A Comparative Study of Methods for a Priori Prediction of MCQ Difficulty

Ghader Kurdi [a,*], Jared Leo [a], Nicolas Matentzoglu [a], Bijan Parsia [a], Uli Sattler [a], Sophie Forge [b], Gina Donato [b] and Will Dowling [b]

[a] *Department of Computer Science, The University of Manchester, Oxford Road, Manchester, M139PL, UK*
[b] *Elsevier, 1600 John F. Kennedy Boulevard, Philadelphia, PA 19103, USA*

Editor: Dagmar Gromann, University of Vienna, Austria
Solicited reviews: Vinu Ellampallil Venugopal, University of Luxembourg, Luxembourg; Three anonymous reviewers

Abstract. Successful exams require a balance of easy, medium, and difficult questions. Question difficulty is generally either estimated by an expert or determined after an exam is taken. The latter provides no utility for the generation of new questions and the former is expensive both in terms of time and cost. Additionally, it is not known whether expert prediction is indeed a good proxy for estimating question difficulty.

In this paper, we analyse and compare two ontology-based measures for difficulty prediction of multiple choice questions, as well as comparing each measure with expert prediction (by 15 experts) against the exam performance of 12 residents over a corpus of 231 medical case-based questions that are in multiple choice format. We find one ontology-based measure (relation strength indicativeness) to be of comparable performance (accuracy = 47%) to expert prediction (average accuracy = 49%).

Keywords: ontologies, semantic web, automatic question generation, difficulty modelling, difficulty prediction, multiple choice questions, student assessment

1. Introduction

Multiple choice question (MCQ) examinations are widely used to assess the knowledge and skills of students and the quality of teaching instruments. Using good-quality questions is essential for achieving these purposes. Several criteria exist for measuring question quality, as discussed in [1–3]. Good quality questions need to be, among other things, 1) valid (i.e., they measure what they are supposed to measure); 2) discriminating (i.e., they discriminate between high- and low-information students); 3) fair (i.e., their results are not biased in favour of a subgroup within the cohort); and 4) of appropriate difficulty. Difficulty of MCQs is usually[1]. defined as the proportion of students solving a question correctly out of the total number of students attempting the question, and is known as *percentage correct*.

The difficulty criterion is of importance, attributed by its effect on the other quality criteria. Knowledge about difficulty level and sources of difficulty in questions provides insights into whether other quality criteria are satisfied or not. With regards to validity, being able to answer the question 'what makes a particular question easy or difficult?' is an important step in understanding 'what does the question measure?' For example, questions that are difficult due to their linguistic complexity are usually not valid in tests other than in language tests. This is because it is not clear whether students' failure in answering these questions is due to the language factor or to their lack of the knowledge or skills of interest. In addition, inappropriately difficult or easy questions tend to have bad discrimination because, either almost none of the students solve them

*Corresponding author. E-mail: ghader.kurdi@manchester.ac.uk.

[1]This is based on a recent systematic review on difficulty prediction of assessment questions [4].

or all of the students solve them correctly. Finally, the difficulty level of questions is a major determinant of the fairness of exams, especially when different exam forms are used (equally difficult forms are needed) or when question selection is allowed (equally difficult questions are needed).

While information about the difficulty of questions is essential for designing exams, percentage correct can only be retrospectively determined. Traditional means of estimating difficulty are by obtaining it from previous administrations of the questions, if previous statistics are available, or by relying on experts' estimation, which is usually the case in small-scale exams.

With recent advances in automated procedures for generating questions [5–11],[2] allowing the ability to generate a huge number of new questions, the need for measures that approximate prospective difficulty becomes ever more vital. These measures can be incorporated into the generation process allowing the generation of questions with the desired difficulty (satisfying the needs of exam developers) or, at least, with appropriate difficulty (filtering inappropriately easy and difficult questions). Furthermore, organising auto-generated questions by difficulty will reduce exam developers' efforts in sorting through, and trying to predict the difficulty of, a large number of questions. Finally, reliable predictive measures will allow moving progress toward the goal of generating exams automatically.

The majority of existing automatic difficulty prediction models are machine-learning based approaches [see, for example, 14–17] that have merely been used for finding correlations in existing data as opposed to prediction. Existing cross-validated models that have been developed for prediction [18–20] are highly domain-specific which limit their utility. However, in a prior work [7, 8], we have developed two ontology-derived measures which are based on a domain-independent model of difficulty.

Since the aforementioned ontology-based measures have neither been evaluated thoroughly nor compared to each other in a systematic way, we continue the work we carried on [7, 8] by evaluating the performance of the previously proposed measures. Specifically, we extend our work by collecting data about student performance on a set of auto-generated questions that were validated in [8]. The data about student performance is used as a gold standard for which expert prediction (available from [8]) as well as the prediction of ontology-based measures are compared to. This allows us to validate our measures and determine whether they are suitable for replacing expert estimations when constructing exams.

This paper aims to address the following research questions:

RQ1: How accurate is expert prediction of difficulty against student performance?
  * How well do experts perform in comparison to guessing?

RQ2: How accurate are automatic difficulty prediction (ADP) methods against student performance?
  * How well does each method perform in comparison to guessing?
  * How well does each method perform in comparison to the other method? and
  * How well does each method perform in comparison to domain experts?

We collected difficulty information for 231 questions through a study involving 15 medical experts and a cohort of 12 residents. Similar to studies conducted on other domains [21, 22] we found that the difficulty of case-based MCQs was moderately predicted by domain experts (average accuracy = 49%). We also found the automated measure proposed in [8] to be of comparable performance to experts (accuracy = 47%) and to represent an economical alternative.

The main contributions of this paper are:

- User studies in the medical domain investigating the predictive performance of domain expert and automated ontology-based measures;
- A detailed analysis of the performance of ontology-based measures for difficulty prediction that show, by example, the minimum set of criteria that need to be considered in evaluating the performance of similar measures; and
- A fairly large question set (231 questions, of which 92 were answered by at least 10 participants) annotated with percentage correct and expert prediction that can be used for testing the performance of new approaches to difficulty prediction.[3].

[2] For an overview of the field of automatic question generation, the reader is referred to the systematic reviews reported in [12, 13].

[3] Available at: https://github.com/grkurdi/A-Comparative-Study-of-Methods-for-a-Priori-Prediction-of-MCQ-Difficulty-dataset

2. Background

2.1. Multiple choice questions

MCQs consist of two components:

- The *stem*: a textual element that represents a problem to be solved, possibly accompanied by supplementary elements such as tables or graphs;
- The *options*: a set of alternatives to select from. Standard MCQs, known as single response questions, have one correct option (known as the *key*) and three or four incorrect options (the *distractors*) [23]. Another popular type of MCQs, known as multiple response questions, have at least two keys. Distractors are added so that the number of options typically sums to four or five.

Writing high-quality MCQs is known to be challenging and expensive. The challenges faced by exam developers are apparent from the low quality of MCQ examinations as indicated by several studies investigating their quality. For example, the authors of [24–28] found more than 50% of investigated MCQs to contain at least one item writing flaw.[4] Other studies [29–31] reported that the percentage of MCQs with all their distractors being considered functional[5] is low (between 5% and 23%).

2.2. Ontology-based MCQ generation and difficulty prediction

Given the challenges faced by test developers in constructing high-quality MCQs, automated approaches for question generation have come into play. Ontologies have been increasingly used, in research contexts, as a source for automatic generation of questions [5–9]. We attribute their increased use to the following reasons. The first reason is the availability of ontologies with potential educational value. These ontologies contain exact facts and represent domains of interest precisely and non-ambiguously in a machine-processable way. Besides that, ontologies are supported by standard reasoning services and the development of further supporting tools and services is an active research area. Another reason is that, compared to texts, the process of finding good distractors is easier. As an example, consider the question 'Which city is located in the UK?', generated from a Wikipedia[6] article about the United Kingdom. The cities mentioned in the articles are most likely to be UK cities. Even if a non-UK city is mentioned, detailed information about it, which is important in deciding whether it serves its intended purpose as a distractor, cannot be found in the same article. For a detailed systematic review of automatic question generation methods, the reader is referred to [12, 13].

One point worth mentioning is that underlying difficulty models are not part of most existing question generation approaches. According to Alsubait [13], apart from the similarity-based approach (outlined in Section 3.1), only two question generation approaches [37, 38] take into account generating questions with controlled difficulty but without providing an experimental evaluation of the performance of difficulty prediction. The automatic measures compared in this paper represent existing, domain non-specific measures of MCQ difficulty. Other measures are either variants of the similarity approach [39], designed for questions with other response formats, or categorised by being domain- or question-specific [18, 19, 37, 40].

2.2.1. Case-based question generation

One of the limitations of current question generation approaches is the simplicity of the generated questions in terms of both cognitive level,[7] with the majority of generated questions in [5, 7, 9] testing recall of information, and structure, with generated stems in [5, 7, 9] containing at most two concepts. In a recent study [8], we tackled the generation of medical case-based questions (see question Q2) using a large medical ontology. What is interesting about these questions is that they are widely used in medical education and that answering these questions requires more than just recall of information [42–44]. From a computational point of view, the complex structure of their stems, consisting of multiple concepts, introduces additional challenges of coordination between these concepts and understanding the role they play in question difficulty. The generation approach was evaluated through expert review of questions generated from four medical specialities. More details on the set of generated questions will be given in Sections 4.1.2 and 4.2.2.

[4] Violations of best practices as suggested in MCQ-writing guidelines.

[5] Functional distractors are those selected by at least 5% of examinees [30, 32, 33].

[6] Wikipedia has been used as a source for question generation by [34–36].

[7] The mental process involved in question-solving as described in Bloom's taxonomy [41], a popular classification of cognitive levels.

3. Competing measures

The target of difficulty prediction is to assign difficulty levels (easy, medium, or difficult), as derived from percentage correct (to be discussed in Section 4.2.4). The two ontology-based measures compared are described in this section.

3.1. Similarity-based measure

A plausible prediction model was proposed in [7], in which the similarity between the key and the distractors was suggested as an indicator of MCQ difficulty. Increasing the similarity between the key and distractors results in increasing the difficulty of MCQs. The rationale is that more knowledge is required to differentiate between key and similar distractors. As an example, consider the following question (Q1) taken from [13]. The most similar distractor to the key, and the most difficult to eliminate, is the option 'the tongue' since this option shares with the key the feature of being a body part. On the other hand, elimination of the options 'disease' and 'glossitis' is easier since they do not have shared features with the key.

Q1: Pyorrhoea occurs in ...:
A. the tongue
B. glossitis
C. the gums **(key)**
D. a disease

To control the difficulty of questions, Alsubait et al. [7] developed a similarity measure that is based on Jaccard similarity [45] and intended to be used with ontologies. The similarity measure is defined as follows:

$$similarity(k,d) = \frac{\#(subs(k,O) \cap subs(d,O))}{\#(subs(k,O) \cup subs(d,O)}$$

where the numerator is the number of common subsumers between the key k and a distractor d (i.e. both are class names selected from an ontology) and the denominator is the number of all subsumers of both k and d. The overall difficulty of the question is then defined as the average similarity between the key and distractors.

Different variants of the similarity measure, each of which uses a different set of subsumers,[8] were defined in [13]. These include:

- Atomic similarity in which only atomic subsumers of k and d are counted and
- Sub-similarity in which both atomic and complex (i.e. sub-expressions) subsumers of k and d are counted. We used this variant of the similarity measure in the experiments reported in this paper.

Preliminary studies showed that the similarity measure has a good difficulty prediction [7, 13]. In the absence of other domain-independent measures that are empirically supported, the similarity measure is considered as the gold standard for automatic difficulty prediction. However, one of the limitations of this measure is that it does not take into account the contribution of the stem to the difficulty of questions. While this did not represent a problem in questions having simple stems (e.g. 'What is X?' where X is a concept name), we believe that the role the stem plays is a major influencer on the difficulty of case-based questions that are characterised by stems that contains multiple terms (i.e. multi-term questions). In addition, the similarity measure is developed based on the assumption that all relational axioms have the same strength (i.e. a disease is either associated or not associated with a clinical finding). However, this is not always the case, especially in the medical domain where relations such as *hasClinicalFinding* have different degrees of strength (e.g. most common, common, or rare clinical finding). These limitations motivate us to develop the new difficulty measure 'Relation Strength Indicativeness' introduced in [8].

3.2. Relation Strength Indicativeness

A measure of question difficulty estimates difficulty by combining several calculations that exploit the relational axioms of an ontology, along with their *strength*. This measure, coined *relation strength indicativeness (RSI)*, requires an ontology to contain existential class axioms, i.e., those axioms of the form $A \sqsubseteq \exists R.B$,[9] where A and B are classes, and their relation R has an associated strength (Figure 1 demonstrates how the strength of relations can be encoded).

The proposed difficulty measure targets more complex types of questions, such as Q2 below, when compared to simple questions, such as Q1. The two main calculations RSI uses involve *stem indicativeness* and *option entity difference*. The former intuitively repre-

[8] Subsumers are retrieved using the OWL API [46].

[9] The corresponding Manchester OWL syntax is: A SubClassOf R some B.

sents the degree to which stem entities are indicative of the key, whilst the latter captures the difference between how indicative the stem entities are to the distractors, when compared to the key. The final difficulty measure is based on an average of these two measures.

Consider the following case-based medical MCQ (Q2), similar to those generated in [8]:[10]

> **Q2:** A 13-year-old female patient presents with Hemorrhage of urethra and Hematuria. What is the most likely diagnosis?
>
> A. Dysmenorrhea
> B. HIV infection
> C. Urethritis **(key)**

RSI's primary data source is an OWL ontology representation of Elsevier's Merged Medical Taxonomy (EMMeT), dubbed EMMeT-OWL [8, 47]. EMMeT content is maintained by a group of medical experts including physicians and nurses. The maintenance includes adding and removing relationship instances as well as manually adjusting rankings on the strength of the relationship instance. This is based on evidence from Elsevier content, which includes books, journals, and First Consult/Clinical Overviews.

RSI uses the EMMeT relation *hasClinicalFinding (hCF)*, which relates Diseases or Symptoms to Diseases, Symptoms, or ClinicalFindings, each of which can be used as a question's stem entities (in this case, the patient's symptoms). A fragment of the ontology from which Q2 was generated is listed in Figure 1:

1) $Urethritis \sqsubseteq \exists hCF.HemorrhageOfUrethra : 10$
2) $Urethritis \sqsubseteq \exists hCF.Hematuria : 10$
3) $Dysmenorrhea \sqsubseteq \exists hCF.HemorrhageOfUrethra : 6$
4) $Dysmenorrhea \sqsubseteq \exists hCF.Hematuria : 7$
5) $HIVinfection \sqsubseteq \exists hCF.HemorrhageOfUrethra : 6$
6) $HIVinfection \sqsubseteq \exists hCF.Hematuria : 6$

Fig. 1. A snippet of EMMeT-OWL used to provide data for Q2 where the annotations (: *n*) represent the strength of the *hCF* relation which range from *most common clinical finding* (10) to *rare clinical finding* (7), including a rank for a known non-relation *not a clinical finding* (6).

[10] A simple and modified version of a question generated in [8] is used for the sake of a non complex example.

Since the question is asking for the *most likely* diagnosis, the option entity[11] that has the strongest relation to the stem entities is the key.

Definition 3.1 (*stemInd*)**.** Let $\mathcal{S}$ be the set of symptoms and $\mathbf{k}$ be the key. Let *rank* be a function that returns the rank of any annotated axiom (i.e. axioms that are annotated with the strength of the relationship instance) and let *min* and *max* be functions that return the minimum and maximum ranks that a given relation can have (usually 7 (rare clinical finding) and 10 (most common clinical finding) respectively). Then *Stem indicativeness* (*stemInd*) is defined as follows:[12]

$$stemInd(\mathcal{S},\mathbf{k}) = 1 - \left(\frac{\sum_{s}^{\mathcal{S}} (rank(\mathbf{k} \sqsubseteq \exists hCF.s) - min(hCF))}{|\mathcal{S}| \times (\max(hCF) - \min(hCF))} \right)$$

The *Option entity difference measure* ($optDiff$) is defined in terms of each individual distractor difference ($disDiff$).

Definition 3.2 ($disDiff$)**.** Let $\mathcal{S}$ be the set of symptoms, $\mathbf{d}$ be a distractor and $\mathbf{k}$ be the key. Then $disDiff$, is defined as follows:

$$disDiff(\mathcal{S}, \mathbf{k}, \mathbf{d}) = \frac{n}{\left(\frac{\sum_{s}^{\mathcal{S}} (rank(\mathbf{k} \sqsubseteq \exists hCF.s) - \mathbf{d}_s) \times \mathbf{d}_s}{|\mathcal{S}|} \right)}$$

where n is the number of stem components (usually the histories and symptoms, however in this example, n = 1 since only symptoms are used) and $\mathbf{d}_s = rank(\mathbf{d} \sqsubseteq \exists hCF.s)$.

Using this measure allows $optDiff$ to be defined:

Definition 3.3 ($optDiff$)**.** Let $\mathcal{D}$ be the set of distractors. $optDiff$ is defined as follows:

$$optDiff(\mathcal{D}, \mathcal{S}, \mathbf{k}) = \sum_{d}^{\mathcal{D}} \left(disDiff(\mathcal{S}, \mathbf{k}, \mathbf{d})^2 \right)$$

The overall question difficulty is simply the average of $optDiff$ and *stemInd*.

[11] An option entity is a class name that was selected as an option for the question.

[12] Note that *hCF* relations used in the equations only serve as an example and it can be replaced by any relations associated with strength.

We demonstrate the use of RSI using Q2. *Stem indicativeness* equates to 0, showing that the stem is indicative of the key, and therefore has a low difficulty score. The more indicative the stem is of the key, the less difficult the question will be, and vice-versa. The *distractor difficulty* for Dysmenorrhea equates to 0.0444 whilst the difficulty of HIVinfection equates to 0.0416, indicating that Dysmenorrhea is more difficult than HIVinfection, or, it would be harder to eliminate Dysmenorrhea as a distractor compared to HIVinfection since the former has stronger relations to the stem entities than the latter. *Option entity difference* then equates to 0.0037, leading to an overall question difficulty of 0.00185. Suppose that instead of axioms 3 and 4 in Figure 1, the following axioms were present:

3) $Dysmenorrhea \sqsubseteq \exists hCF.HemorrhageOfUrethra$: **10**
4) $Dysmenorrhea \sqsubseteq \exists hCF.Hematuria$: **9**

The *distractor difficulty* for Dysmenorreah would instead equate to 0.2222, and thus the *option entity difference* would change to 0.0511. This demonsrates the effectiveness of RSI: the more similar the distractors are to the key, i.e., the more indicative the stem is to the distractors when compared to the key, the more difficult a question is considered, and vice-versa.

The questions studied and reviewed in [8] often use more complex stems. These include multiple types of stem entities such as: risk factors (via the *hasRiskFactor* relation); and patient demographics. The difficulty and similarity calculations are adjusted to account for additional stem entities and relations, where averages are usually taken over each calculation.

4. Method

To evaluate the performance of both experts and automated measures, we conducted two experiments: an expert review and a mock exam. Both experiments are described below.

4.1. Expert review

In a previous study [8], we carried out an expert review to evaluate the ontology-based approach we developed for generating medical case-based MCQs. As part of the review, experts rated the usefulness of generated questions (i.e. whether or not they are ready to use in an exam context) and predicted their difficulty. In what follows, we explain aspects of the review that are centered around expert prediction of difficulty.

4.1.1. Subjects

Fifteen experts were recruited to review the questions and were paid for their participation. The recruitment was conducted by our industrial partner, *Elsevier*, through emailing experts (i.e. authors and contributors knowledgeable in the specialities of interest) asking for participation or recommendation. Other colleagues/fellows recommended by experts were also contacted. Demographic information including education level, practical experience, teaching experience and exam construction experience were collected at the start of the review (Table 1).

Table 1
Demographic characteristics of domain experts.

Demographic characteristics	Number of experts
Speciality	
Internal medicine	5
Gastroenterology	4
Cardiology	5
Orthopedics	1
Level	
Resident	1
Generalist	7
Specialist	7
Experience as a practitioner	
None	2
Less than 1 year	0
1-3 years	4
3-6 years	3
More than 6 years	6
Teaching experience	
None	0
Less than 1 year	1
1-3 years	6
3-6 years	3
More than 6 years	5
Exam construction experience	
None	4
Less than 1 year	6
1-3 years	2
3-6 years	1
More than 6 years	2

4.1.2. Questions

The EMMeT-OWL ontology, which contains definitions of concepts such as diseases, clinical findings, drugs, symptoms, and risk factors, was utilised as a source for question generation. Four physician specialities (internal medicine, cardiology, orthope-

dics, and gastroenterology) were selected and a total of 3,407,493 case-based questions were automatically generated from these specialities. The generated questions belong to four templates: 'What is the most likely diagnosis?', 'What is the most likely clinical finding?', 'What is the drug of choice?', and 'What is the differential diagnosis?' (see Appendix B for examples). A stratified random sample of 435 questions was selected for expert review. Five stratifiers were used: speciality, question template, the number of distractors (key-distractor combinations in the case of differential diagnoses questions), the number of stem entities, and difficulty as predicted by the RSI measure. We aimed for an equal number of questions from each stratum but this was not possible due to the small number of questions in some strata. We obtained expert ratings for these 435 questions as described next.

4.1.3. Procedure

The expert review was conducted through a bespoke web-based questionnaire tool. Each expert reviewed approximately 30 questions belonging to their speciality.[13] To check agreement among experts, questions were reviewed by two experts. However, 119 questions were only reviewed by one expert due to the unavailability of another expert.

Each question was displayed individually and experts were asked to solve the displayed question without a time limit. To facilitate expert decision regarding question quality, the experts were shown the correct answer after answering the question and were shown an explanation of the incorrectness of the selected option(s) if they answered incorrectly. The following data about the performance of domain experts were collected:

- Selected answer(s)
- Score: Each single response question answered correctly is given one mark while an incorrect answer is awarded zero marks. With regards to questions with multiple responses (i.e., differential diagnosis questions), a mark for each correct answer is added to the final mark for each question and a mark of zero is given for fully incorrect answers.[14] The awarded mark is compared to the full mark of each question, which is equal to the number of correct options, in order to distinguish fully correct answers from partially correct answers.
- Time to solve: The time starts by displaying the question on the screen and ends by the expert clicking the 'submit' button.

After answering each question, experts were instructed to rate different aspects of the questions (e.g., usefulness, difficulty, and correctness of explanation)[15] while keeping in mind that the questions were targeting resident specialists or practising specialists. They started by rating the usefulness of the question using the following categories:

- Appropriate: The question is appropriate as a Board exam question; the level of knowledge required to answer the question is that of a resident specialist or practicing specialist.
- Inappropriate/no medical knowledge needed: The question can be answered correctly by people having little to no medical knowledge, (far) below the level of targeted exam audience.
- Inappropriate/guessable: The correct answer is guessable based on syntactic clues. For example, similar words between the stem and the key can clue examinees to the correct answer.
- Inappropriate/confusing: The syntax or terminology is not intelligible and/or the key does not logically follow from the question stem.
- Inappropriate/other: The question is inappropriate for other reasons.

They were then asked to classify the question as belonging to one of the following difficulty levels:

- Easy: More than 70% of examinees would be expected to answer the question correctly;
- Medium: 30% to 70% of examinees would be expected to answer the question correctly; or
- Difficult: Less than 30% of examinees would be expected to answer the question correctly.

They were also provided with an optional comment box for any additional information that they may have wanted to add. The main aim of obtaining expert pre-

[13] Question options were generated such that they belong to a shared speciality (determined using the EMMeT relation *hasSpeciality*).

[14] As the exam was experimental and no marks were displayed for participants, it made no sense to use negative marking. One could argue that participants could get the full mark on multiple response questions (only differential diagnosis questions in our sample) by selecting all options. We ensured that this was not the case by looking for such a pattern in the responses to differential diagnosis questions.

[15] An explanation for each of the options was displayed at this stage of the review.

diction is to compare it with student performance. Therefore, we did not collect their predictions for questions rated as inappropriate to use in an exam context, since these questions would not be used in the mock exam.

4.2. Mock exam

To obtain the empirical difficulty of the selected set of questions (i.e. percentage correct), we administered the questions to a cohort of residents. Details about the cohort, the questions, and the procedure are explained next.

4.2.1. Subjects

To recruit residents, experts who work in universities asked for volunteers. Twelve residents, with a mean age of 32 years (standard deviation = 2.3), participated in this experiment and were paid for their participation. Participants completed a demographics questionnaire, which asked them to indicate their age, sex, and practical experience (i.e. number of years working as a practitioner). Table 2 summarises their demographic information.

Table 2
Demographic characteristics of residents who took the mock exam.

Demographic characteristics	Number of residents
Sex	
Male	10
Female	2
Specialty	
Orthopedics	5
Internal medicine	4
Gastroenterology	2
Cardiology	1
Experience as a practitioner	
None	2
Less than 1 year	0
1-3 years	3
3-6 years	3
6-9 years	2
More than 9 years	2

4.2.2. Questions

We used disproportional stratified random sampling, aiming for equal group proportions whenever possible, to select questions from our sample space which consists of auto-generated questions rated as appropriate by at least one domain expert in the expert study (345 questions). We used this sampling technique to obtain a representative sample of each group in the population which was not possible using other sampling techniques (e.g. random sampling or proportional stratified sampling) due to the large difference in size between groups in the population. We decided to include questions that are rated as appropriate by at least one domain expert because one of the main reasons for disagreement on appropriateness was related to the difficulty of questions. The questions causing disagreement were found to be too easy or too difficult, and therefore inappropriate, by one of the reviewers, which was suggested by their comments. Including these questions in the mock exam allows further investigation of their difficulty.

We based stratification on four stratifiers: speciality, template, difficulty as predicted by our measure, and difficulty as predicted by the domain experts. Stratifying by speciality was necessary to ensure that residents from different specialities were tested on questions covering areas they are expected to be knowledgeable about. In addition, using templates as a stratifier allowed us to investigate the applicability of the measures to different question types and to investigate whether differences in difficulty can be attributed to the intrinsic nature of the templates themselves. Finally, stratifying based on our difficulty measure and the experts' predictions was used to allow investigation of the performance of these measures in predicting empirical difficulty.

The sample size for each speciality was determined considering a reasonable duration of testing (60-minute exam). This resulted in a sample of 231 questions in total to be administered to the residents involved in the experiment. The distribution of these questions is stated in Table 3. Variation in the number of questions across specialities was due to the unequal number of experts in each speciality and, therefore, the unequal number of reviewed questions. The selected questions were reviewed for linguistic issues and minimal edits were applied where necessary. For example, the stem 'A patient with a history of acetaminophen presents with ...' was edited to read: 'A patient who has used acetaminophen presents with ...'. This step was carried out to eliminate the effect of linguistic ambiguity on empirical difficulty.

4.2.3. Procedure

A web-based system was developed to administer the questions and collect performance data. Residents agreed to complete a 60-minute mock exam using their own machines and were assigned questions belong-

Table 3
Distribution of question sample per speciality and question type (Template 1 = What is the most likely diagnosis?, Template 2 = What is the drug of choice?, Template 3 = What is the most likely clinical finding?, and Template 4 = What is the differential diagnosis?).

Speciality	Template 1	Template 2	Template 3	Template 4	Total
Cardiology	41	7	8	7	63
Gastroenterology and hepatology	30	10	4	3	47
Internal medicine	53	14	8	17	92
Orthopaedics	17	9	3	0	29
Total	141	40	23	27	231

ing to their speciality, in addition to internal medicine questions.[16] For example, orthopaedic residents were assigned the 29 orthopaedic questions in addition to the 92 internal medicine questions. The questions were presented in a random order to avoid systematic bias resulting from position effects on difficulty. For example, participants suffering from fatigue which affects their performance at the end of the exam. Residents were not shown feedback indicating whether they answered the questions correctly or not. For each question attempted, the following data were collected:

- Selected answer(s);
- Score: the same as in the expert review; and
- Time to solve: the same as in the expert review.

4.2.4. Data analysis

A standard test theory analysis [48] was conducted for internal medicine questions that were administered to ten residents or more. The possible values that difficulty (percentage correct) can take and how they are interpreted is as follows:

- Easy: percentage correct >70%;
- Medium: 30% $\leqslant$ percentage correct $\leqslant$ 70%; and
- Difficult: percentage correct <30%.

The percentage correct was then compared to difficulty as predicted by the aforementioned measures. However, this type of item analysis was not possible for questions belonging to the other three specialities due to the low number of participants they had been administered to (1 to 5 residents at most).

We designed a new approach for analysing difficulty data for questions answered by less than ten participants. To investigate the relation between expert prediction and empirical difficulty, we grouped the questions based on expert prediction, resulting in three groups: easy, medium, and difficult questions according to the experts. We then computed the percentage correct for each group by dividing the total number of correct responses to all questions in the group by the total number of responses (correct and incorrect) to all questions in the group. One would expect the number of correct responses to difficult questions to be low and therefore the percentage correct for the difficult group to be low. A similar procedure was followed to investigate the relation between automated difficulty measures and percentage correct.

While studies concerned both with investigating expert ability in predicting question difficulty [21, 22] and with building difficulty models [7, 20] use the accuracy metric (Appendix A) for performance evaluation, we extend the evaluation by using approaches and metrics borrowed from the information retrieval and machine learning communities. The analysis was extended to include other metrics because accuracy does not reflect the performance of prediction when the distribution of classes (easy, medium, and difficult questions in our case) is not balanced. Another reason is that difficulty is an ordinal variable; it is therefore important to find how close or far away the prediction is from the empirical difficulty.

The following metrics, which are standard in classification problems, were used to compare measures for difficulty prediction: accuracy, precision, recall, F-score, and Kappa. We also used the evaluation metric, 'average relative error', which was used in the study reported in [20] for evaluating the performance of different machine learning models for predicting the difficulty of reading comprehension questions. We explain how we calculated these metrics in Appendix A.

Since different performance metrics focus on different aspects of the prediction, it is therefore essential to consider all of them, prioritising them based on the problem at hand, to allow comparison between the performances of the different methods. That is, which metrics do we care about in the case that different metrics give contradictory results? For example, it is usually the case that classification methods have a high

[16] Domain experts indicated that all residents are expected to have knowledge in internal medicine.

precision but low recall, or vice versa. Deciding on the superior method depends on the metric that is prioritised, whether it is higher precision or better recall. Our discussion of metrics is guided by the following characteristics of the problem of prediction of question difficulty:

- The distribution of difficulty levels is not balanced, with the difficult questions being the minority class. This is apparent from the distribution of difficulty levels in the test set in addition to the literature about MCQ examinations [for example, see: 26, 33, 49, 50].
- All of the classes are of importance, with little preference for good performance on difficult questions for two reasons: in addition to them being the minority class, appropriately difficult questions play an important role in discriminating between low- and high-information students.

As we were interested in performance for all difficulty levels, we averaged over the precision for each difficulty level, thereby penalising prediction methods that perform well on some of the difficulty levels. A similar calculation was performed for recall and F-score.

To answer the question of 'whether experts and automated measures do better than random guessing?', we compare their performance with the performance of the following three naive baselines:

- Random guesser which assigns difficulty levels arbitrarily;
- Weighted guesser which assigns difficulty levels according to their distribution in the test set; and
- Majority class classifier which assigns the most common difficulty level in the test set (medium) to all questions.

5. Results and Discussion

5.1. Residents' performance

Following the description of the difficulty levels in Section 4.2.4, 39.1% (n = 36) of the 92 internal medicine questions were easy, 44.6% (n = 41) were medium, and 16.3% (n = 15) were difficult. We consider this to be a good indicator of question suitability as a test set, since this distribution of difficulty levels is similar to the distribution of difficulty levels reported in analyses of real exams (for example, see [26, 49]). Residents' scores range from 58.49 to 77.65 with an average of 67.69 ($\pm$ 5.85) (see Table 4 details). Comparing these results to the results achieved by domain experts (range = 63.64 to 80.65 and mean = 72.09 $\pm$ 5.30) indicates that participants are adequately knowledgeable.

5.2. Performance of the measures

5.2.1. Is expert prediction a good proxy for difficulty?

Overall, the accuracy of expert prediction ranges between 46% and 53%. As Table 6 illustrates, the accuracy of experts is close (less than 10% variation in accuracy between experts). However, looking at other metrics, more variation in performance between- and within-experts can be seen. Of interest are the low values for precision, recall, and thus F-score on difficult questions compared to easy and medium questions,[17] which suggests that domain experts are less precise and complete in classifying difficult questions as compared to easy or medium questions. Given that domain experts who are involved in the experiment have teaching and exam construction experience, it is expected that they have more self-training (comparing one's own prediction with student performance) in predicting the difficulty of easy and medium questions since these represent a majority. The amount of self-training is a possible explanation of the difference in performance.

A point of interest is whether or not there are consistent patterns characterising expert prediction. An example of a pattern is experts having a tendency to underestimate or overestimate the difficulty of questions. Looking at the data, we found 44 questions for which experts overestimated the difficulty compared to 21 questions for which experts underestimated the difficulty. This suggests that experts tend to overestimate difficulty as opposed to underestimating it. We ran a further analysis of the relation between experts' performance on questions (getting the question right or wrong) and their prediction. The analysis aimed to answer two questions: 1) Is there a relation between experts' performance and their prediction accuracy? and 2) Is there a relation between experts' performance and overestimation or underestimation of difficulty? Regarding the first question, the data suggest that experts were more accurate in their prediction when they an-

[17] We performed a one way repeated measure ANOVA to compare the effect of actual difficulty of questions on F-scores achieved by experts. The F-score differed significantly between the different difficulty levels ($F(2,8) = 10.96$, $p < 0.05$).

Table 4

Residents' performance on the mock exam. Score is calculated as the percentage of the total possible scores.

Id	No. of questions	Normalised score (out of 100)	% of questions answered correctly
S1	155	77.65	74.19
S2	139	75.47	71.94
S3	92	73.40	69.57
S4	121	71.02	67.77
S5	92	69.73	65.22
S6	92	66.97	61.96
S7	121	65.94	61.98
S8	121	65.22	60.33
S9	92	64.22	57.61
S10	121	62.32	57.85
S11	103	61.86	56.31
S12	139	58.49	53.96
Average	115.67	67.69	63.22

Table 5

Resident performance (in percent) on questions belonging to different difficulty levels as predicted by: a) domain experts; b) relation strength indicativeness measure. Raw numbers are presented between parentheses.

			Correctness of responses (i.e. resident performance)						
			Incorrect		Partially correct		Correct		Total responses
Predicted difficulty	a)	Easy	17.37	(33)	3.68	(7)	78.95	(150)	(190)
		Medium	34.07	(46)	5.19	(7)	60.74	(82)	(135)
		Difficult	11.11	(2)	0	(0)	88.89	(16)	(18)
	b)	Easy	23.78	(39)	4.27	(7)	71.43	(118)	(164)
		Medium	23.23	(23)	0	(0)	76.77	(76)	(99)
		Difficult	28.57	(10)	0	(0)	71.43	(25)	(35)

swered the questions correctly. The prediction of 51% of questions solved correctly was accurate compared to 36% of questions solved incorrectly. Concerning the second question, experts overestimated the difficulty of 63% of the questions they solved correctly, compared to 81% of the questions they solved incorrectly, which hints at an increase in the percentage of overestimation when questions are solved incorrectly. However, the small number of observations, especially the observation about questions solved incorrectly, precludes making a strong conclusion about expert performance and prediction.

Given that expert prediction is considered as a major component of the evaluation framework for difficulty measures, which is apparent from relying heavily on expert prediction as a source of validation in multiple studies [7, 51, 52], the performance of domain experts was lower than anticipated. However, all experts outperform the three baseline classifiers in each of the prioritised metrics (i.e. accuracy, Kappa, average precision, average recall, and average F-score) except for the relative error metric, which is outperformed by the majority classifier. However, this is due to the majority of the questions in the test set belonging to the medium level and therefore the distance between any misclassified level and the actual difficulty level is minimal.

With regards to questions belonging to other specialities, a Fisher's exact test[18] was performed, comparing the frequency of responses to questions belonging to the three difficulty levels (Table 5), as predicted by domain experts. Since the P-value of the test (0.003) is less than the significance level (0.05), we can conclude that a dependency exists between expert prediction and resident performance. As Table 5 illustrates, easy questions have a higher percentage of correct responses and a lower percentage of incorrect responses as compared to medium questions. However, this was not the case for difficult questions. This result, along with the results obtained from internal medicine ques-

[18]The Fisher's exact test was selected because of the low frequencies observed in some cells (Table 5).

tions, indicates that expert precision is worst on difficult questions.

To summarise, the results indicate that experts moderately predicted question difficulty. The results are suggestive of an adverse effect of expert's performance on their accuracy and of experts' tendency to overestimate question difficulty.

5.2.2. *How well did the automated measures perform in comparison with guessing and in comparison with each other?*

While preliminary evaluations of the similarity measure [7] showed that it has potential for predicting question difficulty, the current evaluation shows that the accuracy of this measure on its own is lower than two of the baseline classifiers (Table 6). However, it is important to note that the similarity measure was evaluated in questions that have simple stems (i.e. consist of two concepts at most). Most of the questions in our dataset have more complex stems that contain two to five concepts. It is expected that the complexity of the stem contributes to the difficulty of the questions which is not captured by the similarity measure. This seems a plausible justification for its low performance. Taking into account the contribution of both stem and options into difficulty, as combined in the relation strength indicativeness measure (Section 3.2), increases the performance on all metrics except for recall on difficult questions as shown in Table 6. The performance of the relation strength indicativeness measure is also better than random and weighted guessers.

Another observation we made is that the similarity measure tends to overestimate the difficulty of questions. The predicted difficulty of 45 questions (48.91%) was higher than the empirical difficulty. On the other hand, the predicted difficulty of 14 questions (15.22%) was lower than the empirical difficulty. We observed a similar pattern for the relation strength indicativeness measure. We expect that cohort exposure to examined concepts, particularly when reviewing previous or sample exam papers, to moderate the effect of difficulty factors captured by the automated measures. Investigating the relation between cohort characteristics and difficulty remains an area for future research.

Performing Fisher's exact test on questions belonging to other specialties did not reveal a significant difference between the frequencies of correct and incorrect responses to questions belonging to different difficulty levels (as predicted by relation strength indicativeness measure). Results obtained from internal medicine indicate that the distance between predicted difficulty and empirical difficulty is higher in automatic prediction than in expert prediction. Classifying easy questions as difficult, and vice versa, is expected to have a strong impact on the frequency of correct and incorrect responses in each group (Table 5). Therefore, we attribute the failure in detecting a significant relation to the high value of the average relative error (Table 6).

5.2.3. *How well did the automated measures perform in comparison to domain experts?*

The performance of our measure is competitive compared with the performance of domain experts. Looking at Table 6, the relation strength indicativeness measure ranks higher than low-performing experts on all prioritised metrics except for the relative error metric. This indicates that difficulty levels assigned by domain experts are closer to the actual difficulty levels than the difficulty levels assigned by the automated measure. This can be explained by the ability of domain experts to recognise other features (e.g. linguistic features) that play a role in the difficulty of questions. For example, while the relation strength indicativeness measure predicts questions with indicative stems and low-similarity distractors to be easy, the language complexity of the questions or the use of rare concepts increases the difficulty of the question. In addition, experts have pedagogical content knowledge (i.e knowledge about challenging concepts that students find difficult to understand or have misconceptions about) which gives them an advantage over automated measures.

6. Methodological Reflection

The studies reported in this paper were pilot in nature. Conducting similar studies with a larger number of experts and student cohort would increase confidence in the results. To allow replication and extension of our work, the question set and the associated data were made available online.

While we have investigated expert performance on question difficulty prediction, our investigation was focused on medical questions and therefore the generalisability of these results to other domains is unknown. It is possible that other domains are more mature in the sense that pedagogical content knowledge is well-known. This, in turn, would improve expert prediction which would provide different results. In addition,

Table 6

Performance of different methods on difficulty prediction of internal medicine questions. The rank of each method among others is enclosed in parentheses and boldface indicates the method with the best performance in each metric (Q = questions, Acc. = accuracy, Rel. error = relative error, E = easy, M = medium, D = difficult, and Avg. = average).

					Precision				Recall				F-score			
Method	**#Q**	**Acc.**	**Rel. error**	**Kappa**	**E**	**M**	**D**	**Avg.**	**E**	**M**	**D**	**Avg.**	**E**	**M**	**D**	**Avg.**
Baseline																
Random	-	.33 (9)	.44 (7)	0 (8)	.33 (8)	.33 (8)	.33 (2)	.33 (7)	.33 (6)	.33 (8)	.33 (5)	.33 (6)	.36 (7)	.39 (7)	.22 (6)	.32 (8)
Weighted	-	.38 (7)	.38 (4)	0 (8)	.39 (7)	.45 (4)	.16 (7)	.33 (7)	.39 (5)	.45 (6)	.16 (6)	.33 (6)	.39 (6)	.44 (6)	.19 (7)	.34 (6)
Majority	-	.45 (6)	.28 (**1**)	0 (8)	Na	.45 (4)	Na	.15 (8)	0 (8)	1 (**1**)	0 (7)	.33 (6)	Na	.62 (**1**)	Na	Na
Experts																
Expert 1	22	.46 (5)	.39 (5)	.19 (2)	.80 (2)	.40 (7)	.29 (4)	.50 (3)	.40 (4)	.50 (4)	.50 (2)	.47 (2)	.53 (4)	.44 (6)	.36 (4)	.45 (3)
Expert 2	35	.46 (5)	.36 (3)	.16 (5)	1 (**1**)	.44 (5)	.25 (6)	.56 (**1**)	.42 (3)	.47 (5)	.50 (2)	.46 (3)	.59 (3)	.46 (5)	.33 (5)	.46 (2)
Expert 3	20	.50 (3)	.36 (3)	.18 (3)	.63 (4)	.63 (**1**)	0 (8)	.42 (5)	.71 (**1**)	.42 (7)	0 (7)	.38 (5)	.67 (**1**)	.50 (4)	0 (8)	.39 (4)
Expert 4	23	.52 (2)	.36 (3)	.05 (7)	.63 (4)	.40 (7)	0 (8)	.34 (6)	.71 (**1**)	.29 (9)	0 (7)	.33 (6)	.67 (**1**)	.33 (9)	0 (8)	.33(7)
Expert 5	30	.53 (**1**)	.30 (2)	.24 (**1**)	.67 (3)	.50 (3)	.40 (**1**)	.52 (2)	.55 (2)	.57 (2)	.40 (4)	.51 (**1**)	.60 (2)	.53 (3)	.40 (**1**)	.51 (**1**)
Automatic																
[8]	92	.47 (4)	.42 (6)	.17 (4)	.48 (5)	.54 (2)	.32 (3)	.45 (4)	.39 (5)	.54 (3)	.47 (3)	.47 (2)	.43 (5)	.54 (2)	.38 (3)	.45 (3)
[7]	92	.36 (8)	.50 (8)	.08 (6)	.46 (6)	.41 (6)	.27 (5)	.33 (7)	.28 (7)	.29 (9)	.73 (**1**)	.43 (4)	.35 (8)	.34 (8)	.39 (2)	.36 (5)

we find it worthwhile and interesting to look at domain experts' characteristics (e.g. teaching experience and exam construction experience) and how these contribute to their predictive performance. However, the amount of data that we have was limited for conducting such an analysis. Another factor that is expected to improve expert prediction, and that requires additional studies, is interaction and familiarity with the cohort to be tested.

Similarly, while we believe that research on medical, case-based questions has a major impact due to the heavy use of these question in medical education and in Board exams [53, 54], it would be interesting to investigate the utility of the automatic measures evaluated in this paper in predicting the difficulty of other types of questions.

Automatic measures for difficulty prediction are developed for the purpose of controlling the difficulty of automatically generated questions. This does not preclude the use of these measures for predicting the difficulty of human-authored questions (after parsing these questions). One of the limitations of the current study is that our test set consists of automatically generated questions only. These questions are very similar in terms of their linguistic structure. Difficulty prediction measures might perform worse on human-authored questions that are expected to be inherently more diverse in their linguistic structure. Another difference between auto-generated and human-authored questions is that, as mentioned earlier, the percentage of flawed questions is high among the latter type of questions. This is another expected source of performance variation between different measures on the two sets of questions. However, obtaining human-authored examination questions annotated with student performance was difficult because of exam security issues. Further studies that investigate the consistency of the results for human-authored questions are in high demand.

Another point that needs to be emphasised here is that, although the questions in the test set belong to four templates, these templates have different characteristics (e.g. the number of concepts in the stem and the number of correct answers). In addition, we varied the questions' characteristics within questions belonging to the same template. If the questions had been similar, we would have had no confidence in the generality of the test set and the generalisability of the results. However, at least the different characteristics of the question set increased our confidence in generalizing the results.

Finally, it is worth mentioning that the performance of both automatic measures investigated in this paper is heavily dependent on the completeness and correctness of the ontology in use. Thus, an interesting next step would be investigating the variation in performance when ontologies with different characteristics (e.g. size and expressivity) are used. Taking a different perspective, the performance of these measures can also be used as an indication of ontology quality.

7. Conclusion

To the best of our knowledge, this study is the first to compare the performance of domain experts, naive and automated methods for MCQ difficulty prediction. With respect to RQ1, experts moderately predicted

the difficulty of questions and were more accurate in predicting easy and medium questions compared to difficult questions. Regarding RQ2, the comparison shows that the relation strength indicativeness measure outperforms the similarity-based measure. Moreover, the former difficulty measure is of comparable performance to that of domain experts, who are heavily relied on in practice. We consider this as a major success since it can be used as an economical alternative. We believe that the ability of our model to explain its decisions (why a particular question is classified as belonging to a particular difficulty level), whether the decision is correct or not, is another point of strength. These justified decisions can make exam designers consider new aspects of questions, which in turn provide new insights into the difficulty and validity of questions.

However, investigating additional factors that can be used to predict the difficulty of both automatically generated and human-authored questions is still a subject of ongoing research. While doing this, the criteria presented in this study need to be considered as the minimum set of evaluation criteria.

Finally, while we made an attempt at creating an annotated question set that can be used for testing the performance of prediction methods, a larger question set is needed to cross-validate the results and gain more confidence in their consistency, as well as to provide statistical significance. In addition, a larger question set will allow the use of standard machine learning algorithms for building prediction models and investigating whether these models outperform the ontology-based measures compared in this study.

Acknowledgement

We would like to thank all participants of our experiments for their valuable contributions to our work.

Funding

This work was funded by an EPSRC grant (ref: EP/P511250/1) under an Institutional Sponsorship (2016) for The University of Manchester, along with a partial contribution from Elsevier. The funding acts as a secondment to an initial EPSRC grant (ref: EP/K503782/1) awarded as an Impact Acceleration Account (2016) for The University of Manchester.

Appendix A. Calculation of the evaluation metrics

Let $D = \{e, m, d\}$ be a set of difficulties (e = easy, m = medium, and d = difficult) and let Q be a set of questions $\{q_1, ..., q_n\}$. Let $actDif : Q \rightarrow D$ be a function over Q and D that returns the actual difficulty of a question (as derived from percentage correct) and let $preDif : Q \rightarrow D$ be a function over Q and D that returns the predicted difficulty of a question. Let $Q_{pc} \subseteq Q$ be the set of correctly classified questions, i.e. $q \in Q_{pc}$ if $actDif(q) = preDif(q)$. We can define accuracy as follows:

$$Accuracy = \frac{|Q_{pc}|}{|Q|}.$$

Possible values are between 0 and 1 with 1 indicating that all questions are correctly classified.

For $x \in D$, let $Q_x \subseteq Q$ be the set of questions with the difficulty level x s.t $q \in Q_x$ if $actDif(q) = x$ and let $Q_{px} \subseteq Q$ be the set of questions predicted as being x s.t $q \in Q_{px}$ if $preDif(q) = x$. Precision for Q_x is defined as follows:

$$Precision_{Q_x} = \frac{|Q_x \cap Q_{pc}|}{|Q_{px}|}.$$

The value ranges from 0 to 1 with higher values indicating that the classifier is less likely to identify questions as being x while they are actually not. Next, we define the recall on Q_x as:

$$Recall_{Q_x} = \frac{|Q_x \cap Q_{pc}|}{|Q_x|}.$$

The value ranges from 0 to 1 with a value of 1 indicating that the classifier has identified all questions in Q_x and a value of 0 indicating that it has missed all questions in Q_x. In what follow, we define the $F - score$ on Q_x:

$$F - score_{Q_x} = 2 * \frac{Precision_{Q_x} * Recall_{Q_x}}{Precision_{Q_x} + Recall_{Q_x}}.$$

$F - score_{Q_x}$ ranges between 0 and 1. The closer the $precision_{Q_x}$ and $recall_{Q_x}$ to each other, the greater the value.

Let *max* be a function that returns the maximum possible error where each $x \in D$ is associated with numerical values between [1,3], and maximum possible error is the difference between the maximum and minimum values associated with x (in this case, 3-1=2).

$$\textit{Average relative error} = \frac{\sum_{n=1}^{|Q|} preDif(q) - actDif(q)}{|Q| * max}.$$

The value ranges from 0 to 1. The closer the value to 0, the fewer errors are made by the classifier.

Finally, to define *kappa*, let p_o be the observed agreement and p_e be the agreement by chance. Then,

$$Kappa(Q, p_o, p_e) = \frac{p_o - p_e}{1 - p_e}.$$

The value is less than or equal to 1 with a value of 1 indicating a perfect agreement.

Appendix B. Example questions

Template 1: What is the most likely diagnosis?

Q3: A patient with a history of pericarditis presents with chest pain. What is the most likely diagnosis?

A. cardiac tamponade **(key)**
B. atrioventricular nodal re-entry tachycardia
C. primary pulmonary hypertension
D. end stage renal disease
E. hypertension

Template 2: What is the drug of choice?

Q4: A patient presents with atrioventricular nodal re-entry tachycardia. What is the drug of choice?

A. adenosine **(key)**
B. propafenone
C. flecainide
D. sotalol

Template 3: What is the most likely clinical finding?

Q5: A 75+ year old patient presents with aortic valve stenosis. What is the most likely clinical finding?

A. dyspnea **(key)**
B. Pulsus alternans
C. epistaxis
D. ejection click

Template 4: What is the differential diagnosis?

Q6: A infant patient presents with diarrhea and lethargy. What is the differential diagnosis?

A. pediatric gastroenteritis **(key)**
B. intussusception **(key)**
C. cystic fibrosis
D. intestinal volvulus
E. organophosphate toxicity

References

[1] J. Collins, Writing multiple-choice questions for continuing medical education activities and self-assessment modules, *RadioGraphics* **26**(2) (2006), 543–551. https://doi.org/10.1148/rg.262055145.

[2] J. Considine, M. Botti and S. Thomas, Design, format, validity and reliability of multiple choice questions for use in nursing research and education, *Collegian* **12**(1) (2005), 19–24. https://doi.org/10.1016/S1322-7696(08)60478-3.

[3] J.D. Wasserman and B.A. Bracken, *Psychometric characteristics of assessment procedures*, in: *Handbook of Psychology*, John Wiley and Sons, Inc., 2003. ISBN 9780471264385. http://dx.doi.org/10.1002/0471264385.wei1003.

[4] G. Kurdi, Generation and mining of medical, case-based multiple choice questions, PhD thesis, University of Manchester, 2020.

[5] A. Papasalouros, K. Kanaris and K. Kotis, Automatic generation of multiple choice questions from domain ontologies, in: *IADIS International Conference e-Learning*, 2008, pp. 427–434.

[6] M. Cubric and M. Tosic, Towards automatic generation of e-assessment using semantic web technologies, *International Journal of e-Assessment* **1**(1) (2011).

[7] T. Alsubait, B. Parsia and U. Sattler, Generating multiple choice questions from ontologies: lessons learnt, in: *OWLED*, 2014, pp. 73–84.

[8] J. Leo, G. Kurdi, N. Matentzoglu, B. Parsia, S. Forege, G. Donato and W. Dowling, Ontology-based generation of medical, multi-term MCQs, *International Journal of Artificial Intelligence in Education* (2019). https://doi.org/10.1007/s40593-018-00172-w.

[9] M. Al-yahya, Ontology-based multiple choice question generation, *The Scientific World Journal* **2014** (2014).

[10] N. Afzal, Automatic generation of multiple choice questions using surface-based semantic relations, *International Journal of Computational Linguistics (IJCL)* **6**(3) (2015), 26–44.

[11] M. Heilman, Automatic factual question generation from text, PhD thesis, Carnegie Mellon University, 2011.

[12] G. Kurdi, J. Leo, B. Parsia, U. Sattler and S. Al-Emari, A systematic review of automatic question generation for educational purposes, *International Journal of Artificial Intelligence in Education* (2019), In press.

[13] T. Alsubait, Ontology-based question generation, PhD thesis, University of Manchester, 2015.

[14] V. Crisp and R. Grayson, Modelling question difficulty in an A level physics examination, *Research Papers in Education* **28**(3) (2013), 346–372. https://doi.org/10.1080/02671522.2012.673005.

[15] J.D. Scheuneman, Y.V. Fan and S.G. Clyman, An investigation of the difficulty of computer-based case simulations, *Medical Education* **32**(2) (1998), 150–158. http://dx.doi.org/10.1046/j.1365-2923.1998.00193.x.

[16] V. Mesic and H. Muratovic, Identifying predictors of physics item difficulty: A linear regression approach, *Physical Review Special Topics - Physics Education Research* **7** (2011), 010110. https://link.aps.org/doi/10.1103/PhysRevSTPER.7.010110.

[17] J. Stiller, S. Hartmann, S. Mathesius, P. Straube, R. Tiemann, V. Nordmeier, D. Krüger and A.U. zu Belzen, Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty, *Assessment & Evaluation in Higher Education* **41**(5) (2016), 721–732. https://doi.org/10.1080/02602938.2016.1164830.

[18] R.F. Boldt, GRE analytical reasoning item statistics prediction study, Technical Report, Educational testing services. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1998.tb01786.x.

[19] M.K. Enright and K. Sheehan, Modeling the difficulty of quantitative reasoning items: Implications for item generation, in: *Item Generation for Test Development*, S.H. Irvine and P.C. Kyllonen, eds, Routledge, 2002, pp. 129–157.

[20] D. Hutzler, E. David, M. Avigal and R. Azoulay, Learning methods for rating the difficulty of reading comprehension questions, in: *2014 IEEE International Conference on Software Science, Technology and Engineering*, 2014, pp. 54–62. http://dx.doi.org/10.1109/SWSTE.2014.16.

[21] G. van de Watering and J. van der Rijt, Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items, *Educational Research Review* **1**(2) (2006), 133–147. http://www.sciencedirect.com/science/article/pii/S1747938X06000236.

[22] J.D. Kibble and T. Johnson, Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations?, *Advances in Physiology Education* **35**(4) (2011), 396–401. https://doi.org/10.1152/advan.00062.2011.

[23] S.M. Downing, Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation, *Advances in Health Sciences Education* **7**(3) (2002), 235–241. https://doi.org/10.1023/A:1021112514626.

[24] J.C. Masters, B.S. Hulsmeyer, M.E. Pike, K. Leichty, M.T. Miller and A.L. Verst, Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education, *Journal of Nursing Education* **40**(1) (2001), 25–32.

[25] R. Nedeau-Cayo, D. Laughlin, L. Rus and J. Hall, Assessment of item-writing flaws in multiple-choice questions, *Journal for Nurses in Professional Development* **29**(2) (2013), 52–57.

[26] B.R. Rush, D.C. Rankin and B.J. White, The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value, *BMC Medical Education* **16**(1) (2016), 250. https://doi.org/10.1186/s12909-016-0773-3.

[27] B.M. Hijji, Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study, *Journal of Nursing Education* **56**(8) (2017), 490–496.

[28] J. Pais, A. Silva, B. Guimarães, A. Povo, E. Coelho, F. Silva-Pereira, I. Lourinho, M.A. Ferreira and M. Severo, Do item-writing flaws reduce examinations psychometric quality?, *BMC Research Notes* **9**(1) (2016), 399. https://doi.org/10.1186/s13104-016-2202-4.

[29] Y. Sarin, M. Khurana, M. Natu, A.G. Thomas and T. Singh, Item analysis of published MCQs, *Indian pediatrics* **35**(11) (1998), 1103–1105.

[30] M. Tarrant, J. Ware and A.M. Mohammed, An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis, *BMC Medical Education* **9**(1) (2009), 40. https://doi.org/10.1186/1472-6920-9-40.

[31] J. Ware and T. Vik, Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations, *Medical Teacher* **31**(3) (2009), 238–243. https://doi.org/10.1080/01421590802155597.

[32] M.R. Hingorjo and F. Jaleel, Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency, *JPMA-Journal of the Pakistan Medical Association* **62**(2) (2012), 142–147.

[33] S.K. Namdeo and S.D. Rout, Assessment of functional and nonfunctional distracter in an item analysis, *International Journal of Contemporary Medical Research* **3**(7) (2016), 1891–1893.

[34] K. Gautam, I. Gupta and K. Chandramouli, Conceptual extraction of questions from Wikipedia, in: *ET LACNEM*, 2013.

[35] A. Singh Bhatia, M. Kirti and S.K. Saha, Automatic generation of multiple choice questions using Wikipedia, in: *Pattern Recognition and Machine Intelligence*, P. Maji, A. Ghosh, M.N. Murty, K. Ghosh and S.K. Pal, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 733–738. ISBN 978-3-642-45062-4.

[36] M. Liu, R.A. Calvo, A. Aditomo and L.A. Pizzato, Using Wikipedia and conceptual graph structures to generate questions for academic writing support, *IEEE Transactions on Learning Technologies* **5**(3) (2012), 251–263. https://doi.org/10.1109/TLT.2012.5.

[37] S. Williams, Generating mathematical word problems, in: *The Association for the Advancement of Artificial Intelligence AAAI Fall Symposium: Question Generation*, 2011, pp. 61–64.

[38] L. Kovacs and G. Szeman, Complexity-based generation of multi-choice tests in AQG systems, in: *the IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, 2013, pp. 399–402. https://doi.org/10.1109/CogInfoCom.2013.6719278.

[39] E.V. Vinu and P.S. Kumar, A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption, *Web Semantics: Science, Services and Agents on the World Wide Web* **34** (2015), 40–54. http://www.sciencedirect.com/science/article/pii/S1570826815000475.

[40] E.V. Vinu and P.S. Kumar, Difficulty-level Modeling of Ontology-based Factual Questions, *Semantic Web Journal* (2018), In press.

[41] B.S. Bloom, M.D. Engelhart, E.J. Furst, W.H. Hill and D.R. Krathwohl, *Taxonomy of educational objectives, handbook I: The cognitive domain*, Vol. 19, New York: David McKay Co Inc, 1956.

[42] J.P.W. Cunnington, G.R. Norman, J.M. Blake, W.D. Dauphinee and D.E. Blackmore, *Applying learning taxonomies to test items: Is a fact an artifact?*, in: *Advances in Medical Education*, A.J.J.A. Scherpbier, C.P.M. van der Vleuten, J.J. Rethans and A.F.W. van der Steeg, eds, Springer Netherlands, Dordrecht, 1997, pp. 139–142. ISBN 978-94-011-4886-3.

[43] M.E. Abdalla, A.M. Gaffar and R.A. Suliman, *Constructing A-type multiple choice questions (MCQs): step by step manual*, 2011.

[44] L.W.T. Schuwirth, M.M. Verheggen, C.P.M. Van Der Vleuten, H.P.A. Boshuizen and G.J. Dinant, Do short cases elicit different thinking processes than factual knowledge questions do?, *Medical Education* **35**(4) (2001), 348–356. http://dx.doi.org/10.1046/j.1365-2923.2001.00771.x.

[45] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bull Soc Vaudoise Sci Nat* **37** (1901), 547–579.

[46] M. Horridge and S. Bechhofer, The OWL API: A java API for OWL ontologies, *Semantic Web* **2**(1) (2011), 11–21.

[47] B. Parsia, T. Alsubait, J. Leo, V. Malaisé, S. Forge, M. Gregory and A. Allen, *Lifting EMMeT to OWL getting the most from SKOS*, in: *Ontology Engineering: 12th International Experiences and Directions Workshop on OWL, OWLED 2015, co-located with ISWC 2015, Bethlehem, PA, USA, October 9-10, 2015, Revised Selected Papers*, V. Tamma, M. Dragoni, R. Gonçalves and A. Ławrynowicz, eds, Springer International Publishing, Cham, 2016, pp. 69–80. ISBN 978-3-319-33245-1. https://doi.org/10.1007/978-3-319-33245-1_7.

[48] L. Crocker and J. Algina, *Introduction to classical and modern test theory.*, ERIC, 1986.

[49] M. Mukhopadhyay, K. Bhowmick, S. Chakraborty, D. Roy, P.K. Sen and I. Chakraborty, Evaluation of MCQs for judgement of higher levels of cognitive learning, *Gomal Journal of Medical Sciences* **8**(2) (2010).

[50] B.S. Malau-Aduli and C. Zimitat, Peer review improves the quality of MCQ examinations, *Assessment & Evaluation in Higher Education* **37**(8) (2012), 919–931. https://doi.org/10.1080/02602938.2011.586991.

[51] F.-l. Lee and R. Heyworth, Problem complexity: A measure of problem difficulty in algebra by using computer, *Education Journal* **28**(1) (2000), 85–108.

[52] S. Banerjee, N.J. Rao and C. Ramanathan, Rubrics for assessment item difficulty in engineering courses, in: *2015 IEEE Frontiers in Education Conference (FIE)*, 2015, pp. 1–8. http://dx.doi.org/10.1109/FIE.2015.7344299.

[53] T. Freiwald, M. Salimi, E. Khaljani and S. Harendza, Pattern recognition as a concept for multiple-choice questions in a national licensing exam, *BMC Medical Education* **14**(1) (2014), 232. https://doi.org/10.1186/1472-6920-14-232.

[54] M.C. Rodríguez-Díez, M. Alegre, N. Díez, L. Arbea and M. Ferrer, Technical flaws in multiple-choice questions in the access exam to medical specialties (“examen MIR") in Spain (2009–2013), *BMC medical education* **16**(1) (2016), 47.

[55] S.M. Case and D.B. Swanson, Extended matching items: A practical alternative to free response questions, *Teaching and Learning in Medicine* **5**(2) (1993), 107–115. https://doi.org/10.1080/10401339309539601.

[56] E. Ibrahem, Automated MCQ generation: An ontology-based approach for Java knowledge assessment and ontology validation, Master’s thesis, The University of Manchester, 2016.

[57] E.V. Vinu and P.S. Kumar, Automated generation of assessment tests from domain ontologies, *Semantic Web* **8**(6) (2017), 1023–1047. "https://content.iospress.com/articles/semantic-web/sw252".

[58] S. Gajjar, R. Sharma, P. Kumar and M. Rana, Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat, *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine* **39**(1) (2014), 17–20. http://dx.doi.org/10.4103/0970-0218.126347.