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

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, 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. 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 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 or all of the students solve them correctly. Finally, the difficulty level of the 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).

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

1This is based on a recent systematic review we conducted on difficulty prediction and on a survey of studies investigating MCQ examination quality.
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’ evaluation, which is usually the case in small-scale exams.

With recent advances in automated procedures for generating questions [4–10], 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 experts’ efforts in sorting through, and trying to predict the difficulty of, a large number of questions. Finally, good 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, 11–14] 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 [15–17] are highly domain-specific which limit their utility. However, in a prior work [6, 7], 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 extend our work in [6, 7] by evaluating the performance of the previously proposed measures. Specifically, we compare their prediction, along with expert prediction (collected in our previous study [7]), against a gold standard of student performance (collected in this study). 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?
- How well does each method perform in comparison to domain experts?

We have 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 [18], we found that the difficulty of case-based MCQs was moderately predicted by domain experts (average accuracy = 49%). We also found the automated measure developed in [7] 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 difficulty prediction measures that show, by example, the minimum set of criteria that need to be considered in evaluating the performance of similar measures;
- 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.2

2 The question set is available upon request from the authors.

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) [19]. Another popular type of MCQs, known
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 [20–24] found more than 50% of investigated MCQs to contain at least one item writing flaw. Other studies [25–27] report that the percentage of MCQs with all their distractors being considered functional 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 [4–8]. 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 text, the process of finding good distractors is easier. As an example, consider the question ‘Which city is located in UK?’, generated from a Wikipedia 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 [33].

One point worth mentioning is that underlying difficulty models are not part of most existing question generation approaches. According to Alsubait [33], apart from the similarity-based approach (outlined in Section 3.1), only two question generation approaches [34, 35] take into account generating questions with controlled difficulty but without providing experimental evaluation of the performance of difficulty prediction. The automatic measures compared in this study represent existing domain nonspecific measures of MCQ difficulty. Other measures are either variants of the similarity approach [36], designed for questions with other response formats, or categorised by being domain- or question-specific [15, 16, 34, 37].

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, with the majority of generated questions in [4, 6, 8] testing recall of information, and structure, with generated stems in [4, 6, 8] containing at most two concepts. In a recent study [7], we have 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 [39–41]. 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 specialties. More details on the set of generated questions will be given in Section 4.2.2.

3. Competing measures

The target of difficulty prediction is to assign difficulty levels (easy, medium, 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 has been proposed in [6], in which the similarity between the key and

---

3 Violations of best practices as suggested in MCQ-writing guidelines.
4 Functional distractors are those selected by at least 5% of examinees [26, 28, 29].
5 Wikipedia has been used as a source for question generation by [30–32].

---

6 The mental process involved in question-solving as described in Bloom’s taxonomy [38], a popular classification of cognitive levels.
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 [33]. 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. [6] developed a similarity measure that is based on Jaccard similarity [42] and intended to be used with ontologies. The similarity measure is defined as follows:

\[
\text{sim}(k, d) = \frac{\text{Com}(k, d)}{\text{Union}(k, d)}
\]

where \(\text{Com}(k, d)\) is the number of common subsumers between the key \(k\) and a distractor \(d\), and \(\text{Union}(k, d)\) is the number of all subsumers of both \(k\) and \(d\).\(^7\) The overall difficulty of the question is then defined as the average similarity between the key and distractors.

Preliminary studies have shown that the similarity measure has a good difficulty prediction [6, 33]. 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 a stem plays is a major influencer on the difficulty of case-based questions that are characterized 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. common clinical finding and rare clinical finding). These limitations motivate us to develop the new difficulty measure described below.

3.2. Relation Strength Indicativeness

A new measure of question difficulty was introduced in [7] which 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 \(A \sqsubseteq \exists R.B\).\(^8\) 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 represents 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 [7]:\(^9\)

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 [7, 43]. RSI uses the EMMeT relation hasClinicalFinding (hCF), which relates Diseases or Symptoms to Diseases, Symptoms

---

\(^7\) Different ways of counting subsumers have been defined in [33].

\(^8\) The corresponding Manchester OWL syntax is: A SubClassOf R some B.

\(^9\) A simple and modified version of a question generated in [7] is used for the sake of a non complex example.
or ClinicalFindings, each of which can be used as a question’s stem entities (in this case, the patient symptoms). A fragment of the ontology from which the question was generated is listed in Figure 1:

1) Urethritis  
   \( \exists hCF. HemorrhageOfUrethra : 10 \)  
2) Urethritis \( \sqsubseteq \exists hCF. Hematuria : 10 \)  
3) Dysmenorrhea  
   \( \exists hCF. HemorrhageOfUrethra : 6 \)  
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 EMMcT-OWL used to provide data for Q2 where the annotations (\( \prec \) 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).

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

**Definition 3.1 \((sInd)\).** Let \( S \) be the set of symptoms and \( k \) be the key. Let \( rank \) be a function that returns the rank of any annotated axiom 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* \((sInd)\) is defined as follows:

\[
sInd(S, k) = 1 - \left( \frac{\sum(rank(k \sqsubseteq \exists hCF.s) - \min(hCF))}{|S| \times (\max(hCF) - \min(hCF))} \right)
\]

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

**Definition 3.2 \((disDiff)\).** Let \( S \) be the set of symptoms, \( d \) be a distractor and \( k \) be the key. Then \( disDiff \),

\[
isDiff(S, k, d) = \frac{1}{\sum(rank(k \sqsubseteq \exists hCF.s) - d_s) \times d_s}
\]

where 1 is the number of stem components (usually the histories and symptoms, however in this example only symptoms are used) and \( d_s = rank(d \sqsubseteq \exists hCF.s) \).

Using this measure allows \( optDiff \) to be defined:

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

\[
optDiff(D, S, k) = \sum_d (disDiff(S, k, d)^2)
\]

The overall question difficulty is simply the average of \( optDiff \) and \( stemInd \).

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 Dysmenorrhea would instead equate to 0.2222, and thus the option entity difference would change to 0.0511. This demonstrates 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 [7] 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 [7], 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 (see Table 1).

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 specialties (internal medicine, cardiology, orthopedics, and gastroenterology) were selected and a total of 3,407,493 case-based questions were automatically generated from these specialties. 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: specialty, 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 to target an equal number of questions from each strata 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 specialty. 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 if they answered correctly, and were shown an

---

Table 1

Demographic characteristics of domain experts.

<table>
<thead>
<tr>
<th>Demographic characteristics</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty</td>
<td></td>
</tr>
<tr>
<td>Internal medicine</td>
<td>5</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>4</td>
</tr>
<tr>
<td>Cardiology</td>
<td>5</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>1</td>
</tr>
<tr>
<td>Level</td>
<td></td>
</tr>
<tr>
<td>Resident</td>
<td>1</td>
</tr>
<tr>
<td>Generalist</td>
<td>7</td>
</tr>
<tr>
<td>Specialist</td>
<td>7</td>
</tr>
<tr>
<td>Experience as a practitioner</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>2</td>
</tr>
<tr>
<td>Less than 1 year</td>
<td>0</td>
</tr>
<tr>
<td>1-3 years</td>
<td>4</td>
</tr>
<tr>
<td>3-6 years</td>
<td>3</td>
</tr>
<tr>
<td>More than 6 years</td>
<td>6</td>
</tr>
<tr>
<td>Teaching experience</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td>Less than 1 year</td>
<td>1</td>
</tr>
<tr>
<td>1-3 years</td>
<td>6</td>
</tr>
<tr>
<td>3-6 years</td>
<td>3</td>
</tr>
<tr>
<td>More than 6 years</td>
<td>5</td>
</tr>
<tr>
<td>Exam construction experience</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>4</td>
</tr>
<tr>
<td>Less than 1 year</td>
<td>6</td>
</tr>
<tr>
<td>1-3 years</td>
<td>2</td>
</tr>
<tr>
<td>3-6 years</td>
<td>1</td>
</tr>
<tr>
<td>More than 6 years</td>
<td>2</td>
</tr>
</tbody>
</table>
They started by rating the usefulness of the question targeting resident specialists or practising specialists. The following data about explanation of the incorrectness of the selected option 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. 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 from 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) 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;
- **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 prediction 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.

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

<table>
<thead>
<tr>
<th>Demographic characteristics</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>10</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
</tr>
<tr>
<td>Orthopedics</td>
<td>5</td>
</tr>
<tr>
<td>Internal medicine</td>
<td>4</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>2</td>
</tr>
<tr>
<td>Cardiology</td>
<td>1</td>
</tr>
<tr>
<td>Experience as a practitioner</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>2</td>
</tr>
<tr>
<td>Less than 1 year</td>
<td>0</td>
</tr>
<tr>
<td>1-3 years</td>
<td>3</td>
</tr>
<tr>
<td>3-6 years</td>
<td>3</td>
</tr>
<tr>
<td>6-9 years</td>
<td>2</td>
</tr>
<tr>
<td>More than 9 years</td>
<td>2</td>
</tr>
</tbody>
</table>

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: specialty, template, difficulty as predicted by our measure, and difficulty as predicted by the domain experts. Stratifying by specialty was necessary to ensure that residents from different specialties 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 specialty 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 specialties was due to the unequal number of experts in each specialty 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 belonging to their specialty, in addition to internal medicine questions. For example, orthopedic residents were assigned the 29 orthopedic questions in addition to the 92 internal medicine questions. The questions were presented in a random order in 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;
- Time to solve: the same as in the expert review.

4.2.4. Data analysis
A standard test theory analysis [44] 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% ≤ percentage correct ≤ 70%
- 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 specialties 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 parti-

---

13 Domain experts indicated that all residents are expected to have knowledge in internal medicine.
Distribution of question sample per specialty 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?).

<table>
<thead>
<tr>
<th>Specialty</th>
<th>Template 1</th>
<th>Template 2</th>
<th>Template 3</th>
<th>Template 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiology</td>
<td>41</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>63</td>
</tr>
<tr>
<td>Gastroenterology and hepatology</td>
<td>30</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>47</td>
</tr>
<tr>
<td>Internal medicine</td>
<td>53</td>
<td>14</td>
<td>8</td>
<td>17</td>
<td>92</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>17</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>141</strong></td>
<td><strong>40</strong></td>
<td><strong>23</strong></td>
<td><strong>27</strong></td>
<td><strong>231</strong></td>
</tr>
</tbody>
</table>

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 [18, 45] and with building difficulty models [6, 17] 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, and 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 has been used in the study reported in [17] 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 precision but low recall, or vice versa. The winning 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: 22, 29, 46, 47].
- 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 methods as baselines:

- random guesser which assigns difficulty levels arbitrarily;
- weighted guesser which assigns difficulty levels according to their distribution in the test set;
- 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 (as examples, see [22, 46]). Residents’ scores range from 58.49 to 77.65 with an average of 67.69 (± 5.85) (see Table 4 for details). Comparing these results to the results achieved by domain experts (range = 63.64 to 80.65 and mean = 72.09 ± 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 one can see from Table 6, 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,14 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 under- 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? 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 answered the questions correctly. The prediction of 51% of questions solved correctly was accurate compared to 36% of question 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 [6, 48, 49], 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 specialties, a Fisher’s exact test15 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 student performance. As can be seen in Ta-

---

14 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).

15 The Fisher’s exact test was selected because of the low frequencies observed in some cells (Table 5).
Table 4
Residents’ performance on the mock exam. Score is calculated as the percentage of the total possible scores.

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

Table 5
Frequency of responses to questions belonging to difficulty levels predicted by: a) domain experts; b) relation strength indicativeness measure. Raw numbers are presented between parentheses.

<table>
<thead>
<tr>
<th>Correctness of responses</th>
<th>Percentage correct</th>
<th>Total responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>Partially correct</td>
<td>Correct</td>
</tr>
<tr>
<td>a) Easy</td>
<td>17.37 (33)</td>
<td>3.68 (7)</td>
</tr>
<tr>
<td>Medium</td>
<td>34.07 (46)</td>
<td>5.19 (7)</td>
</tr>
<tr>
<td>Difficult</td>
<td>11.11 (2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>b) Easy</td>
<td>23.78 (39)</td>
<td>4.27 (7)</td>
</tr>
<tr>
<td>Medium</td>
<td>23.23 (23)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Difficult</td>
<td>28.57 (10)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Table 5, 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 questions, indicates that expert precision was 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 [6] show 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 has been 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 these complex stems contribute to the difficulty of the questions which is not captured by the similarity measure. This seems a plausible justification for its low performance. Combining the similarity measure with stem indicativeness, as explained in Section 3.2, increases the performance on all metrics except for recall on difficult questions as can be seen in Table 6. The performance of the relation strength indicativeness measure is also better than random and weighted guessers.

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

Table 6

<table>
<thead>
<tr>
<th>Method</th>
<th>#q</th>
<th>Acc.</th>
<th>Rel. error</th>
<th>Kappa</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.53 (2)</td>
<td>.33 (8)</td>
<td>.57 (2)</td>
</tr>
<tr>
<td>Weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.57 (2)</td>
<td>.33 (8)</td>
<td>.40 (4)</td>
</tr>
<tr>
<td>Majority</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.40 (4)</td>
<td>.50 (4)</td>
<td>.50 (2)</td>
</tr>
<tr>
<td>E M D Avg.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.47 (4)</td>
<td>.47 (2)</td>
<td>.33 (7)</td>
</tr>
</tbody>
</table>

Performance of our measure was competitive in comparison to domain experts. 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 the 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

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, 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 [50, 51], it would be interesting to investigate the utility of automatic measures 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 hand-written 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 hand-crafted questions that are expected to be inherently more diverse in their linguistic structure. Another difference between auto-generated and hand-crafted 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 for the two sets of questions. However, obtaining hand-crafted examination questions annotated with student performance was difficult because of exam security issues. Further studies that investigate the consistency of the results for hand-crafted 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., number of concepts in the stem, number of keys). 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 generalising 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 used ontology. Thus, an interesting next step would be investigating the variation of 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 and 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 new ontology-based 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 about the difficulty and validity of questions.

However, investigating additional factors that can be used to predict the difficulty of both automatically generated questions and hand-written 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 $\{q_1, \ldots, q_n\}$. Let $actDiff : Q \to D$ be a function over $Q$ and $D$ that returns the actual difficulty of a question (as derived from percentage correct) and let $preDiff : Q \to D$ 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 $actDiff(q) = preDiff(q)$. 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 $actDiff(q) = x$ and let $Q_{px} \subseteq Q$ be the set of questions predicted as being $x$ s.t. $q \in Q_{px}$ if $preDiff(q) = x$. 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 a perfect agreement.

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).

$$Average\ relative\ error = \frac{\sum_{x=0}^{n} \frac{|Q_{pc} \cap preDiff(q) - actDiff(q)|}{|Q|} \cdot max}{n}.$$ 

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. Atrialventricular 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

[27] 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. doi:10.1080/01421590802155597.


3. Reviews

Insert review response letter

3.1. Reviewer 1 response

3.1.1. Comment
In general, this is a well written paper. However, I have concerns regarding the focus of the paper w.r.t the audiences of this journal. The paper is felt to an extended version of 2 of their previous works [6-7]. Mainly the experimental section was extended. No major contribution on the conceptual-side, however, this paper, especially due to their detailed experimental study, would definitely be useful for communities who work in the areas related to ontology-based question generation and automated educational assessment generation.

Response
Thanks for your comment. Indeed, our focus was on the experimental framework and we have clarified the relation between this paper and our previous paper in the introduction.

3.1.2. Comment
The title of the paper itself says: "Experts Vs Automata" why this study is so important especially when there are previous studies both in the educational and physiological science communities that experts’ predication of difficulty level is not so precise?

Response
Thanks for raising this point. While there are some studies about expert prediction of difficulty, none of these studies has investigated the prediction of medical experts or, more specifically, their prediction of the difficulty of case-based questions. Since these questions require a specific type of reasoning, it is not clear whether the results of previous studies are transferable to these questions.

In addition, while some studies showed that expert prediction is not precise for individual items, their estimation is still useful in determining the overall difficulty of exams (see “4.5.1. The implications of teachers’ estimations of difficulty” in [18]). As such, we believe that understanding whether automatic measures can approximate expert prediction is an important research area that has not been investigated.

However, we agree that the title is perhaps misleading. We have removed the first part of the title "Experts vs. Automata;" because the main focus of the paper is on the performance of automatic measures.

3.1.3. Comment
This paper compares two existing ontology-based measures for predicting difficulty-level w.r.t to experts’ predication. I wonder, if these were the only possible measures. Even though the authors mentions about machine learning based methods, a comparison with such methods would have been given this manuscript a better reach. However, I understand that this may not be possible if the domain specific data-set is not available.

Response
We agree with you that this would be interesting. However, the machine learning measures we found (through systematic reviews) rely on manual extraction of features. In addition, these features are also not well defined which make the reliability questionable. Both issues preclude including these measures in the comparison.
3.1.4. Comment: The authors could clearly say that this work is an extension of their previous work [6-7] and they may rewrite the new or main contributions accordingly. Also, I feel that Paper 7 is cited too many times in the current manuscript. At one point, I felt that there is too much overlap between these two papers.

Response: Thank you. As per your suggestion, we have clarified the relation between this paper and our previous papers in the introduction. We have also made this paper more self-contained and removed unnecessary citations to our previous work (for example, see Section 4.1.2 and 4.1.3).

3.1.5. Comment: *Introduction* RQ1&2 questions appear to be ambiguous while using *compared to* – Rewrite!

Response: As per your suggestion, we have rewritten these questions.

3.1.6. Comment: We found that MCQ difficulty was moderately predicated by domain experts – This results is previously well studied and reported in earlier works (mainly in educational communities). The author should cite such works instead, and may be they can report their observations for the current domain under consideration.

Response: We have added a citation to a review on expert prediction of difficulty and connected our results with the previous work (see the introductions). Also, see our response to the point 3.1.2.

3.1.7. Comment: In the main contributions: 1. It is mentioned that the new ontologies based measures are proposed. However, I see that the measures were already published in their previous works.

Response: Thanks for your comments. We have removed the contribution you mentioned from our contribution list.

3.1.8. Comment: 2. A fairly large question set – is it available online?

Response: We have pointed out that the question set is available upon request from the authors.

3.1.9. Comment: 3. The experimental setup that should be followed for evaluating ontology-based metrics would be a main contribution.

Response: Thanks for your comment.

3.1.10. Comment: *Methods* More details about each type of question and... Please add a reference. As I mentioned above, please avoid too many references to previous work; it is really affecting the flow and readability of the paper. Make it more self-contained.

Response: As per your suggestion, we have added more details about the questions (see Appendix B) and made the paper more self-contained.
3.1.11.

Comment 4.1.3. Procedure* To check agreement among experts, questions were reviewed by two experts whenever possible – What would be the criteria for a tie-break?

Response We did not attempt to resolve the disagreement. We have explained reasons for disagreement in our previous study [7] and cases where there was disagreement was analysed separately. For example, regarding the usefulness of questions, we have reported on the number of questions evaluated and thought to be appropriate by one reviewer and the number of questions evaluated and thought to be appropriate by two reviewers separately. We did similarly for the number of questions evaluated by two reviewers but thought to be appropriate by one reviewer.

3.1.12.

Comment 4.2.1. Subject* Are their any assumption regarding the level of expertise of these residents?

Response The participants have finished medical school and are thought to have the required knowledge to answer the questions. While it would be interesting to analyse the relation between their performance and their level of expertise (i.e. experience as a practitioner), the data we have is small for such comparison. Unfortunately, we were not able to recruit more participants.

3.1.13.

Comment What is the significance of their "practical experience" for assigning their expert-levels: say, as high, medium, low?

Response Please see response to 3.1.12

3.1.14.

Comment Please detail about the reasons for choosing the mentioned demographic characteristics of the residents. This details is importation because based on the definitions of the various classes of difficulty-level mentioned in Section 4.1.3, there is an explicit assumption that expertise levels of the subjects under consideration should be a uniform distribution.

Response Please see response to 3.1.12.

3.1.15.


Response Thank you for this reference.

3.1.16.

Comment 5.2.1. Is the expert predication a good proxy for difficulty?* the data suggest that experts were more accurate in their predication when they answers the questions correctly – From the perspective of forming a "best-practice" for collecting experts observation, this a good observation even though it is very obvious. But, while detailing more about how the experts’ predication should be collected, make sure that the focus of the paper should not go off. Being a SW related journal, the readers would be more interested in the comparison of predicated difficulty levels with the actual levels collected from students’ performance than with experts’ predication.

Response Thanks for your comment. Regarding the importance of collecting difficulty prediction from experts, please see our response to point 3.1.17
### 3.1.17.

<table>
<thead>
<tr>
<th>Comment</th>
<th>Please elaborate the reasons for why expert predication is considered as a major component of the evaluation framework. I think this is an importance point which determines the relevance of the paper.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>We believe that expert review is a major component of the framework for two reasons: 1) their judgement of appropriateness is needed for filtering inappropriate questions and 2) while some studies showed that expert prediction is not precise for individual items, their estimation is still useful in determining the overall difficulty of the question set (see “4.5.1. The implications of teachers’ estimations of difficulty” in [18]”). We believe that having an appropriate question set (i.e. error-free and varied in terms of difficulty) is important to keep residents motivated and interested in completing the mock exam. In contrast, having questions that are too easy, too difficult or contain errors could make the experiment boring or inappropriately demanding for participants. Additionally, since experts are still relied on in difficulty prediction (which is mostly the case in low stake exams), we believe that it is essential to use it as a baseline to compare against other methods.</td>
</tr>
</tbody>
</table>
3.2. Reviewer 2 response

3.2.1. Comment
The topic of the paper is quite interesting and relevant for the journal. Automatic creation of exams, especially using structured domain knowledge could improve significantly reliability of exams and their quality across different domains. The paper is well written and organized. The main comments I have with respect to the paper are mainly related to the experimental procedure, which requires significant in depth clarification additions:

Response
Thanks for your encouraging comments. We have responded to other comments regarding experimental design below.

3.2.2. Comment
reference [7] has been continuously referred to in the paper, however such paper is not to be found online. It is inappropriate to cite papers which are in review and not accepted at the time of submission. This is a major drawback as critical explanations for understanding the methodology are supposedly explained there. It is also not clear how the current study and the study in [7] are different.

Response
Apologise for any inconvenience caused. We have clarified the relation between the paper under review and our previous papers in the introduction. Our previous work is now published and can be found at: https://link.springer.com/article/10.1007/s40593-018-00172-w

3.2.3. Comment
when subjects are described in the experimental section it is also important to know how were they recruited and what was the payment. Only the demographics are not sufficient to judge the appropriateness of both experts and residents.

Response
As per your suggestion, we have added information about the recruitment (Sections 4.1.1 and 4.2.1). Regarding payment, our industrial partner “Elseveir” took care of the payment. Experts were paid between $100-$300 while students received a gift card for $50-$100. We already mentioned that participants were paid for their participation but we don’t believe that reporting the exact payment is necessary.

3.2.4. Comment
it is interesting to observe that out of 15 experts 10 had no or less than a year experience in constructing exams. Experience with exam construction is, in my opinion, crucial in such kind of study, as the judgement on appropriateness, difficulty, etc can vary significantly with experience

Response
Out of those 10 reviewers who have no or less than a year experience in constructing exams, 5 reviewers have between 1-3 years teaching experience and 4 have more than 3 years of teaching experience. We believe that this is realistic and mimic the procedure for exam construction.

3.2.5. Comment
"out of 435 questions, 375 questions were rated as appropriate by at least one reviewer" - here it is important to explain what was wrong with the 60 questions, why did they judge them inappropriate and was there a specific scale used, or just a binary answer.

Response
Thanks for your comment. We have added information about the categories used for judging the appropriateness (Section 4.1.3) and the reasons for inappropriateness (Section 4.2.2).
3.2.6.

**Comment**
what happened when the reviewers disagreed? Were these questions studied, why did the experts disagreed? Why did you decide to take a question if one of the two reviewers disagreed, and not just omit the question?

**Response**
Thanks for raising this point. One of the main reason 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. These questions were included in the mock exam experiment so that we can further investigate their difficulty. We have clarified this now in Section 4.2.2

3.2.7.

**Comment**
"Questions were reviewed by two experts whenever possible" - please explain. What does it mean 'whenever possible'. Provide more details how many questions were reviewed by how many reviewers. See the comment above - explain also how did you deal with the disagreement and with questions reviewed by only one expert. How reliable this could be? If 'orthopedics' had only one expert - shouldn’t you just not included this domain in the experiment?

**Response**
Apologies for not making this point clear. We have added the requested information to section 4.1.3. With regards to orthopedics', we were able to recruit only one reviewer for the expert review and we reported this honestly. In addition, we were able to recruit 5 residents for the mock exam. Since the main focus of the paper is the comparison against student performance, the later was of interest more than the former, which was the reason for keeping orthopaedic questions.

3.2.8.

**Comment**
It seems that there is misbalance between the domain expertise of experts and residents. Many of the residents were from expertise not covered by most of the experts. How does this effect the experimental design reliability and the results.

**Response**
Thanks for raising this point. Two issues may arise here: 1) inability to have multiple rating of difficulty which was the case for orthopedics questions and some questions from other specialities 2) Inability to calculate percentage correct due to the small number of students. Regarding the first issue, we agree that this is one of the limitations of our study but we have reported honestly that 119 (out of 435) questions were evaluated by one expert. Questions reviewed by one reviewer were analysed separately in our previous paper. With regards to the second issue, having a low number of participants is a problem facing most similar studies. We accounted for this using a new analysis approach which is useful for other studies facing the same problem (see section 4.2.4).

3.2.9.

**Comment**
Explanation was was provided to experts only if they got the answer wrong. However, experts we asked to judge the explanations to the questions. This might create bias to which explanations experts saw and the overall judgement of reliability

**Response**
Apologies for not making this point clear. If experts selected an incorrect option as an answer, they were shown an explanation of why the selected option is wrong. The intention was to facilitate their decision about question usefulness. At a later stage of the evaluation, they were shown the explanations of all options (independent of their selected option). We have clarified this now (Section 4.1.3)
3.2.10.

Comment  With respect to the research approach, I wonder how transferrable this is to other domains. The authors have only published in medical field on this topic. For this research to have a more general impact it is important to study the transferability to other domains, as well as to study the impact it can have on the difficulty estimation when domains are not well structured, or multiple taxonomies exist. As such Section 3, 6 and 7 need a substantial clarification on these points.

Response  While we agree with you that transferability to other domain is always interesting, we believe that research on medical, case-based questions (CBQs) has a high impact due to the heavy use of these question in the medical education and in board exams. For example, from among 1,750 questions used in the German national medical licensing exam over six years (2006-2012), 51.1% were CBQs [50]. Similarly, out of 1,143 questions presented in the access exam for medical specialities in Spain over five years (2009-2013), the percentage of CBQs ranges between 50% and 58% [51].

Regarding your second point, a major assumption of using ontology-based measures for prediction is the use of good quality ontologies. We have already pointed to these two points in the methodological reflection section.

3.2.11.

Comment  How were the different aspects of the questionnaire rated by the experts, was there a scale, what was the agreement / disagreement on this among experts.

Response  Thanks for your comment. We have added information on the scale for question usefulness. The focus of this study is on comparing the prediction of experts and automated measures with student performance. Agreement/disagreement on the different aspects of the questionnaire was investigated and reported on our previous study which is now published at https://link.springer.com/article/10.1007/s40593-018-00172-w