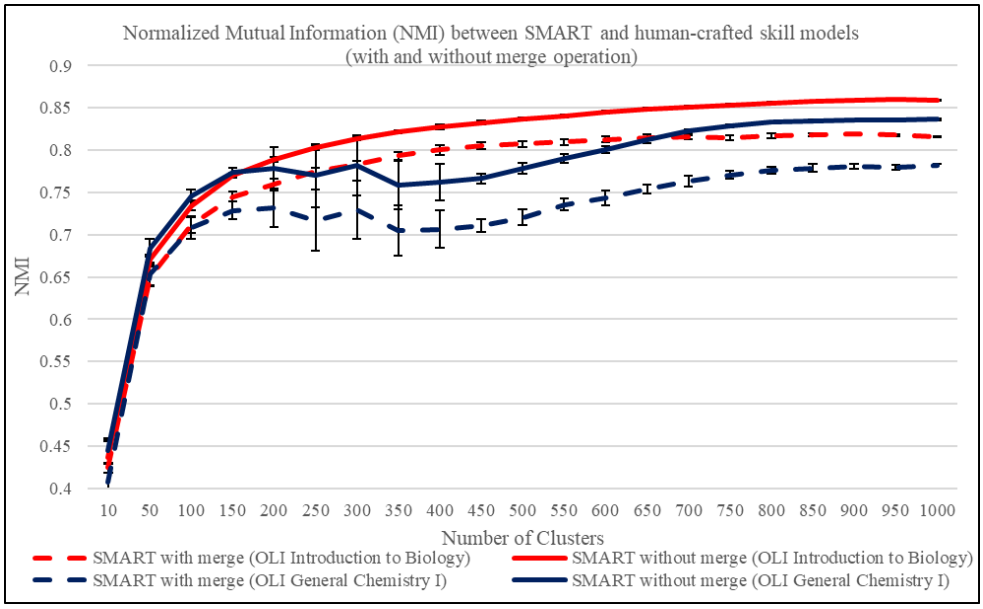

Figure 5. Normalized Mutual Information (NMI) between SMART skill models (with and without merging clusters) and the human-generated skill model for OLI Introduction to Biology and OLI General Chemistry I.

Regarding RQ3, to determine the effect of dense, contextual-based embeddings on the performance of our approach for skill model generation, we compared the model fit for student

learning using the Additive Factor Model based on skill models generated by SMART with different representation types (TF-IDF, BERT, and Sentence-BERT). The results are shown in Figure 6 and Figure 7 —the model fit by varying representation type for the OLI Introduction to Biology and OLI General Chemistry I courses, respectively. An ANOVA was conducted with AIC from the model fit as the dependent variable, representation, focal strategy, and number of clusters as the independent variables, and course as a random variable. The results showed that representation (TF-IDF, BERT, or Sentence-BERT) was a main effect; $F(1, 354) = 70.12$, $p < 0.001$.

To our surprise, both Sentence-BERT and BERT representation generated skill models with the model fit worse than TF-IDF —on average aggregated across focal strategy, number of clusters, and course, AIC for BERT and Sentence-BERT were 1011.3 and 351.7 points lower than TF-IDF, respectively. *These results suggest that either the decrease in clustering performance frequently observed when using high dimensional, sparse data does not apply to our application and/or contextual-based embeddings have no advantage over frequency-based representation when clustering a set of relatively short assessment items for a specific domain*. We have yet to understand why this occurs, but it is beyond the scope of the current study.

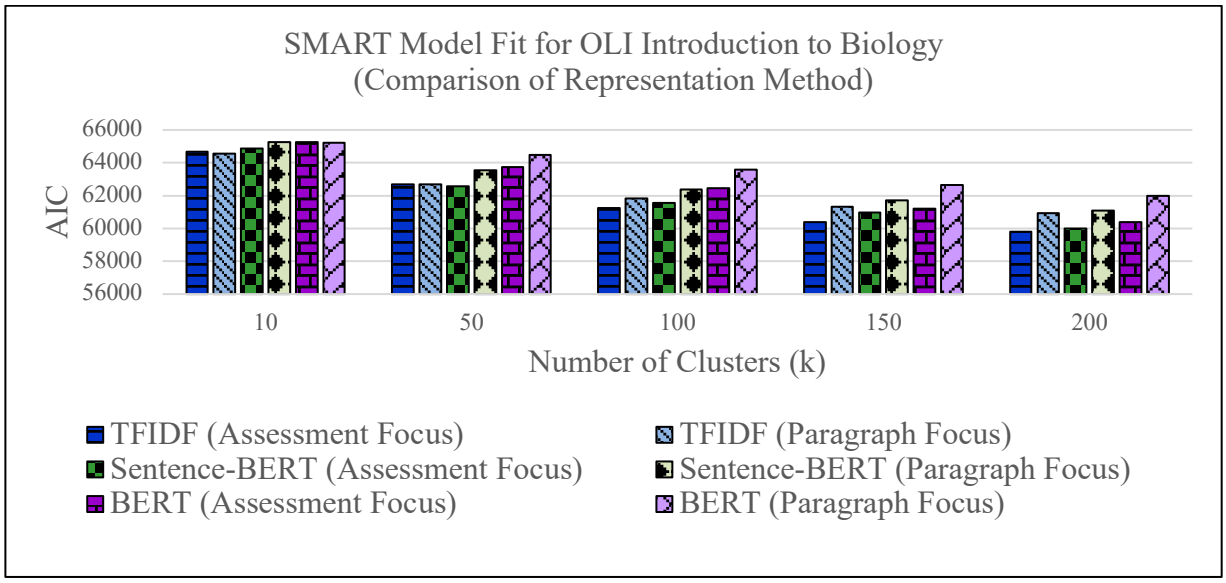

Figure 6. Comparison of the results of the model fit for SMART with varying representation methods on the OLI Introduction to Biology course.

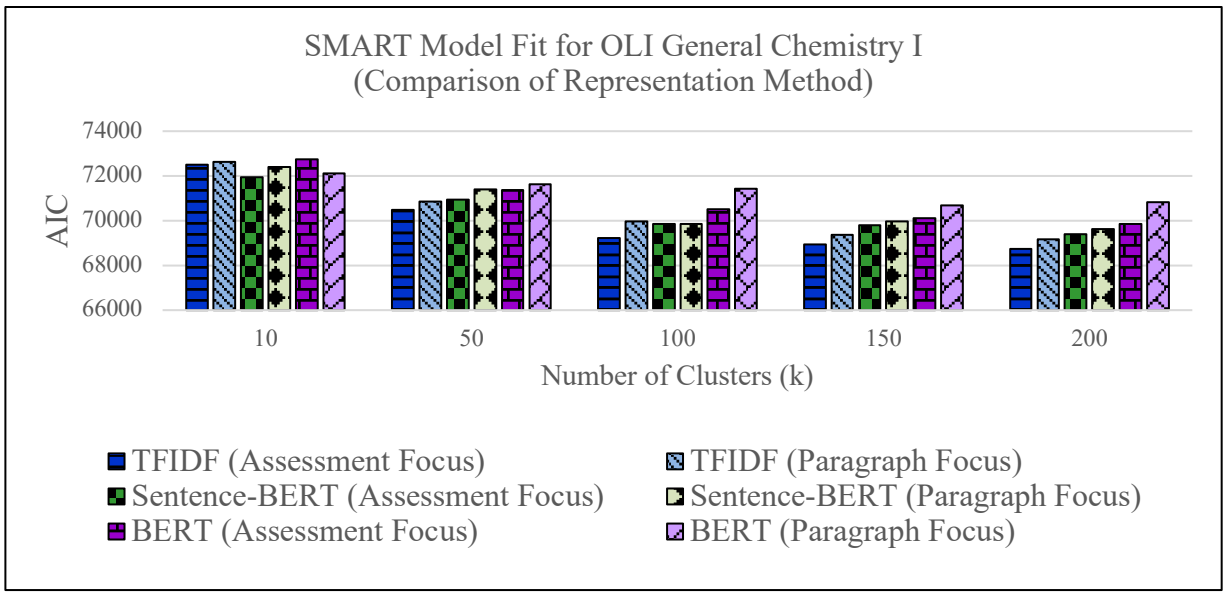

Figure 7. Comparison of the results of the model fit for SMART with varying representation methods on the OLI General Chemistry I course.

In sum, our evaluation of the best hyperparameter of SMART resulted in the selection of first-level TF-IDF representation, assessment items as the focal text, a user-specified number of clusters, and skill labeling without merging clusters. This combination of the hyperparameters was used for the comparison of SMART-generated and human-crafted skill models as described in the next section.

6.2. COMPARISON OF SMART AND HUMAN SKILL MODELS

As mentioned in the method section (5.2.2), the silhouette score was used to determine the number of clusters (i.e., a hyperparameter that specifies the target number of skills) in conjunction with the elbow method. The silhouette score values for the OLI Introduction to Biology and OLI General Chemistry 1 courses (as the number of clusters increase) are shown in Figure 8 and Figure 9, respectively.

Based on the elbow method, we determined the number of skills for the Biology course to be 325 and the number of skills for the Chemistry course to be 215 since the improvement of the silhouette score related to increasing the number of clusters diminishes beyond those values. We acknowledge that selection of the number of clusters using this method involves some subjectivity since it does not exactly pinpoint a value for the number of clusters.

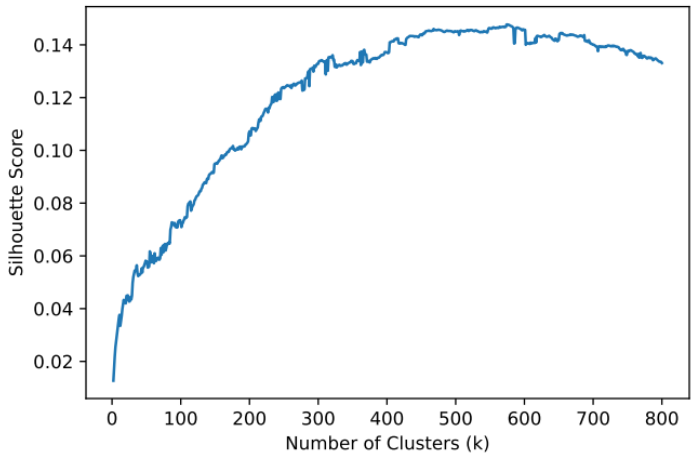

Figure 8. Measurement of the silhouette score for clustering assessment items as the number of clusters increases (OLI Introduction to Biology course).

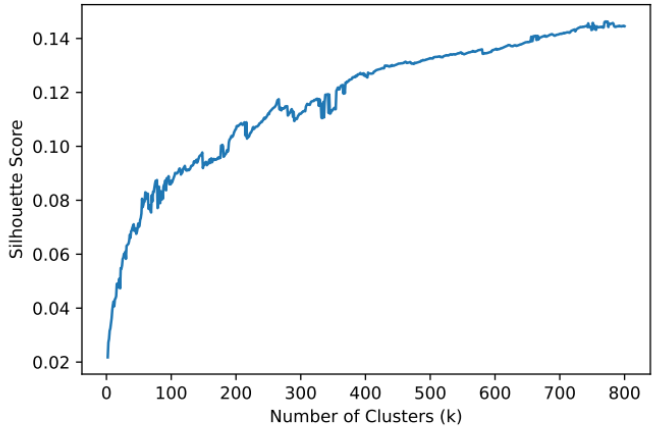

Figure 9. Measurement of the silhouette score for clustering assessment items as the number of clusters increases (OLI General Chemistry 1 course).

As noted in Section 5.2, both the SMART-generated skill model and the human-crafted skill model for each course were applied to student response data to build predictive models using AFM. The results of comparing the model fit in predicting student response for the OLI Introduction to Biology and OLI General Chemistry I courses are depicted in Table 2. Note that since not all students responded to all assessment items, the student response data taken from DataShop do not necessarily include all assessment items. Consequently, the number of skills discovered by SMART does not match the number of skills in the student response data. The results in Table 2 show that *for both courses, SMART was able to discover skill models that equally well or better predicted student response than the human-crafted skill models*.

Further, we evaluated the clustering similarity between the SMART-generated and the human-crafted skill models using Normalized Mutual Information (NMI) as shown in Table 3. In both cases, the SMART-generated skill model had a relatively high similarity with the human-crafted skill model for the entire courseware based on the NMI values. Although, there seems to be no common agreement on the goodness of NMI values, we would argue that an NMI of 0.8 shows a fair amount of agreement between human experts and SMART. The values we observed are comparable to NMI values achieved by top-performing clustering methods used for tasks such as

grouping health-related tweets and emails (Lossio-Ventura et al., 2021) and clustering texts such as news articles and Wikipedia entries (Sherkat et al., 2018).

Table 2. Model fit comparison between SMART and the human-crafted skill model for OLI Introduction to Biology and OLI General Chemistry 1 courses. The number in parenthesis represents the user-specified number of clusters for SMART on the entire courseware.

	OLI Introduction to Biology		OLI General Chemistry 1	
	SMART (325)	Human-crafted	SMART (215)	Human-crafted
# Skills identified and associated to student response data	257	186	187	185
AIC	**216,131.95**	217,561.55	**187,212.10**	187,571.83

Table 3. Clustering similarity between SMART and the human-generated skill models using Normalized Mutual Information (NMI).

Course	NMI between SMART and human-crafted skill models
OLI Introduction to Biology	0.816
OLI General Chemistry I	0.791

7. DISCUSSION

7.1. TEXT-MINING AS A METHOD FOR LATENT SKILL DISCOVERY

The current study demonstrated that even only using instructional text (i.e., instruction paragraphs and assessment items), latent skills covered on the courseware can be identified, labeled, and associated to individual written elements. The proposed model for the skill discovery, SMART, is the first in the current literature to automatically mine the *labeled* latent skills from authentic courseware contents. The current study also provided rigorous evaluations of the proposed method using existing authentic learning data.

The study also showed that the machine-generated skill models performed equally well or better than elaborated skill-models iteratively crafted by human subject-matter experts in terms of the fit to students' learning data. Admittedly, the larger the number of skills, the better the model fit. Indeed, SMART ended up identifying a larger number of skills than human experts to achieve a better performance. However, the number of skills reported in the current study is arguably still plausible.

The current version of SMART lacks a precise method for determining the number of clusters for the *k*-means clustering algorithm used to identify latent skills. Since the attempt to identify the number of clusters through an iterative approach performed poorly, we relied on a common method for estimating the value based on cluster analysis (the elbow method). Further study is required to more precisely select a pedagogically relevant number of skills for a given courseware content.

We plan to extend our study to include additional clustering algorithms and corresponding methods for inferring the number of clusters, e.g., DBSCAN with Gaussian Mixture Models, as well as exploring other methods for approximating the skill distributions made by experts.

The current study also showed that the skill model generated by SMART generally agreed with skill models defined by subject matter experts based on the Normalized Mutual Information between the models. A cursory review of the output of SMART, specifically the assessment item to skill mapping, confirmed that the clustering of assessment items appears logical. Examples from each course evaluated in the study can be found in Appendix A, Section A.1.

In sum, *SMART can generate a skill model that is comparable to the skill models proposed by subject matter experts given the knowledge about the appropriate estimation of the number of skills in the target courseware*. At a minimum, the machine-generated skill model from SMART can serve as an automated basis for skill-model refinement through existing methods, such as Learning Factor Analysis. Even as an initial skill model for further refinement based on expert input, SMART may significantly reduce the workload required of courseware developers.

7.2. KEYWORD EXTRACTION FOR SKILL LABELING

A unique practical strength of the SMART method is its capability to provide a supposedly instructor-friendly interpretation to the machine-discovered skills. While human readable skill labels are significant to the intended application of our approach, the current study was focused on evaluation of whether assessment items were meaningfully clustered based on semantic features. Therefore, we have yet to fully explore and evaluate the possible range of algorithms to extract or generate skill labels that have a utility for instructors and learners. Future research is needed to investigate how to best integrate courseware developers into the engineering pipeline.

For example, SMART might be used to validate and refine skill models used in the existing courseware. As a slight extension, SMART might be a part of analytic tools to optimize online learning by validating the size of an embedded skill model —assuming that SMART can actually make an adequate estimation of the number of skills related to the courseware. Finally, when SMART is fully mature, we may want to directly apply the resulting skill models in an authentic adaptive online environment and investigate the effect on learning.

The current study also shows some issues on skill labeling to be addressed by future research. First, we observed that SMART sometimes generated a skill label that did not discernibly represent all assessment items associated to a single cluster. Specific examples of this situation are included in Appendix A, Section A.2.

This suboptimal labeling behavior might be due to poor classification of the focal text, which results in mixing up texts for different skills. Alternatively, the poor discernability of the suggested skill labels might be due to a limitation of the TextRank algorithm for keyword extraction. Since the text from all assessment items in a single cluster is input into the TextRank algorithm, words that appear closely together in *many* of the assessment items, but not all, can be selected to summarize the entire set of assessment items. In this case, the behavior of the TextRank algorithm may produce a suboptimal skill label. In the current study, we observed approximately 10% of clusters with suboptimal labeling hence we would argue that SMART is still a valid technique with a pragmatic impact. Future research is necessary to produce text clusters with better (or "tighter" if you will) bond among the focal texts *and/or* to identify keywords that provide coverage of all texts in a given cluster.

Second, our skill labeling approach may generate skill labels that is too broad (or too "vague" if you will) to clearly identify or distinguish the skill required for the assessment items. For

example, SMART extracted seven ‘atom’s (‘atom_1’ through ‘atom_7’) and ‘atomic’ for the OLI Introduction to Biology course. Similarly, eight ‘electron’s were extracted for the OLI General Chemistry I course. It appeared that a specific keyword was identified for multiple different clusters of text while each should be associated with different skills. Future research is necessary to study alternative approaches for skill labeling to address the inability to distinguish between skills that are closely related.

Due to the identified shortcomings with using the TextRank algorithm for skill labeling, we plan to extend our evaluation to include additional keyphrase extraction algorithms, as well as supervised, abstractive techniques. For instance, more complex keyphrase extraction algorithms may produce a better result by considering longer phrases as keyphrase candidates that are more relevant and meaningful. Alternatively, sequence models, such as GPT-3, may be able to successfully generate a skill label given the text for a group of assessment items. However, the approach might also require formation of a broad set of training examples from courses across several domains to fine-tune the sequence model.

Once we achieve adequate performance for skill labeling, another research direction involves the extrinsic evaluation of the validity of the automatically generated skill labels. A survey of experts in course development, for example, would better assess the utility of our automated skill labels for application by instructors in managing an adaptive online course when compared to manually formed skill labels.

8. LIMITATIONS

We acknowledge the limitations of the current study that must be further investigated. First, we lack empirical evidence to support the generality of our approach. The current study explored the effect of hyperparameter settings on a subset of the data used for evaluation and was applied to only two STEM courses. Therefore, unforeseen factors may influence the performance of our approach for skill mining and labeling when applied to a diverse set of courses.

For example, our proposed approach is likely to work well for courses that depend on verbose, comprehension-type tasks since the assessment items are clustered according to term frequency-based features. However, the approach may not be as relevant for courses with a high sensitivity to syntactic differences (e.g., programming). Additionally, decisions regarding pre-processing of the text, such as the removal of numbers and specific stopwords, may degrade performance for courses where numbers are prevalent and highly relevant to the learning objectives (i.e. math). In the current study, we have studied two STEM courses from different domains. We will explore more divergent, non-STEM courses, to study domain specific and domain general features that influence skill mining, if any.

Second, regarding the evaluation of the machine-generated skill models during hyperparameter selection, we have reduced the student learning data to only include the top 100 students based on their number of responses to reduce the computational cost. However, this may have inadvertent side effects. For example, it is unknown how sampling different populations of students —e.g., “average” vs. “low” performers— would impact the result. Further study is necessary to validate the impact of the selection bias, if any.

9. CONCLUSION

We found that clustering instructional text based on semantic features resulted in a skill model with predictive power for students' learning comparable to the efforts of the subject matter experts. We have also found that the machine-generated skill models generally agree with the human-crafted skill models in terms of assessment-skill association. At the same time, the current observation of the results indicates that using the TextRank algorithm to extract skill labels sometimes fails to differentiate closely related skills as resulting in having multiple skill clusters with the same label. In the current evaluation study, the machine-generated skill models were compared to the human-crafted skill models for two existing authentic online courses.

In summary, SMART is a method for automatically mining, labeling, and mapping skills for existing online courseware using existing machine learning technologies. It is aimed to apply to authentic large-scale online courseware, aka, Massive Open Online Course (MOOC). With the rapidly growing interest in effective MOOCs, studying assistive technologies to develop valid adaptive instruction is of utmost importance and critical research for the next generation of online education. It is our belief that the current work has provided a significant contribution toward automating the generation of scaffolding for adaptive online content with potential for increasing the reliability, hence pedagogical value, of online education.

In its current state, SMART produces an initial skill model for refinement by experts, such as course developers and instructors. Should the aim of the approach be fully achieved, automated skill models produced by SMART may have the potential to reduce the workload of course developers while enabling adaptive online content at the launch of a course. As student response data becomes available, then existing methods for model refinement —such as Learning Factor Analysis— may be deployed to further improve the validity of the SMART-generated skill model. Through integration into open data sharing platforms, such as DataShop and its successor Learnsphere, researchers can run data analytics using built-in functionalities with the skill models automatically generated by SMART. Yet, we believe that SMART is an important step towards our goal to automate the generation of skill models with instructor-friendly labels from courseware content without the use of an existing skill model or student learning data.

The entire code base for the algorithms reported in the current paper is available on GitHub[1].

[1] https://github.com/IEClab-NCSU/SMART

APPENDIX

A. ASSESSMENT ITEM TO SKILL MAPPINGS FROM SMART

In this appendix, we provide a sample of the assessment item to skill mappings generated by SMART. In the first subsection, we present examples where the grouping of assessment items appears logical at face value. Next, we present examples where the skill label does not discernably apply to all assessment items assigned to the skill.

A.1. EXAMPLES OF ASSESSMENT ITEM TO SKILL MAPPINGS

From the OLI Introduction to Biology courseware:

Skill Name	Assessment Items (Clustered by Skill)
phosphate	How are ATP and ADP structurally different? ATP has three phosphates attached to a single adenosine, while ADP has two phosphates attached to an adenosine, Correct. What does the T in ATP stand for? What about the D in ADP? D stands for diphosphate (two) and T stands for triphosphate (three).
	The products of ATP hydrolysis, ADP and inorganic phosphate, are more stable because: Electrostatic repulsion between the phosphates is relieved. Correct.
	When ADP is converted to ATP, the new phosphate is added to which position on the molecule? a. Correct
	ATP is formed from ADP by the addition of _ to ADP _ phosphate. Correct.
protein	Moving single molecules across the membrane using a transport protein involves: uniport. Correct.
	Which of the following can cross the cell's plasma membrane without the assistance of protein transporters? Select all that apply. gasses small hydrophobic molecules Correct. Gases are small enough to cross the membrane without assistance. The inner portion of the membrane is hydrophobic, so hydrophobic molecules also can cross on their own provided they are small enough.
	Which component of the plasma membrane has the widest variety of functions? proteins. Correct. Proteins help transport material, are involved in metabolism and adhesion, and play a number of other roles.
	In which of the following organisms is it necessary to transport the mRNA across a membrane prior to protein synthesis? Select all that apply. Animals Plants Correct.
	Molecules that can not cross the phospholipid bilayer on their own are transported across the cell membrane using proteins. Which of these molecules would NOT need to be transported across the membrane using a protein? carbon dioxide Correct. Carbon dioxide and other gasses can cross the membrane on their own and do not require transport using a protein.
	Oxygen is transported from the lungs to the tissues by dissolving directly in blood. false Correct. Oxygen binds to the protein hemoglobin, which is found in red blood cells. Proteins perform most of the biological functions in organisms. Proteins are involved in oxygen transport.

glycolysis	When ATP levels are high, do you expect glycolysis to be operating? No Correct. High levels of ATP act like a stoplight to glycolysis. There is no need for more ATP, so glycolysis stops. Glycolysis produces ATP. If glycolysis produces ATP, should it be operating when the cell has plenty of ATP? Which of the images above would best represent the pathway when ATP levels are low? A. Correct. If ATP levels are low, more ATP is needed. This image shows the green light, meaning that glycolysis will proceed, and more ATP will be produced. If ATP levels are low, the cell has to generate ATP using the energy in glucose. Glucose has to proceed through the entire pathway to pyruvate in order to produce ATP. Glycolysis does not require O2 to generate: energy Correct. Glycolysis can (and often does) occur in the absence of oxygen. Glycolysis takes place in: cytosol Correct. In addition to ATP, which of the following are end products of aerobic glycolysis? pyruvate and NADH Correct. Phosphofructokinase, the major flux-controlling enzyme of glycolysis is allosterically inhibited by _ and activated by _. ATP, AMP Correct, Since glycolysis generates ATP, if [ATP] is high, less carbon is sent through the pathway by inhibiting PFK; the converse is true for high levels of AMP, which is produced by hydrolysis of ATP. Look through all of the information on this page.

From the OLI General Chemistry I courseware:

Skill Name	**Assessment Items (Clustered by Skill)**
energy	The skater has the greatest kinetic energy _? Correct. The skater will have the greatest velocity at the bottom of the track. Which of the following forms of electromagnetic energy has the greatest energy? X rays Correct. Energy is directly proportional to frequency. The greater the frequency the greater the energy. Which of the following compounds would have the greatest lattice energy? NaF Correct. With ionic charges being equal, compounds with smaller distances between ions have higher lattice energy. All of the compounds contain the same ionic charges, +1 and -1. With ionic charges being equal, compounds with smaller distances between ions have _ lattice energy? greater Correct. Which of the following compounds would have the greatest lattice energy? MgI2 Correct. With magnitude of the ionic charges being +2 and -1 and this compound has a smaller distances between its ions, this compound will have the greatest lattice energy. The magnitude of charge of the ions in the compound has a _ impact on lattice energy than the distance between the ions? greater Correct. Look for the compound with the greatest magnitude of charge. With the magnitude of charge being equal, look for the compound with the smallest distance between the ions. Which of the following compounds would have the lowest lattice energy? NaBr Correct. With a magnitude of charge of +1 and -1 and larger distance between its ions than NaCl, NaBr will have the lowest lattice energy. Place the following in order from least repulsive force to greatest repulsive force.

A.2. EXAMPLES OF SUBOPTIMAL SKILL LABELLING

From the OLI Introduction to Biology courseware:

Skill Name	Assessment Items (Clustered by Skill)
equal concentration	A solution has an equal concentration of H+ and OH-. This solution is probably: neutral Correct. Neutral substances have equal concentrations of H+ and OH-. Acids have a higher concentration of H+ than OH-. Bases have a lower concentration of H+ than OH-. Neutral substances have equal concentrations of H+ and OH-.
	A solution has a *higher* concentration of H+ than OH-. This solution is probably: acidic Correct. Acids have a higher concentration of H+ than OH-. Acids have a higher concentration of H+ than OH-. Bases have a lower concentration of H+ than OH-. Neutral substances have equal concentrations of H+ and OH-.
	A solution has a *lower* concentration of H+ than OH-. This solution is probably: basic Correct. Bases have a lower concentration of H+ than OH-. Acids have a higher concentration of H+ than OH-. Bases have a lower concentration of H+ than OH-. Neutral substances have equal concentrations of H+ and OH-.
bacterial cell membrane	*Require(s) DNA polymerases that are stable at higher temperatures. PCR Correct. PCR includes a step at high temperatures, which requires special DNA polymerases.*
	A typical PCR cycle consists of steps at different temperatures, typically 55, 78 and 98o C. The correct order of these steps is? 98, 55, 78 Correct. First you denature the DNA at 98o C and then let the primers anneal at 55o.
	What would you expect to happen to a bacterial cell membrane under high temperatures? It would probably get looser and break apart. Correct. Think about how lipids melt at high temperatures.

From the OLI General Chemistry I courseware:

Skill Name	Assessment Items (Clustered by Skill)
chemical formula	What is the chemical formula for copper(I) carbonate?
	Enter the chemical formula for copper(II) iodide.
	Which of the following statements is true of the chemical equation below? $Cu(s) + 2\ AgNO_3(aq) \rightarrow 2\ Ag(s) + Cu(NO_3)_2(aq)$
	Copper undergoes oxidation when placed in a solution of silver nitrate. If 6.2 g of copper is placed into 50.0 mL of a 2.5 M AgNO3 solution, which is the limiting reactant?
	What is the concentration of Cu(NO3)2 when the reaction is complete? (Assume the change in volume from the added Cu(s) is negligible and can be ignored.)

ACKNOWLEDGEMENTS

The research reported here was supported by National Science Foundation Grant No. 1623702 and No. 2016966 to North Carolina State University.

REFERENCES

BANSAL, M., AND SHARMA, D. 2021. A novel multi-view clustering approach via proximity-based factorization targeting structural maintenance and sparsity challenges for text and image categorization. *Information Processing & Management*, 58(4), Elsevier, 102546.

BARNES, T. 2010. Novel derivation and application of skill matrices: The q-matrix method. In *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizkiy and R. S. J. d. Baker, Eds. CRC Press, Boca Raton, FL, 159-172.

BIER, N., AND RINDERLE, J. 2011. Openness and Learning Analytics. *Open Education Annual Conference*, Park City, UT. Routledge.

BIER, N., STRADER, R., AND ZIMMARO, D. 2014. An approach to Skill Mapping in Online Courses. *Learning with MOOCs*, Cambridge, MA.

BIER, N., MOORE, S., AND VAN VELSEN, M. 2019. Instrumenting courseware and leveraging data with the Open Learning Initiative (OLI). In *Companion Proceedings 9th International Learning Analytics & Knowledge Conference*, J. Cunningham, N. Hoover, S. Hsiao, G. Lynch, K. McCarthy, C. Brooks, R. Ferguson, and U. Hoppe, Eds. Tempe, AZ, 990-1001.

CEN, H., KOEDINGER, K., AND JUNKER, B. 2006. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, M. Ideka, K.D. Ashley, and T.W. Chan, Eds. 4053, Springer, Berlin, 164–175. DOI: https://doi.org/10.1007/11774303_17

CHAPLOT, D. S., MACLELLAN, C., SALAKHUTDINOV, R., AND KOEDINGER, K. 2018. Learning Cognitive Models Using Neural Networks. In *Proceedings of International Conference on Artificial Intelligence in Education*, C. Penstein Rosé, R. Martínez-Maldonado, U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, and B. du Boulay, Eds. Vol 10947, Springer, Cham, 43-56.

CHEN, Y., LI, X., LIU, J., AND YING, Z. 2018. Recommendation system for adaptive learning. *Applied Psychological Measurement*, 42(1), Sage Publications, 24-41.

CLARK, R., FELDON, D., VAN MERRIENBOER, J. J. G., YATES, K., AND EARLY, S. 2008. Cognitive task analysis. In *Handbook of Research on Educational Communications and Technology*, J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, and M. P. Driscoll, Eds. Macmillan/Gale, New York, NY, Routledge, 577–593.

CRANDALL B, KLEIN G, HOFFMAN RR. 2006. Working Minds: A Practitioner's Guide To Cognitive Task Analysis. MIT Press, Cambridge, MA.

DAI, Y., ASANO, Y., YOSHIKAWA, M. 2016. Course Content Analysis: An Initiative Step toward Learning Object Recommendation Systems for MOOC Learners. In *9th Proceedings of International Conference on Educational Data Mining*, T. Barnes, M. Chi, and M. Feng, Eds. International Educational Data Mining Society, 347–52.

DESMARAIS, M. C. 2012. Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter*, 13(2), ACM, 30–36. DOI: https://doi.org/10.1145/2207243.2207248

DESMARAIS, M. C., AND BAKER, R. S. J. D. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1–2), Springer, 9–38. DOI: https://doi.org/10.1007/s11257-011-9106-8

DESMARAIS, M. C., AND NACEUR, R. 2013. A Matrix Factorization Method for Mapping Items to Skills and for Enhancing Expert-Based Q-Matrices. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, 7926, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-39112-5_45

DEVLIN J, CHANG MW, LEE K, TOUTANOVA K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 4171-4186.

GAVRILOVIĆ, N., ARSIĆ, A., DOMAZET, D., AND MISHRA, A. 2018. Algorithm for adaptive learning process and improving learners' skills in java programming language. *Computer Applications in Engineering Education*, 26(5), Wiley Online Library, 1362-1382.

GONZALEZ-BRENES, J. P., AND MOSTOW, J. 2012. Dynamic Cognitive Tracing: Towards Unified Discovery of Student and Cognitive Models. In *Proceedings of the 5th International Conference on Educational Data Mining*, K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, and J. Stamper Eds. International Educational Data Mining Society, 49-56.

HARIS, S. S., AND OMAR, N. 2012. A rule-based approach in Bloom's Taxonomy question classification through natural language processing. In *2012 7th International Conference on Computing and Convergence Technology (ICCCT)*, K. D. Kwack, S. Kawata, S, Hwang, D. Han, and F. Ko, Eds. IEEE, 410–414.

HARTIGAN, J. A., AND WONG, M. A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), Wiley, 100–108. DOI: https://doi.org/10.2307/2346830

IMHOF, C., BERGAMIN, P., AND MCGARRITY, S. 2020. Implementation of Adaptive Learning Systems: Current State and Potential. In *Online Teaching and Learning in Higher Education*, P. Isaias, D. G. Sampson, and D. Ifenthaler, Eds. Springer, Cham, 93–115. DOI: https://doi.org/10.1007/978-3-030-48190-2_6

JOVANOVIC, D., AND JOVANOVIC, S. 2015. An adaptive e-learning system for java programming course, based on Dokeos LE. *Computer Applications in Engineering Education*, 23(3), Wiley Online Library, 337-343.

KETCHEN, D., AND SHOOK, C. 1996. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17(6), Wiley Online Library, 441-458.

KOEDINGER, K. R., BAKER, R., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., AND STAMPER, J. 2010. A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechenizkiy and R. S. J. d. Baker, Eds. CRC Press, Boca Raton, FL.

KOEDINGER, K. R., CORBETT, A. T., AND PERFETTI, C. 2012. The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, 36(5), Wiley Online Library, 757–798.

KOEDINGER, K. R., MCLAUGHLIN, E. A., AND STAMPER, J. C. 2012. Automated Student Model Improvement. In *Proceedings of the 5th International Conference on Educational Data Mining*,

K. Yacef, O. Zaïane, A. Hershkovitz, M. Yudelson, and J. Stamper Eds. International Educational Data Mining Society, 383-395

KOEDINGER, K. R., AND NATHAN, M. J. 2004. The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences*, 13(2), Taylor & Francis, 129-164.

KULLBACK, S., AND LEIBLER, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), Institute of Mathematical Statistics, 79-86.

LIU, M., MCKELROY, E., CORLISS, S. B., AND CARRIGAN, J. 2017. Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development*, 65(6), Springer, 1605-1625.

LOSSIO-VENTURA, J. A., GONZALES, S., MORZAN, J., ALATRISTA-SALAS, H., HERNANDEZ-BOUSSARD, T., AND BIAN, J. 2021. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*, Elsevier, 117.

MARTIN, B., MITROVIC, T., MATHAN, S., AND KOEDINGER, K. R. (2005). On using learning curves to evaluate ITS: Automatic and semi-automatic skill coding with a view towards supporting on-line assessment. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education* C. K. Looi, G. McCalla, B. Bredeweg, and J. Breuker, Eds. Springer, Cham, 419-426.

MARTIN, B., MITROVIC, A., KOEDINGER, K. R., AND MATHAN, S. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, 21(3), Springer. 249-283. doi: 10.1007/s11257-010-9084-2

MATSUDA, N., FURUKAWA, T., BIER, N., AND FALOUTSOS, C. 2015. Machine Beats Experts: Automatic Discovery of Skill Models for Data-Driven Online Course Refinement. In *Proceedings of the 8th International Conference on Educational Data Mining*, O.C. Santos, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, Eds. International Educational Data Mining Society. 101-108.

MATSUDA, N., SHIMMEI, M., CHAUDHURI, P., MAKAM, D., SHRIVASTAVA, R., WOOD, J., AND TANEJA, P. (in press). PASTEL: Evidence-based learning engineering methods to facilitate creation of adaptive online courseware. In *Artificial Intelligence in STEM Education: The Paradigmatic Shifts in Research, Education, and Technology*. F. Ouyang, P. Jiao, B. M. McLaren and A. H. Alavi, Eds. New York, NY: CSC Press, 1-16.

MIHALCEA, R., AND TARAU, P. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 404–411.

MISLEVY, R. J., ALMOND, R. G., AND LUKAS, J. F. (2003). A Brief Introduction to Evidence-centered Design. *ETS Research Report Series, 2003*(1), Wiley Online Library, 1-29. DOI: https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

PAQUETTE, G., MARIÑO, O., ROGOZAN, D., AND LÉONARD, M. 2015. Competency-based personalization for massive online learning. *Smart Learning Environments*, 2(1), Springer, 4. DOI: https://doi.org/10.1186/s40561-015-0013-z

PELÁNEK, R. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, 27(3), Springer, 313-350.

REIMERS, N., GUREVYCH, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 3973-3983.

RIHÁK, J., AND PELÁNEK, R. 2017. Measuring Similarity of Educational Items Using Data on Learners' Performance. In *Proceedings of the 10th International Conference on Educational Data Mining*, X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, Eds. International Educational Data Mining Society. 16-23.

ROUSSEEUW, P. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, Elsevier, 53-65.

SALTON, G., AND BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), Elsevier, 513–523. DOI: https://doi.org/10.1016/0306-4573(88)90021-0

SALTON, G., AND MCGILL, M. J. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York, NY.

SHEN, J. T., YAMASHITA, M., PRIHAR, E., HEFFERNAN, N., WU, X., MCGREW, S., AND LEE, D. 2021. Classifying math knowledge components via task-adaptive pre-trained BERT. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education*, I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, Eds. Springer, Cham, 408-419.

SHERKAT, E., VELCIN, J., AND MILIOS, E. E. 2018. Fast and Simple Deterministic Seeding of KMeans for Text Document Clustering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction,* P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, Eds. Springer, Berlin, Heidelbert, 76–88.

SHMUELI, G. 2010. To Explain or to Predict? *Statistical Science*, 25(3), Institute of Mathematical Statistics, 289–310.

SHUTE, V. J., TORREANO, L. A., AND WILLIS, R. E. 2000. DNA: Providing the blueprint for instruction. In *Cognitive Task Analysis*, Psychology Press, 85-100.

ŚMIEJA, M., HAJTO, K. & TABOR, J. 2019. Efficient mixture model for clustering of sparse high dimensional binary data. *Data Mining and Knowledge Discovery*, 33, Springer, 1583-1624.

STAMPER, J., AND KOEDINGER, K. 2011. Human-machine student model discovery and improvement using data. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Springer, Berlin, Heidelberg, 353–360.

SUPRAJA, S., HARTMAN, K., TATINATI, S., AND KHONG, A. W. H. 2017. Toward the Automatic Labeling of Course Questions for Ensuring Their Alignment with Learning Outcomes. In *Proceedings of the 10th International Conference on Educational Data Mining*, X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, Eds. International Educational Data Mining Society. 56-63.

TATSUOKA, K. K. 1983. Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement*, 20(4), JSTOR, 345–354.

THORNDIKE, R. L. 1953. Who belongs in the family? *Psychometrika*, 18(4), 267-276.

TYTON PARTNERS. 2020. Time for Class 2020. Tyton Partners and Bay View Analytics in Partnership with Every Learner Everywhere, posted July 2020, www.everylearnereverywhere.org/resources/time-for-class-2020/

VINH, N. X., EPPS, J., AND BAILEY, J. 2009. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, A. Danyluk, L. Bottou, and M. Littman, Eds. Association for Computing Machinery, 1073–1080. DOI: https://doi.org/10.1145/1553374.1553511

WALKINGTON, C. A. 2013. Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, 105(4), American Psychological Association, 932-945.

WANG, W., SONG, L., DING, S., WANG, T., GAO, P., AND XIONG, J. 2020. A Semi-supervised Learning Method for Q-Matrix Specification Under the DINA and DINO Model With Independent Structure. *Frontiers in Psychology*, 11(2120). Frontiers. https://doi.org/10.3389/fpsyg.2020.02120

WINTERS, T., SHELTON, C., PAYNE, T., AND MEI, G. 2005. Topic extraction from item-level grades. In *American Association for Artificial Intelligence 2005 Workshop on Educational Datamining*. AAAI.

YANG, Y. C., GAMBLE, J. H., HUNG, Y., AND LIN, T. 2014. An online adaptive learning environment for critical-thinking-infused English literacy instruction. *British Journal of Educational Technology*, 45(4), Wiley Online Library, 723-747.

ZAMORA, J. 2017. Recent Advances in High-Dimensional Clustering for Text Data. In *Claudio Moraga: A Passion for Multi-Valued Logic and Soft Computing*, Springer, 323-337.