Automated Prediction of Examinee Proficiency from Short-Answer Questions

Le An Ha[1] **Victoria Yaneva**[1,2] **Polina Harik**[2]
Ravi Pandian[2] **Amy Morales**[2] **Brian Clauser**[2]

[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
ha.l.a@wlv.ac.uk
[2]National Board of Medical Examiners, Philadelphia, USA
{vyaneva, pharik, rpandian, amorales, bclauser}@nbme.org

Abstract

This paper brings together approaches from the fields of NLP and psychometric measurement to address the problem of predicting examinee proficiency from responses to short-answer questions (SAQs). While previous approaches train on manually labeled data to predict the human ratings assigned to SAQ responses, the approach presented here models examinee proficiency directly and does not require manually labeled data to train on. We use data from a large medical exam where experimental SAQ items are embedded alongside 106 scored multiple-choice questions (MCQs). First, the latent trait of examinee proficiency is measured using the scored MCQs and then a model is trained on the experimental SAQ responses as input, aiming to predict proficiency as its target variable. The predicted value is then used as a "score" for the SAQ response and evaluated in terms of its contribution to the precision of proficiency estimation.

1 Introduction

The automated scoring of Short Answer Questions (SAQs) has been a longstanding research area in both psychometric measurement and NLP applications. Typically, a SAQ consists of a description of a problem followed by an open question which requires some form of a free short response by the test-taker. An example of a SAQ testing medical knowledge is presented in Table 1. The task at hand is to score the responses given by the test-takers, either manually or automatically. What makes the problem challenging is the fact that the answers, even the correct ones, tend to vary a lot in their expression and level of detail. The educational measurement literature has shown that the manual scoring of SAQs requires significant resources and often suffers from low rater agreement (Section 2.1). In NLP, the automated scoring of SAQ responses is still far from being solved and requires large amounts of expensive expert-rated data (Section 2.2). The lack of affordable and reliable SAQ scoring solutions perpetuates the assessment field's reliance on less informative but easy to score formats like Multiple Choice Questions (MCQs).

Despite the fact that the automated scoring of SAQs is a vibrant research area in both psychometrics and educational NLP, there are almost no interdisciplinary efforts that capitalize on the strengths of both fields (Section 2). In psychometrics, the majority of the effort is focused on assessing the reliability, validity and fairness of the scoring procedures for measuring examinee proficiency, with low exposure to state-of-the-art NLP techniques. By the same token, NLP studies utilize sophisticated language technology but do not evaluate the scoring systems in a way that answers the main question of *"Does this scoring method help with the more precise measurement of examinee proficiency?"*.

This paper presents an interdisciplinary study on automated SAQ scoring positioned in the intersection between NLP and psychometric measurement. Unlike previous NLP studies, the scoring method is evaluated on its ability to produce better assessment of examinee proficiency. As a step further, the proposed approach does not rely on manually rated responses for training, instead utilizing information available at the stage of the evaluation of new SAQs for their inclusion in standardized exams, as explained below.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 28th International Conference on Computational Linguistics, pages 893–903
Barcelona, Spain (Online), December 8-13, 2020

A previously healthy 26-year-old man is brought to the emergency department because of a tingling sensation in his fingers and toes for 3 days and progressive weakness of his legs. He had an upper respiratory tract infection 2 weeks ago. He has not traveled recently. He was unable to get up from bed this morning and called the ambulance. Temperature is 37.3C (99.1F), pulse is 110/min, respirations are 22/min, and blood pressure is 128/82 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 99%. Physical examination shows weakness of all four extremities in flexion and extension; this weakness is increased in the distal compared with the proximal muscle groups. Deep tendon reflexes are absent throughout. Sensation is mildly decreased over both feet. What is the most likely diagnosis?
Sample of correct answers: Guillain-Barré syndrome, acute immune-mediated polyneuropathy

Table 1: An example of a practice SAQ item

High-stakes standardized tests[1] need to ensure that the items included in the test are not too easy or too difficult for the intended population, and that each item contributes to the overall test score. For this reason, any newly developed item is first embedded in a live exam without being scored and with the examinees not knowing that it is not going to be scored (a phase known as *item pretesting*). The examinees are then scored on the live[2] items and their level of proficiency is estimated using Item Response Theory (IRT) modeling (Section 3). Once response data for a given pretest item has been collected (usually for a whole annual cycle), its psychometric charateristics are computed and it is later used as a scored item or discarded depending on its quality. Therefore, in the context of standardized testing: i) each newly developed SAQ would first be embedded in a live exam as a pretest item, and ii) there would be an examinee proficiency variable measured using their performance on live items.

In this study, we propose the use of proficiency as a dependent variable in a machine-learning model which takes pretest SAQ responses as input and aims to predict proficiency as output. Notably, the proficiency is measured independently using the *live items*, while the model predicts it using the responses to the *pretest items*. During the training phase, the model learns to map similar responses to one another and identify patterns associated with the responses of high-proficiency students versus those of lower proficiency students. The underlying hypothesis is that *people with similar proficiency would provide similar responses to the SAQs*. The predicted proficiency for a given new response is then used as its "score" and is evaluated in terms of its contribution to estimating the overall proficiency of the examinee who gave the response to the experimental question. We use data from a large medical exam, where in each form two pretest SAQs are presented alongside 106 scored MCQs.

Contributions The main contribution of this study is an approach for automated SAQ scoring in the context of standardized testing that does not rely on manually labeled training data. The evaluation of the results is done using psychometric approaches investigating whether or not the developed scoring procedure improves the measurement of proficiency for the examinee who gave the response. Core psychometric concepts that relate to NLP for automated scoring are explained in detail.

2 Related Work

This section describes the use of SAQs in testing, followed by studies on automated SAQ scoring.

2.1 Short-Answer Questions (SAQs)

While the format of open-ended questions was common at the beginning of the 20th century, by the 1950s, MCQs became the assessment tool of choice for many testing organizations. MCQs restricted an examinee's response set to a fixed number of options. This was quantitatively pleasing and lent itself naturally to an elegant and defensible set of mathematical measures. Statistics like item difficulty, item reliability and examinee proficiency formed the core of the emerging discipline of psychometrics (Gulliksen, 1950) and drove the wide use of MCQs in many assessments. Scoring MCQs was highly effi-

[1]Examples of well-known high-stakes exams include TOEFL (Test of English as a Foreign Language) (https://www.ets.org/toefl), SAT (Scholastic Assessment Test) (https://collegereadiness.collegeboard.org/sat) and USMLE (United States Medical Licensing Examination) (https://www.usmle.org/).

[2]"Live" items are ones that operationally scored, while "experimental" or "unscored" items are those being pretested

cient and was less susceptible to bias than the expert-based scoring processes utilized in prior assessment formats.

However, MCQs are not without their own drawbacks. Precisely because the examinee has been limited to a fixed set of responses, test-taking strategies and cueing become an important consideration. With regards to medical exams in particular (the exam domain used in this study), a patient does not present to a doctor with a fixed set of possible diagnoses (Veloski et al., 1999; Newble et al., 1979). In this context, SAQ items offer several benefits. They tend to be more difficult, reinforce long-term retention, and students, knowing that they will be taking an SAQ assessment, prepare in a way that facilitates more optimal learning (Sam et al., 2018; Pinckard et al., 2009). For these reasons and others, SAQs are of particular interest for many testing organizations, even though the two core issues involved in their scoring, cost and human bias, remain.

2.2 Automated SAQ scoring

The problem of SAQ scoring has received considerable attention, with several shared tasks and competitions orgnaized in the past (e.g., a SemEval shared task (Dzikovska et al., 2013), or the ASAP 2 Kaggle competition[3]). The task is to predict the human labels for each instance and, traditionally, this has been done using n-grams (Heilman and Madnani, 2015) or a wide variety of linguistic features as in Tack et al. (2017). Leacock and Chodorow (2003) use predicate-argument structure, pronominal reference, morphological analysis and synonyms to rate the questions. Other approaches include semi-supervised learning based on clustering, which are mainly effective for very short answers (single words or phrases) but do not generalize well to longer responses (Zesch et al., 2015).

In addition to the variety of methods used for SAQ scoring, several other problems relating to other aspects of the task have been investigated. Heilman and Madnani (2015) explore the effects on performance of training sample size to help answer the question of how much data needs to be gathered and labeled before automated scoring for a new prompt is deployed. Ramachandran and Foltz (2015) tackle the problem of rubric coverage by generating reference texts from summarized top-scoring student responses. Rudzewitz (2016) combines the task of SAQ scoring with that of plagiarism detection showing that incorporating features from other domains (with lexical and semantic ones revealed as the most informative) can improve the results. Willis (2015) use NLP to assist human raters in incrementally developing a set of automatic marking rules, which can be further used for automatic scoring.

In all of these applications, the aim is to predict the human scores assigned to the responses. As the ASAP 2 challenge puts it: "Your success depends upon how closely you can align your scores to those of human expert graders" (`https://www.kaggle.com/c/asap-sas`). This framing of the problem does not take into account the fact that human graders themselves may be biased or make errors. In other words, these models may be useful to partially solve the problem of effort associated with SAQ scoring but do not help with solving the problem of rater accuracy. In addition, the framing of the problem as one of human label prediction still requires the labeling of a training set, which is an expensive and time-consuming procedure, thus addressing the need for cost reduction only partially. As we discuss in the next sections, the approach proposed in this study aims to address both of these gaps.

3 Data

An example SAQ is presented in Table 1. As can be seen, the responses consist of very short phrases (less than 60 characters), hence the format is sometimes referred to as very-short-answer questions (VSAQ).

Item writing: The SAQs used in this study were directly derived from existing MCQs by removing their answer options and slightly rewriting the lead in question (e.g., "*Which of the following is the most likely diagnosis?"* is converted to *"What is the most likely diagnosis?"*). The item text for the initial MCQs was developed by experienced item-writers following strict guidelines for content, information structure and formatting. All MCQs were formerly pretested and met quality criteria for live use.

Item selection: When choosing MCQs to be converted to SAQs, concern was paid to a breadth of topics (as opposed to choosing several items that covered the same diagnosis), difficulty, and discriminatory

[3] `https://www.kaggle.com/c/asap-sas`

power. The latter two are common measures used in psychometrics, computed as follows:

i) Item difficulty: The difficulty of an item was defined by the proportion of its responses that are correct, commonly referred to in the educational-testing literature as its *P-value*, calculated as follows:

$$P_i = \frac{\sum_{n=1}^{N} U_n}{N},$$

where *Pi* is the P-value for item *i*, *Un* is the 0-1 score (incorrect-correct) on item *i* earned by examinee *n* , and *N* is the total number of examinees in the sample. For example, a P-value of .3 means that the item was answered correctly by 30% of the examinees. It is important that items that are too difficult or too easy be filtered out, as they would not bring valuable information about examinee proficiency (Yaneva et al., 2020). The criteria for the exclusion was a P-value below 0.4 and above 0.95 and was defined on the basis of industry standards.

ii) Item discriminatory power or item biserial correlation (often referred to as Rb parameter in educational testing), is the correlation between examinees' responses on the given item and examinees' total test score. The purpose of this metric is to ensure the quality of the item, where examinees who perform better on the test overall, must be more likely to succeed on the item than those examinees who performed worse. The crierion for the Rb parameter used in this study is that it had to be greater than 0.1, in order to exclude items which do not contribute to the overall test score.

Filtering on these criteria resulted in an initial pool of 150 MCQs. The next step was to conduct an editorial review and have a subject matter expert answer the items in the new format, as explained below.

iii) Answer length: Of the 150 items, 30 were eliminated based on editorial review from experienced item writers which indicated that answering these items would require a significantly longer response. The rationale behind excluding those was to eliminate questions that would tests different skills such as the ability to write well, which were not part of the construct of clinical knowledge.

iv) SAQ difficulty as determined by a subject matter expert: The remaining items were presented to a subject matter expert who solved the questions in the SAQ format. This expert was an experienced physician and had no prior knowledge of the MCQ versions of the items. Based on the expert's judgement, items that were considered too difficult to answer without having an option set were excluded.

After following the elimination steps presented above, the final number of SAQ items included in the study was 100. Data for those were collected for one full year (August 2018 through August 2019) and amounted to 50,894 individual responses from 25,447 test-takers.

Exam administration: The data for the study were collected during the administration of Medicine Clinical Science subject exam. This exam was developed at the National Board of Medical Examiners and distributed to a large number of medical schools in the US and Canada[4] to use as a subject exam at the end of a semester. Students and faculty members were notified in advance that SAQs would be included in some exam forms and were given ample time to prepare for the new format. The test takers had no knowledge that these SAQ items were pretest items and approached them as scored ones. During the exam, each test-taker saw 110 questions in total, four of which were experimental questions. These experimental questions were presented in SAQ or MCQ format (two of each) such that no examinee would see the two versions of the same item. The rest of the test comprised 106 MCQs, the responses to which were scored and used to form the final grade for each examinee. The response field was limited to a 60 character input and the following instructions were presented: *Your answer should be brief and should: respond directly to the question and only to the question (i.e., do not provide a rationale); be as specific as possible; consist of no more than a few words (e.g., antibiotic therapy).*

Proficiency estimation The responses given to the 106 live MCQs per form were first automatically scored by mapping the letter of the selected option to the letter of the correct response for each examinee and item. After that, the scores for the 106 MCQs were used to compute examinee proficiency.

We used one-parameter Rasch model, the most common application of IRT, to estimate examinee proficiency (Rasch, 1960). Within the IRT framework, probability of a correct response on a given item is

[4]Accredited by the Liaison Committee on Medical Education (LCME)

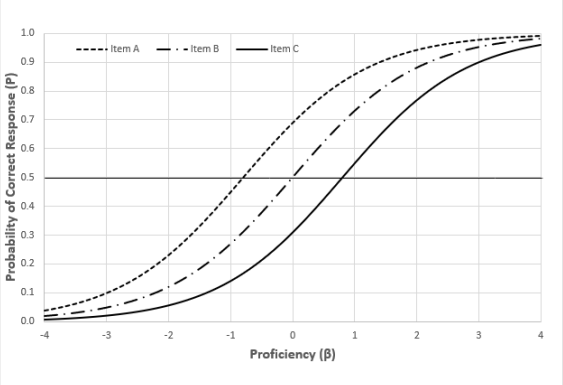

Figure 1: Hypothetical example of the relationship between probability of correct response for items A, B, and C and proficiency

conceptualized as influenced by both examinee proficiency and item characteristics. More specifically, in Rasch IRT model, probability of a correct response on an MCQ item is a function of examinee proficiency and one item parameter - the difficulty of that item. The higher an examinee's proficiency relative to the difficulty of an item, the more likely the examinee to answer that item correctly. When examinee υ applies their proficiency β_υ to answer item ι of difficulty $\delta\iota$, the probability of a correct response (1) can be expressed as follows:

$$P\{X_{\upsilon\iota} = 1|\beta_\upsilon, \delta\iota\} = \exp(\beta_\upsilon - \delta\iota) / (1 + \exp(\beta_\upsilon - \delta\iota))$$

Figure 1 is a hypothetical example of the relationship between probability of correct response for items A, B, and C and proficiency. Within IRT framework, item difficulty is defined as the location on the proficiency scale where the chance of answering an item correctly is 50% (P=0.5). In our example, item difficulties are -0.8. 0 and 0.8 for items A, B and C, respectively. Typically, the probability of correct response is greater than 0.5 for examinees whose proficiency is above item difficulty. In our example, an examinee with proficiency 1 is very likely to answer all 3 items correctly, while examinee with proficiency 0.3 is likely to correctly respond to items A and B only. In our study, examinee proficiency β_υ was obtained by Joint Maximum Likelihood Estimation[5], an estimation process that iterates through data for all items and all examinees in the sample. The estimation sample consisted of about 2K-4K examinee responses to each MCQ item.

4 Method

This section presents the approach for automated prediction of examinee proficiency. First, various types of embeddings are generated and compared (Section 4.1) in order to select the best-performing ones for the main experiments. After that, several baselines are developed using the correct answer from the MCQ version of the items as a rubric (section 4.2). Next, the proposed approach is presented in Section 4.3. Finally, the metrics used to compare different approaches can be found in Section 4.4.

4.1 Embedding Selection

We train various embedding models on medical and generic corpora in order to select those that perform best. The models include BERT (Devlin et al., 2018), ELMo (Peters et al., 2018), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013). For BERT we compare the model-produced pooled vectors that represent the whole phrase (henceforth "BERT pooled"), and the embedding vectors of individual words, which can then be mean-pooled (henceforth "BERT embedding"). In terms of domain-specific embeddings, we train BERT and ELMo on approximately 22,000,000 MEDLINE medical abstracts[6]. For ELMo we used three different types of heads: the first layer (tokens), the last layer

[5]https://www.winsteps.com/index.htm

[6]`https://www.nlm.nih.gov/bsd/medline.html`

(lm), and the mean of the three LSTM layers (means). In addition, we experiment with the pre-trained clinical BERT embeddings provided by Alsentzer et al. (2019) and train Word2Vec on the PubMed Central Open Access database[7] and clinical notes. In terms of generic corpora, we experiment with BERT (whole word masking (wwm) configuration [8]) and ELMo extracted from the 1B Word corpus (Chelba et al., 2013), as well as GloVe (Pennington et al., 2014) (Wikipedia 2014 + Gigaword 5) and Word2Vec (Mikolov et al., 2013) (Google News Corpus[9]). The embedding evaluation results are presented in Table 2 and discussed in Section 6.

4.2 Baselines

For automated scoring baselines, we experimented with mapping the examinee responses to the correct answer to the item in its MCQ form (also known as MCQ key). This approach allows automated SAQ scoring without the need of human rating but is limited to items that have an MCQ form or a correct answer phrase used as a rubric. We also experiment with several methods to perform that match:

Exact match This is an exact match between the new instance and the MCQ key. If the new instance matches an answer key, then it is considered *correct*, otherwise it is considered *incorrect*.

Exact match + Synonymy relation We use a heuristic to identify whether a response can be considered an equivalent to the MCQ key. The two are considered equivalent to each other if they are either a typo of each other, or the customised edit distance between the two strings is zero. In the latter, the cost of deletion of stop words[10] is zero, the cost of replacing a word or a phrase with its WordNet (Miller, 1995) or UMLS (Unified Medical Language System) Meta-thesaurus (Schuyler et al., 1993) equivalent is zero, and the cost of replacing a typo with its correct form is also zero. Such heuristic is designed to be flexible and would determine that: *"discontinue the simvastatin", "stop statin" "discontinue simastatin"* and *"discontinue simvastatin"* are all equivalent, and similarly for *"overnight polysomnography"* and *"overnight sleep study"*. The scoring is done as with the exact match approach.

Embedding similarity Each item was represented as a vector using mean pooling of all tokens. The embedding similarity approach is based on the cosine similarity between the vector representation of a new instance and an MCQ key. The assigned score is the cosine similarity between the two.

4.3 Predicting Examinee Proficiency from SAQ Responses

This subsection describes an approach to proficiency estimation from SAQ responses.

Model input and output The input for the model are the differences between the responses' embeddings vectors and the correct answer embedding vectors. The model aims to predict the proficiency level (model outputs) given the responses. We train the model on the first half of the responses together with the proficiency estimates for each test-taker based on their performance on the 106 MCQs. The model is then asked to predict proficiency scores for the test set (the remaining 50% of the data). Specifically, given a response to a specific item, the model is asked to estimate the most likely proficiency level of somebody who would produce such a response. This number (the most likely proficiency level) could then be used as a "score" for that response, or used directly in the IRT equations.

Training and test split The somewhat unusual division of the data into training and test sets presented here (50/50) is done with a specific practical consideration in mind. In practice, as the new items are first embedded in the exam and administered, there would initially be very few responses. As the annual cycle of the exam progresses and more data is accumulated, the data collected up to a certain point in time forms the training set and is used to score the data coming after this temporal cutoff point (the test set). Here, we present a conservative scenario where half of the data is used for training and the other half for testing, which would allow for an implementation of the automated system sometime in the middle of the annual cycle. While the exploration of the effects on training and test set sizes is not the focus

[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[8] https://storage.googleapis.com/bert_models/2019_05_30/wwm_uncased_L-24_H-1024_A-16.zip

[9] https://news.google.com

[10] Stop words in our case include the generic stopwords, and domain specific stop words such as "study", "level", "test", etc.

of this paper (see (Heilman and Madnani, 2015) for a study on this topic), it is conceivable that model performance would improve as the year progresses and more training data becomes available.

What the model learns Conceptually, what the model aims to learn is how examinees of different proficiency levels answer SAQs, so that an instance similar to the responses of examinees with high proficiency levels is scored high and vice-versa. Put differently, instead of using a human scorer to provide the ground truth, the answers of the high-proficiency students from the training set provide a measure of what are likely high-scoring responses and the answers of the students with lower proficiency levels provide a measure of what are likely low-scoring responses. It is important to note that the answers from the training set do not need to be scored by human raters; it is their relationship to the proficiency level estimated from the 106 MCQs that is used to train the model and this relationship is later extrapolated to the test set responses by virtue of their semantic similarity to the training set ones. In practical terms, as the input to the model is a vector representing the difference between the response and the correct answer, the problem is to learn how different dimensions in the vector representation influence the outputs, or in other words, their weights. This contrasts the embedding similarity approach, in which the similarity is calculated under the assumption of equal dimension weights. To learn the weights of each dimension of the embedding vectors we use the default-parameter RidgeCV regressor from the sci-kit learn Python library (Pedregosa et al., 2011), which performs ridge regression with Generalized Cross Validation (also known as efficient leave-one-out cross validation).

When training the model we experiment with two settings. In the first one, we train one model using the responses to all items, and, as a result, we learn a set of weights for all items. In the second setting, we train a model for each item, thus learning different sets of weights for different items.

4.4 Evaluation Metrics

This section presents commonly used metrics from the field of psychometrics for expressing the relationships between individual scores and proficiency, computed as follows:

Item-level correlation is represented as $r_i = pearsonr(score_{i,j}, \theta_j)$, where $score_{i,j}$ is the score of the response of examinee j for item i and θ_j is the proficiency level of the examinee j, as measured using 106 operationally scored items. The correlation is calculated using all examinees js that responded to the item i. A higher correlation indicates the item can contribute more to the scoring and vice versa.

Examinee-level correlation is represented as $r = pearsonr(sum(score_{i,j}, score_{k,j}), \theta_j)$, where the $score_{i,j}$, and $score_{k,j}$ are the scores of the responses of examinee j for the two experimental items i and k. The higher the correlation, the more predictive power of the scores to the proficiency level. The correlation is calculated using all examinees j.

We apply these two metrics to both the experimental MCQ items and the SAQ items. For the MCQ items, the scores are either 0 (incorrect choice), or 1 (correct choice) for the two items that were presented to the examinees. For SAQs, the "scores" are produced using the proposed method (Section 4.3).

5 Results

Table 2 shows that most embedding types perform in a similar way, with no statistically significant differences between those with higher correlations. Best results were achieved by ELMo trained on Medline clinical abstracts and BERT embeddings trained on Wikipedia + BookCorpus. Interestingly, no clear pattern emerged with respect to the usefulness of domain-specific embeddings. While there is no clear explanation for this lack of an effect, it may possibly be related to the size of the corpora or it could mean that most of the variations found in the answers were everyday variations, rather than medicine-specific ones (e.g., *"drug side effects"* and *"effect of medication"*). Based on these results, the embeddings presented hereafter refer to the best performing embedding type, namely ELMo (token, 1024 dimensions) trained on Medline clinical abstracts. Next, the results for item-level correlations to proficiency are presented in Table 3, while examinee-level correlations are presented in Table 4.

As can be seen from the item-level results (Table 3), the approach proposed in this study has the highest mean correlation to proficiency (.26), which is higher than all baselines and compared to .21 for the MCQ version of the items. Its SD (.072) is slightly higher compared to that of MCQ scores (.065). It is also

		Item level		Examinee level	
Embedding type	Corpus	Mean	SD	Mean	SD
Bert wwm (embedding)	Medline	.19	.044	.23	.058
Bert wwm (pooled)	Medline	.17	.044	.20	.059
Bert wwm (embedding)	Wikipedia + BookCorpus	**.19**	.047	**.23**	.062
Bert wwm (pooled)	Wikipedia + BookCorpus	.13	.042	.17	.058
Clinical Bert (embedding)	MIMIC-III v1.4	.13	.053	.16	.077
Clinical Bert (pooled)	MIMIC-III v1.4	.18	.044	.22	.058
ELMo (512 dimensions)	One Billion Word Benchmark	.18	.039	.21	.052
ELMo lm (1,024 dimensions)	Medline	.18	.041	.22	.054
ELMo mean0 (1,024 dimensions)	Medline	.18	.04	.22	.054
ELMo token (1,024 dimensions)	Medline	**.19**	.039	**.23**	.054
GloVe (300 dimensions)	Wiki 2014 + Gigaword 5	.17	.039	.20	.054
Word2Vec PMC (300 dimensions)	PubMed Central OA	.16	.042	.20	.054
Word2Vec PMC (600 dimensions)	PubMed Central OA	.16	.044	.20	.057
Word2Vec Generic (300 dimensions)	Google news	.17	.039	.20	.054

Table 2: Mean correlation and SD for various embedding types and training corpora. The results are calculated across all tasks (baselines + proficiency modeling) at both the item-level and the examinee-level. In this experiment, we run the experiments on more settings than reported in this paper.

Method	Mean	STD	Min	25%	50%	75%	Max	N
MCQ items	.210	.065	.039	.179	.210	.253	.341	53,206
Automated Scoring Baselines								
Exact match	.104	.100	-.061	.028	.082	.175	.361	37,107*
Exact match + synonyms	.103	.110	-.061	.006	.104	.191	.337	43,231*
Embedding similarity	.198	.088	-.021	.154	.196	.263	.409	50,389
Proficiency modeling (using ELMo embeddings)								
Joint model for all items	.262	.084	.047	.219	.270	.319	.448	26,626
One model per item	**.267**	.079	.059	.222	.265	.323	.457	26,626

Table 3: **Results per item:** Pearson correlation between various approaches and the examinee proficiency metric. The column N represent the number of scored responses using each method. *: We ignore items to which nobody responded either exactly as or synonymous to the MCQ key, as such scores (all zeros) are not informative.

interesting to note that the embedding similarity approach performs almost as good as MCQ scores. In terms of the coverage for the different approaches, in the matching-MCQ-key baseline, there are only 73 items (37,107 responses) for which at least one examinee responded with an exact match to the MCQ key, and there are 83 items for which there exists at least one response that either exactly matches the MCQ key, or is considered by the heuristic to be equivalent to the MCQ key (43,231 responses). The embedding similarity approach covers the whole set of responses (50,389). The proficiency modeling approach proposed by this study covered 26,626 responses due to having 50% of the responses in the training set. However, once trained, the model can be applied to an unlimited number of new responses.

The examinee-level results (Table 4) reveal a similar pattern, where the proposed proficiency-modeling approach provides a correlation of 0.316 compared to 0.261 for the MCQ scores approach. This result refers to training one model per item and it outperforms the model trained on data from all items together (0.288). This conclusion is in accordance with the results for item-level correlations.

6 Discussion

Probably the most important finding from this study is that, for our sample, SAQ "scores" assigned through the proposed approach can contribute more to the precision of proficiency measurement than the MCQ form of the same items (indicated by the higher correlation between the former to proficiency). In addition, the method outperformed baselines relying on matching the response to a rubric (the MCQ key), which is the approach of choice in most existing approaches to SAQ scoring. For the baselines, the moderate success of matching the responses to a single key phrase is likely a consequence of the

Method	Examinees	r
MCQ items	26603	0.261
Baselines		
Exact match	13,371*	0.159
Exact match + synonyms	18,110*	0.154
Embedding Similarity	24,659	0.212
Proficiency modeling (ELMo)		
Joint model for all items	12,918	0.288
One model per item	12,918	0.316

Table 4: **Results per examinee:** Pearson correlation between the sum of two scores per examinee and proficiency. *: We ignore items to which nobody responded either exactly as or synonymous to the MCQ key, as such scores (all zeros) are not informative.

shortness of our responses and is somewhat similar to the observation made by Zesch et al. (2015) that semi-supervised clustering approaches work for very short but not for lengthy responses. It is therefore possible that this approach may not generalize well over formats requiring longer responses and the baseline results for those could be lower. Notably, the presented results are a product of a conservative scenario where only 50% of the data is used for training. Overall, these preliminary results show that a model which learns the associations of different responses to proficiency has the potential to improve the precision of proficiency measurement, while also eliminating the need for labeled training data.

Validity: An interesting question related to the validity of this approach is the argument that the proficiency metric is developed based on MCQs, which may measure different kinds of skills compared to SAQs. In fact, it is this hypothesis that ignited the interest in using SAQs in the first place. While acknowledging that such subtle differences may exist, it is not far-fetched to assume that both formats measure the generic construct of clinical skills knowledge. Proficiency was measured using 106 MCQs with known psychometric characteristics and can therefore be regarded as a reliable metric of clinical skill knowledge. It is also possible that the output would correlate better if proficiency is measured using other SAQs rather than MCQs, but this remains to be tested empirically.

Limitations and operational feasibility: While offering better estimations of proficiency, the approach has the drawback of having low interpretability which is an important prerequisite for high-stakes standardized exams. In addition, its relevance needs to be regularly assessed to ensure that there are no major differences between the responses of different-year cohorts. With these caveats in mind, the approach can be easily applied to other standardized exams. Since large-scale standardized tests typically require that any new item is pretested, there will always be other items (in MCQ, SAQ or other format), which can be used to measure the ground-truth proficiency. Once trained on the pretested data, the model can be applied for scoring an unlimited number of new responses and the item can be used as a live one. The method could also be useful for predicting examinee proficiency in formative assessments such as those in Massive Open Online Courses (MOOCs), which do not offer manual scoring due to the high volume of responses (in fact, they rarely use the SAQ format for the same reason). In these cases, high-proficiency students could be defined through other criteria such as likelihood of course completion, quiz answers, peer ratings of forum responses, etc. While this scenario would require research on defining proficiency in the context of MOOCS, the proposed approach could be well suited to provide short-answer scoring when manual labeling is not feasible.

It is important that the findings presented in this study are regarded as preliminary until a large-scale evaluation is conducted across different samples of questions and examinees, as well as compared to results from human scoring. The validity of the approach needs to be further tested with proficiency measured using other SAQs. In spite of its limitations and the preliminary nature of the findings, this study presented a proof of concept showing that pretest SAQ scoring through modeling proficiency is a viable way to assign meaningful scores to new responses. These were shown to contribute to better measurement of overall proficiency for the pretest items without relying on a labeled training set.

7 Conclusion

This paper presented experiments towards modeling examinee proficiency as an alternative to the human scoring of SAQs. First, an IRT proficiency metric was computed based on the test-takers' performance on 106 MCQs. After that, the responses to SAQs were used as an input to a model aiming to predict examinee proficiency. The results were compared to those achieved through MCQ scores and through several automated-scoring baselines in terms of the correlation between the "scores" given by each method and the overall examinee proficiency. The results indicated that the predicted proficiency for items in the test set used as a response score for these items has a higher correlation to the proficiency metric at both the item level and the examinee level compared to the MCQ scores and the automated scoring baselines.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, NORTH TEXAS STATE UNIV DENTON.

H Gulliksen. 1950. *Theory of Mental Tests*. New York: Wiley.

Michael Heilman and Nitin Madnani. 2015. The impact of training data on automated short answer scoring performance. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 81–85.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

DI Newble, Avril Baxter, and RG Elmslie. 1979. A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education*, 13(4):263–268.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

R Neal Pinckard, C Alex McMahan, Thomas J Prihoda, John H Littlefield, and Anne Cale Jones. 2009. Short-answer examinations improve student performance in an oral and maxillofacial pathology course. *Journal of Dental Education*, 73(8):950–961.

Lakshmi Ramachandran and Peter Foltz. 2015. Generating reference texts for short answer scoring using graph-based summarization. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–212.

G Rasch. 1960. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. copenhagen: Danmarks paedagogiske institut.

Björn Rudzewitz. 2016. Exploring the intersection of short answer assessment, authorship attribution, and plagiarism detection. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 235–241.

Amir H Sam, Samantha M Field, Carlos F Collares, Cees PM van der Vleuten, Val J Wass, Colin Melville, Joanne Harris, and Karim Meeran. 2018. Very-short-answer questions: reliability, discrimination and acceptability. *Medical education*, 52(4):447–455.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated cefr-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

J Jon Veloski, Howard K Rabinowitz, Mary R Robeson, and Paul R Young. 1999. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Academic medicine: journal of the Association of American Medical Colleges*, 74(5):539–546.

Alistair Willis. 2015. Using nlp to support scalable assessment of short free text responses. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 243–253.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2020. Predicting item survival for multiple choice questions in a high-stakes medical exam. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6812–6818.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132.