A survey on recent approaches to question difficulty estimation from text

LUCA BENEDETTO and PAOLO CREMONESI, Politecnico di Milano, Italy

ANDREW CAINES and PAULA BUTTERY, Computer Laboratory & ALTA Inst., University of Cambridge, U.K.

ANDREA CAPPELLI, ANDREA GIUSSANI, and ROBERTO TURRIN, Cloud Academy Sagl., Switzerland

Question Difficulty Estimation from Text (QDET) is the application of Natural Language Processing techniques to the estimation of a value, either numerical or categorical, which represents the difficulty of questions in educational settings. We give an introduction to the field, build a taxonomy based on question characteristics, and present the various approaches that have been proposed in recent years, outlining opportunities for further research. This survey provides an introduction for researchers and practitioners into the domain of question difficulty estimation from text, and acts as a point of reference about recent research in this topic to date.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Applied computing** → **Education**; • **General and reference** → **Surveys and overviews**.

Additional Key Words and Phrases: question difficulty estimation, question calibration, student assessment

ACM Reference Format:

Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2018. A survey on recent approaches to question difficulty estimation from text. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 35 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Question Difficulty Estimation (QDE) – also referred to as "question calibration" – consists of estimating a value, either numerical or categorical, representing the difficulty of a question, and is of crucial importance in the educational domain. The best way to intuitively understand the importance of an accurate estimation of question difficulty is through some use cases. An example is Computerized Adaptive Testing [62], an examination format in which the students are provided with questions whose difficulty is targeted to their proficiency, that was shown to be highly beneficial to the learning outcome [18]. In case of miscalibrated questions (i.e., whose difficulty has been erroneously estimated), students are provided with questions inappropriate to their level, which affects the learning outcome [108]. Another example is the fact that question difficulty is leveraged for accurately assessing students: indeed, in some testing frameworks, students' skill levels are estimated based on their past answers to exam questions and the known difficulty of those questions. A student which correctly answered a very difficult question will have an estimated knowledge level higher than the one of a student who correctly answered a lower difficulty question; therefore, miscalibrated items may affect the accuracy of students' assessment. Lastly, regardless of the testing theory that is used in designing the exams, a test that is too easy or too difficult for a particular group results in a limited range of scores, which is not informative [3].

Traditionally, QDE is performed with either i) manual calibration [1] or ii) pretesting [58]. Manual calibration consists of having one (or more) domain experts manually selecting a numerical or categorical value representing the difficulty of each question, which is not scalable, intrinsically subjective and inconsistent. The other approach, pretesting, consists of deploying the new questions in an exam, as if they were standard questions, but without using them to assess students.

Manuscript submitted to ACM

The other questions in the exam are used to assess the students, and their answers – together with the estimated skill levels – are used to calibrate the questions under pretesting. Even though this approach leads indeed to an accurate and reliable estimation of question difficulty, it introduces a long delay between the time of question generation and when the questions can be used to assess students. Also, it requires the new questions to be shown to students before being actually used to score them, which is undesirable, since they might be leaked or exposed too often [110].

In order to overcome the limitations of traditional approaches to question calibration, recent research has attempted to leverage the textual content of questions with Natural Language Processing (NLP) techniques to automatically estimate their difficulty. Indeed, question text is the only information that is always available at the time of question creation and, being able to estimate question difficulty from it, we would overcome the need for pretesting, manual calibration, and their limitations. No surveys have been carried out about Question Difficulty Estimation from Text (QDET), which is a research direction that has received increased attention in recent years, thanks to concurrent advancements in NLP. Two existing surveys focused on question generation have referred to some works which addressed the task of QDET [15, 57]. However, such survey papers did not consider papers which performed QDET without focusing on question generation; therefore, a comprehensive review of recent literature on QDET is still lacking.

With this survey paper, we have the goal of presenting a comprehensive review of recent (i.e., from 2015) approaches to QDET. Even though the research of techniques to perform QDET and to modify questions difficulty in a controllable manner has a fairly long history [12, 60, 79], we focus only on recent works since there has been rapid development compared to previous years. Indeed, the last few years have seen an improvement in the capabilities of NLP techniques and this has been reflected in the progress on QDET. Overall, we find that there has been a shift from the usage of theoretically supported features such as readability and word-complexity measures towards approaches which rely upon modern NLP techniques based on machine learning.

In this survey, we aim at creating a single point of reference which can be looked at by any researcher and practitioner working on the task of QDET or approaching it for the first time. We propose a taxonomy based on question format to organize all the approaches published so far, and analyze the techniques which have proven effective or not effective in certain scenarios. We do not perform a quantitative comparison of the different approaches as that is not feasible for several reasons. First of all, different approaches are generally designed to work on different scenarios – different educational domains, different types of questions, different question formats, different definitions of difficulty, etc. – and, secondly, due to the value of exam material, there is a lack of publicly available resources, which makes it difficult to exactly reproduce all the proposed approaches. Indeed, most of the papers presented in this survey only compare themselves with simple baselines (random or majority), rather than with previously proposed models.

The contributions of this work can be summarized in the following points: i) we perform a comprehensive review of recent work on QDET, ii) we propose a taxonomy to organize such works, iii) we discuss future research directions and limitations of the proposed approaches. We envision two ways of reading this survey: i) a read from start to finish, which informs on all the approaches that were proposed in previous research for different types of questions and compares them, and ii) a read focused on specific types of questions, which can be guided by the proposed taxonomy.

This document is organized as follows. Section 2 describes the research method. Section 3 introduces the testing theories featuring in this survey. Section 4 presents the proposed taxonomy. Section 5 and Section 6 dive into the details and describe the approaches proposed in the literature. Section 7 presents some additional analyses, along different dimensions, of the papers presented in this survey. Section 8 concludes the paper.

2 RESEARCH METHOD

In order to retrieve the works included in this survey, we proceeded as follows.

- We performed a comprehensive search on digital libraries (AAAI, ACL, ACM, Elsevier, IEEE, Springer) using relevant keywords; in order not to miss relevant papers, we also performed the same search on Google Scholar, since it offers wider coverage of works on the Internet, but we only retained peer-reviewed research publications.
- We manually filtered all the papers, keeping the ones that satisfy the following criteria: i) they propose and evaluate approaches that leverage textual information to perform QDE, either as the final target or as an intermediate step; ii) they focus on the educational domain; iii) they have been peer reviewed; iv) they were published in 2015 or later; v) they are written in English.
- We collected all the papers which cited or were cited by the remaining papers; we made use of the citation network because we observed that the initial keywords – although chosen to include as many relevant works as possible – still failed to retrieve some papers which are relevant to this survey.
- We filtered the resulting papers using the same criteria as before.

Several recent works proposed models to create question embeddings from text, claiming that they are capable of capturing several question characteristics, including difficulty. The final target of these papers is not QDET and no experiments are performed to support the claim that such embeddings are capable of capturing question difficulty, therefore they are not included in this survey (e.g., [49, 50, 65, 75, 95, 103]). Similarly, we consider papers from the recent literature on difficulty-controllable question generation only if they can be used for QDET of already existing question. For instance, we do not analyze [33], which proposed an approach for generating questions of a given difficulty but cannot be used to calibrate existing questions. Lastly, we do not include the papers that perform QDET in domains other than education, such as community question answering systems [64], the ones that have the target of estimating reading complexity of a piece of text (e.g., [20]), and the ones that consider question difficulty for question answering models instead of human learners [34].

3 THEORETICAL BACKGROUND: THEORIES OF TESTING

All the papers presented in this survey perform QDET, but the definition of difficulty can be diverse. Indeed, three approaches are used: i) Classical Test Theory, ii) Item Response Theory, and iii) manual definitions. Regardless of the theory (if any) used to obtain it, question difficulty can be either a continuous value or a discrete value, therefore the task of QDET can be seen either as a regression task or as a classification (discrete regression) task. It is important to remark here that the decision of which theory to use is an exam design choice outside the scope of this survey. However, below we summarise the theoretical testing frameworks which feature in this survey.

3.1 Classical Test Theory (CTT)

CTT [38] is a well established testing theory that predicts outcomes of psychological testing, such as the difficulty of items or the ability of test-takers. The term "classical" refers to the contrast with modern psychometric theories such as IRT, compared to which CTT has the advantage of being simple to compute and to understand.

CTT assumes that each individual is associated with a true score T, which would be the expected correctness of an infinitely long run of repeated independent administrations of the same test. In practice, the observed score X is used, which is the sum of the true score T and an error E: $X = T + E$, where T and E are two unobservable (or latent) variables. The major assumptions of CTT are that i) T and E are not correlated, ii) E is normally distributed with zero mean, and

iii) the errors of different tests are not correlated. The concept of item difficulty of an item in CTT is expressed by the *p-value*, which is a continuous value in the range $[0; 1]$. The p refers to "probability" and is the fraction of correct responses in the considered population. The p-value is typically referred to as *correctness*: the higher the p-value, the easier the item. Similarly, we can define the*wrongness* as $1 -$ p-value: the higher the value, the more difficult the item.

The main limitation of CTT is that it does not leverage the students' skills when estimating the item difficulty: in practice, it simply uses the fraction of students that wrongly answer a question without considering their skill level.

3.2 Item Response Theory (IRT)

IRT [39] is another well established technique that associates latent traits to both students and questions. Its simplest implementation, the one-parameter model (named the "Rasch Model" [82]), associates a skill level θ to each student and a difficulty level b to each question. An important property of IRT is "invariance": item latent traits do not depend on the ability distribution of test takers and a given question is assigned the same difficulty regardless of the skill levels of the students answering it (in contrast to CTT, which simply considers the fraction of correct and wrong answers). Two important assumptions of IRT are that i) individuals are independent from each other and that ii) the item responses of a given individual are independent from each other.

For a given question j and its latent trait b_j, we can define the item response function (i.r.f.) which indicates the probability ($\mathrm{P_C}$) that a student i with skill level θ_i answers the question correctly. The formula of the i.r.f. is as follows:

$$\mathrm{P_C} = \frac{1}{1 + e^{-1.7\cdot(\theta_i - b_j)}} \tag{1}$$

where the coefficient 1.7 was empirically found in previous research to generally lead to accurate results. The background intuition is that a student with a given skill θ_i has a lower probability of correctly answering more difficult questions: if a question is too difficult or too easy (i.e., $b_j \longrightarrow \infty$ or $b_j \longrightarrow -\infty$), all the students answer in the same way (i.e., $\mathrm{P_C} \longrightarrow 0$ or $\mathrm{P_C} \longrightarrow 1$), which shows why it is important to have assessment items that are not too easy nor too difficult.

Question difficulties obtained in IRT are real values in a given range (selected at the time of calibration) but, in practice, are sometimes converted to discrete values, thus representing difficulty in a discrete manner.

3.3 *Manual definition*

In some cases, question difficulty is not based upon any learning theories and it is just manually selected by educational experts. In all these cases – at least considering the papers presented in this survey – difficulty is a discrete value, and the number of possible classes can vary, depending on the specific implementation.

4 TAXONOMY

Figure 1 presents the taxonomy we propose for categorizing all the papers presented in this work. We group the papers depending on the characteristics of the questions that the proposed models work on, since the type of question heavily affects the models that can be used in each application scenario. We provide here a brief overview of the proposed approaches and their categorization, and will describe them in detail in Section 5 and Section 6.

The first distinction is the educational domain considered by each work; there is a crucial difference between i) Language Assessment (LA), both first and second language, and ii) Content Knowledge Assessment (CKA), e.g., math. Indeed, question difficulty comes from different sources in these two scenarios: in LA, the difficulty comes from the linguistic demands of the task and the topic being assessed along with any stimulus text, while in CKA the difficulty mostly comes from the topics which are being assessed and the question format has a smaller importance. Moreover,

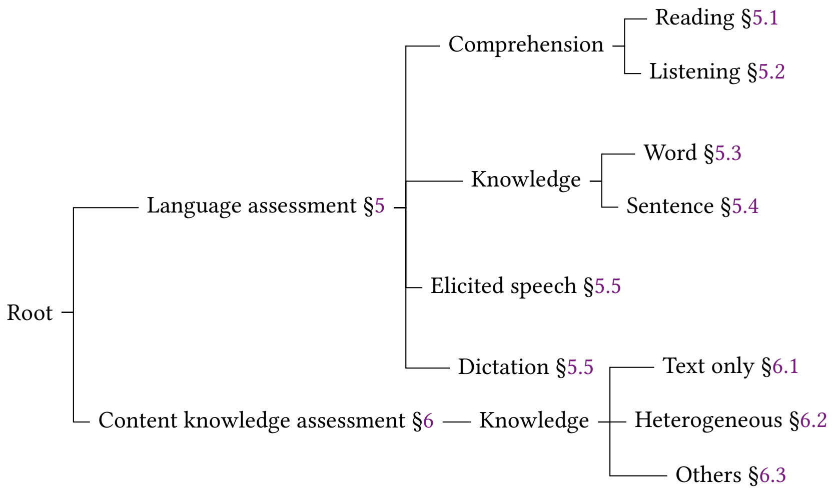


Fig. 1. The taxonomy based on question format we use for categorizing the papers presented in this survey.

CKA questions are often built in order to minimize the effects of language on the difficulty [114]. This difference has an influence on the approaches to QDET in the two domains, as they focus on different features: approaches developed for LA often rely upon theoretically supported measures such as readability formulas and predefined word complexity measures, which are rarely used in CKA; on the other hand, CKA works often leverage learnt features, such as TF-IDF (i.e., Term Frequency–Inverse Document Frequency [52]) and word embeddings (e.g., word2vec [69], ELMo [80]), or end-to-end neural networks, which are much less common in LA. This difference also shows that – generally – research in LA does not focus on semantic word representations, which is instead very important for CKA.

Almost all the proposed approaches to perform QDET, in both domains, address the task as a supervised problem: a training set containing texts and difficulties of exam questions is used to train a model which is capable of performing QDET for previously unseen questions. In some cases, additional textual datasets are used, to pre-train the model or part thereof. In such cases, the models built for LA leverage general purpose datasets (e.g., Wikipedia), while the ones built for CKA leverage datasets related to the topics that are assessed by the questions (e.g., books, lecture transcripts).

4.1 Language Assessment (LA)

Focusing on LA, most approaches to QDET deal either with i) *comprehension* questions or ii) *knowledge* questions. Comprehension questions are provided to the student together with a passage (either written or spoken) which contains the answer to the question, meaning that the task of a comprehension question involves finding in the passage (or inferring from it) the answer to a question. On the contrary, knowledge questions assess the knowledge of the student at a certain time and the answer to the question is not found in (or inferred from) a related passage[1].

4.1.1 Comprehension questions. These questions can be categorized into reading comprehension and listening comprehension. Only one of the works presented here focuses on listening comprehension questions [67], while reading comprehension questions received slightly more consideration in previous research, as we found four relevant works

[1] For clarity, we remark here that this definition of "comprehension questions" and "knowledge questions" is different from that given in Bloom's taxonomy [11]. Indeed, Bloom's taxonomy delineates a hierarchy of cognitive-learning levels, ranging from the knowledge of specific facts to more advanced levels of synthesis, while here we categorize the questions depending on their format.

which focused on it [10, 47, 48, 61]. We also notice that the papers on comprehension questions are very recent (the first one is dated 2017), and this is most likely due to the fact that recent advancements in NLP techniques based on neural networks have enabled new ways of leveraging the accompanying texts.

4.1.2 Knowledge questions. This type of questions has received more attention than comprehension questions, and this interest also started before the first research on comprehension questions. This also has an impact on the types of models that are used for QDET of knowledge questions for LA. Indeed, many of the models use fairly simple and theoretically-grounded features such as word-complexity for learners of specific languages and readability measures. No end-to-end neural networks were proposed so far and most of the works did not experiment with word embeddings or word frequency features. Knowledge questions for LA can be further divided depending on their format: some are vocabulary questions made of single words [23, 26, 88, 116], others represent whole sentences [5, 6, 31, 44, 47, 59, 74, 88, 97–99, 104].

4.1.3 Others. There are two types of questions which are explored in one paper only [88] and do not really fall in any of the previous categories: i) elicited speech and ii) dictation exercises. The elicited speech task evaluates reading and speaking skills of students by requiring them to produce a sentence out loud, while the dictation task consists of asking the students to transcribe an audio recording and thus evaluates both listening and writing skills[2].

4.2 Content Knowledge Assessment (CKA)

In CKA, all items are knowledge questions, and can be categorized depending on the content of the questions. Specifically, they can be divided into i) text only questions, and ii) heterogeneous questions, which contain information – such as images – that cannot be captured at text level. Equations and formulas are generally considered as "text", since they can be expressed in LaTeX-like verbal format [118]. Questions with images are quite rare and this is reflected by the fact that only three works [30, 94, 119] experimented on QDET for heterogeneous questions. The research focused on text only questions can be categorized depending on the type of information that is leveraged by the models. Specifically, we can distinguish between i) models that only consider the question text for the task of QDET [8, 9, 27], ii) models that also leverage texts from other sources (e.g., lecture content, books, etc.) [45, 81, 114, 115, 122], and iii) models that leverage non-textual information (e.g., ontologies [29, 55, 89, 107], knowledge components [21, 102], and others [94, 113]).

Lastly, there are two works which do not belong to any of the previous categories because they deal with specific types of questions and can be used only in the niches they were designed for. One of them [78] deals with questions whose answers are in the form of First Order Logic formulas and leverages such formulas for QDET. The other [72] performs QDET for short-answer questions and leverages the text of the students' answers (not of the question).

5 LANGUAGE ASSESSMENT

In this section, we present and discuss the approaches that have been proposed for QDET in the language assessment domain: we focus on reading comprehension questions in subsection 5.1, on listening comprehension questions in subsection 5.2, on word knowledge questions in subsection 5.3, on sentence knowledge questions in subsection 5.4, and on elicited speech and dictation items in subsection 5.5.

[2] Considering the additional information that is available (i.e., text to read and audio to listen to) these types of questions might seem to belong to the category of comprehension question. However, they do not require the students to infer the answer to a specific question from the text/audio, but only to perform a transformation from written to spoken form or vice-versa.

Text:
Halloween [...] is one of the most unusual and fun holidays in the United States. [...] This tradition comes from an old Irish story about a man named Jack who was very stingy. [...] But he also could not enter hell, because he had once played a trick on the devil. [...]

Question: What is the reason why Jack could not enter hell?

a. Because he had once played a trick on the devil. b. Because he walked on the earth carrying a lantern.
c. Because he had once played a trick on a witch. d. Because he was very stingy.

Fig. 2. Example of reading comprehension question from [47].

5.1 Reading comprehension questions

In reading comprehension questions, students are given a textual passage and one (or more) questions associated with it, as shown in the example in Figure 2. The reading passage is an important component – although not the only one – of question difficulty, and this is reflected in the four models proposed in recent years. Indeed, one of them [47] completely bases the estimation of question difficulty on the reading complexity of the reading material, while the others [10, 48, 61] leverage both the text of the question and the text of the accompanying passage. An overview of the four models is shown in Table 1.

Paper	Year	Sources of text	Approach
[47]	2018	Reading passage only.	Reading difficulty directly used as an indication of question difficulty.
[10]	2021	Reading passage, question text.	Five features computed from the text of the question and the passage are normalized, averaged, and then compared to a threshold.
[61]	2019	Reading passage, question text.	Words are embedded with word2vec, the sequences of word embeddings are embedded with LSTMs, the final estimation is done with an FCNN.
[48]	2017	Reading passage, question text, and distractors.	Words are embedded with word2vec, the sequences of word embeddings are embedded with a sentence CNN, an attention mechanism is used to detect the relevant parts of the passage, and the final estimation is done with a FCNN.

Table 1. Overview of the approaches proposed for estimating the difficulty of reading comprehension questions.

Being more specific, in [47] (2018) the authors assume that examinees correctly answer a reading comprehension question only if they can understand the whole textual passage, therefore they directly use reading complexity as an indicator of question difficulty. For the estimation of reading complexity, the authors adopt a measure designed for learners of English as a foreign language [46]. This can be considered a fairly simple approach, and it leaves plenty of room for improvement, since it estimates the same difficulty for all questions associated with a certain passage. By observing the correctness of students' answers across a set of questions of different difficulty, the authors observe that there is a relation between question difficulty and average correctness.

In [10] (2021), the authors propose a complexity-controllable question generation model, which has a complexity estimator that can be used on already existing questions as well. The proposed approach is fairly simple, as it computes five features and compares their values with a trained threshold. To be precise, it computes i) number of clauses in the question, ii) number of dependency relations in the question, iii) topic coherence of sentences in the passage, iv) frequency of question entities in the passage, and v) distance between entities in the question and in the passage. Then, it computes their average (after normalization), and compares the result with the threshold: if it is larger than the threshold, the question is labelled as difficult, otherwise it is labelled as easy.

In [61] (2019), the authors propose a neural model to estimate the difficulty of Chinese reading comprehension items. First of all, each word is transformed into a semantic vector of 300 dimensions with word2vec (trained on the Sinica Balanced Corpus [19]), and there is no distinction between the words of the document and the words of the question. The embedding vectors are input into two uni-directional Long Short-Term Memory networks (LSTMs) [42]. Then, the output derived from the two LSTMs is input into the a Fully Connected Neural Network (FCNN) made of three layers that outputs a value in the range $[0; 1]$ representing the difficulty of the item. The experimental dataset is made of 334 items and the model reaches an accuracy of 37%, with a random baseline of 20%.

The model proposed in [48] (2017) is the only one that explicitly takes into consideration the relation between the reading passage and the question. It does so by using an attention mechanism [106] to model the importance of each sentence in the reading document for a specific question. The proposed model is made of four components: i) input component, ii) sentence CNN (Convolutional Neural Network) component, iii) attention component, and iv) prediction component[3]. All the questions are Multiple Choice Questions (MCQ), and the model leverages both the text of the question (i.e., the stem) and the text of the options.

In the input component, all the text material of a question (i.e., document, stem, and options) is converted into pretrained embeddings using word2vec (with 200 dimensions) trained on the English Gigaword dataset [36]. The sentence CNN component reduces the dimensionality of the input data by applying a series of convolution and max-pooling operations. The attention component aims at finding which parts of the text are relevant for each question. In practice, there are two attentions involved in the model, both computed using cosine similarity: the first one measures the similarity between the text stimulus and the question, the second one measures the similarity between the question and the available answers. Lastly, the prediction component concatenates the two outputs of the attention components and uses a FCNN to learn the difficulty.

The proposed model outperforms the baselines proposed by the authors, which are simplified versions of this same model, but a random or majority baseline is not evaluated. Even though this model is arguably more advanced than the other two, it is the oldest one from a temporal point of view, therefore the authors do not compare with the other models explicitly built for QDET in reading comprehension questions. The attention mechanism is a particularly interesting feature of this model, since it is an important step towards explainability: the authors do perform some analysis of the text spans attended to, but without diving into details, which would certainly be worth doing in future research.

5.2 Listening comprehension questions

Only one paper [67] about QDET for listening comprehension questions was published in recent years. Specifically, the authors focus on MCQs from an English language proficiency test. Item difficulty depends on both the audio transcript and the text of the question, and indeed the proposed approach leverages both sources of information. First, the authors compute 339 features from the text (written and spoken) using *TextEvaluator*, an automated text complexity prediction system [92]. Then they experiment with several regressors for estimating item difficulty from the features. The features can be categorized into the following groups: academic vocabulary, concreteness, word familiarity, syntactic complexity, cohesion, argumentation, conversational style, and narrative structure.

Using the Pearson's correlation coefficient between the true difficulty and the estimated difficulty, the authors show that a random forest regressor consistently outperforms all the other models[4]. All the groups of features seem to bring

[3] The authors refer to these as "layers" instead of "components"; we change notation to clarify that these components can themselves be composed of several hidden layers.

[4] They experimented with least squares linear regression, LASSO regression, decision tree regression, elastic net, k-neighbours regression, stochastic gradient descent regression, linear and non-linear support vector regression.

Select the real English words:
A. Frequently B. Positively C. Apply
D. Morride E. Shampoo F. Brican

(a) Example of Yes/No question.

Word: apply

Options:
1. I have not seen this word before
2. I have seen it, but I don't know what it means
3. I have seen it before and I think it means:
   ____________ (synonym or translation)
4. I know this word. It means
   ____________ (synonym or translation)
5. I can use this word in a sentence: ____________

(c) Example of Vocabulary Knowledge Scale question.

Word: Microphone

Definitions:
A. Machine for making food hot
B. Machine that makes sounds louder
C. Machine that makes things look bigger
D. Small telephone that can be carried around

(b) Example of Vocabulary Level Test with four definitions and one word.

Definitions:
A. Set of beliefs
B. Having a very close relationship
C. Separate parts of something larger

Words:
1. Intimacy 2. Doctrine 3. Section
4. Focus 5. Volume 6. Mathematics

(d) Example of Vocabulary Level Test with three definitions and six words.

Fig. 3. Examples of vocabulary questions.

valuable information for QDE, and the most highly ranked features were related to the lexical content of the item text, for all item types. These highly ranked features covered three aspects of vocabulary: i) vocabulary diversity measured as type-to-token ratio in the item texts[5], ii) the difficulty of vocabulary in the item text as measured by the frequency of the words in different corpora, and iii) the concreteness and imageability of the text [91].

5.3 Single word knowledge questions

The single word knowledge questions are all vocabulary questions and they have the target of assessing the vocabulary breadth of students. They can have different formats, as shown in Figure 3. In Yes/No tests (Figure 3a) students receive a list of words, and have to select the ones that are real words. In Vocabulary Knowledge Scale (VKS, Figure 3c), students are asked to report how well they know a word and – if they report knowing it – they have to provide a synonym, a translation, or an example of the word in context [51]. Lastly, in Vocabulary Level Test (VLT) students are shown one or more definitions together with one or more target words, and they have to match the definitions with the target words [4, 71, 85]. The number of definitions and target words can vary, as shown in Figure 3b and 3d.

Differently from comprehension questions, no information is available in addition to the target word and (possibly) some definitions, therefore the proposed models are generally simpler from an architectural point of view, as visible in the overview in Table 2. Interestingly, no one of the proposed approaches leveraged the definitions for QDET.

Neural networks are rarely used and, when they are, there is generally little focus on semantics, as the difficulty is assumed to come mainly from other aspects of the words. Indeed, [26] (2018) is the only work that leverages word2vec embeddings without any other features for QDET for this type of question. Specifically, the author focuses on VLT, with one word and four definitions. The approach is made of a word2vec model – pretrained on all English Wikipedia texts – for embedding the target questions, and a Support Vector Machine (SVM) regression model with linear kernel for the numerical estimation. The author experiments with questions whose difficulty (obtained with IRT) is in the

[5] Type-token ratio (TTR) is a measure of vocabulary variation within a text, and it is shown to be a helpful measure of lexical variety.

Paper	Year	Approach	Question format
[26]	2018	SVM that uses as features word2vec embeddings.	VLT
[116]	2018	SVM that uses as features: word length, word frequency, utilization on the web, Age-of-acquisition, concreteness rating, number of POS tags, most frequent POS tag, word2vec embeddings, number of double consonants, number of vowels, presence of shorter homophones.	Yes/No, VKS, VLT
[88]	2020	Weighted softmax model that uses as features: word length, log-likelihood from character-level language model, Fisher score.	Yes/No

Table 2. Overview of the approaches proposed for estimating the difficulty of single word knowledge questions.

range [−5; +5], and the proposed model is only capable of performing QDET with a RMSE of 3.632. No baselines are evaluated, but the results show that there is clear room for improvement. There might be several reasons for this: one of them is most likely the small dataset that was used for the experiments (92 words, 22 being held-out for testing), but maybe word2vec is simply not the most suited model for this task.

The first work that evaluated the correlation between the difficulty of vocabulary questions (Yes/No, VLT, and VKS) and some textual features is [23] (2015), which found that character length and corpus frequency significantly correlate with vocabulary difficulty. However, this work did not have the task of performing difficulty prediction, and it is therefore mostly used as a starting point by more recent research.

An example is [116] (2018), in which the authors propose an approach that can be used for VKS, VLT with one word, and Yes/No items (although only for real words). The approach consists of i) computation of features related to the word difficulty level, ii) reduction of these features with Principal Components Analysis (PCA), and iii) classification with a SVM. The model uses the following features: word length, word frequency (obtained from NLTK corpora [66]), utilization on the web (i.e., number of relevant documents retrieved by Google), Age-of-acquisition from [56], concreteness rating from [14], number of part-of-speech (POS) tags (obtained from NLTK corpora), most frequent POS tag, word2vec embeddings (100 dimensions), number of double consonants in the word, number of vowels, and existence of shorter homophones. The second step of the proposed approach consists of reducing the dimensionality of the data using PCA [76]: specifically, the authors reduce the dimensionality of the input data from 111 features to 2 features; no experiments were performed with other dimensions. Using manually defined difficulties as gold standard, the authors report an accuracy of 73.5%, with random baseline of 33.3% (three difficulty levels).

The most recent paper in this section [88] (2020) focuses exclusively on Yes/No items. The proposed model uses three groups of features: i) character length of the target word, ii) corpus frequency, and iii) "Fischer score". While character length is straightforward to calculate, corpus frequencies may only be obtained for real words, whereas the pseudo-words found in Yes/No items inherently do not occur in corpora and therefore have no frequency value. Therefore, the authors propose a character level Markov chain language model to compute the log-likelihood of a word (or pseudo-word), and use this as feature instead of the corpus frequency; this language model is trained on the OpenSubtitles corpus [63] Lastly, the Fischer score of a word is a vector representing the gradient of its log-likelihood under the language model (conceptually similar to trigrams weighted by TF-IDF [28]). The authors experiment both with a linear regression model and a weighted-softmax, and observe that the first appears to overfit the training data. The weighted-softmax does not overfit and leads to estimated difficulties which are in agreement with expert judgements. They also find that the Fischer score features are the most useful for QDET, while character length has little impact.

His characteristic talk, with its keen _____ of detail and subtle power of inference held me amused and enthralled.

1. instincts 2. presumption 3. observance
4. expiation 5. implements

(a) Example of cloze item.

Vacc___ like penic___ and ot___ antibiotics th___ were disco___ as a dir___ result are lik___ the grea___ inventions o___ medical sci___.

(b) Example of c-test.

___ines like ___illin and ___er antibiotics ___at were ___vered as a ___ect result are ___ely the ___test inventions ___f medical ___nce.

(c) Example of prefix deletion item.

The Taj Mahal __________________ (build) around 1640.

(d) Example of Cued Gap-Filling Item.

Text: "[...] The exact role of other factors is much more difficult to pinpoint; for instance, [...]"
Question: The word "pinpoint" in the paragraph is closest in meaning to:
1. identify precisely 2. make an argument for 3. describe 4. understand

(e) Example of closest-in-meaning item.

Fig. 4. Examples of sentence knowledge questions.

5.4 Sentence knowledge questions

Knowledge questions that are presented to students in the form of one or more sentences can be divided into i) reduced redundancy testing, ii) grammar questions, and iii) vocabulary questions.

Reduced redundancy testing [93] is based on the idea that natural language can be redundant thanks to contextual cues, and more advanced learners can be distinguished from beginners by their ability to deal with reduced redundancy. In language testing, a standard approach to reduce redundancy are cloze tests (Figure 4a), consisting in removing some words from a test and asking the learner to fill the gaps. Similar approaches are c-tests (Figure 4b), which have more gaps but provide the first half of the words as a hint [25], and prefix deletion tests (Figure 4c), where the first half of the word is masked and the second part is used as a hint. In this survey, we present seven papers that dealt with the task of QDET for reduced redundancy testing [6, 31, 44, 47, 59, 88, 104].

Grammar questions aim at assessing the grammar knowledge of students rather than their vocabulary breadth. Two papers dealt with QDET for grammar questions in recent years [47, 74], focusing on Cued Gap-Filling Items (CGFI), where learners read a short text and fill in the gap(s) using cues consisting of a single word which must be transformed to fit the context, as shown in Figure 4d.

Even though vocabulary questions are generally single words, as discussed in Section 5.3, in some cases they are made of whole sentences. That is the case for the Closest In Meaning (CIM) questions considered in [97–99]. As shown in Figure 4e, students are given a text passage and are asked to pick, from a set of possible choices, the word that is closest in meaning to a word highlighted in the text.

5.4.1 Reduced redundancy testing. QDET for reduced redundancy testing has received a fair amount of research attention, and the proposed approaches have different levels of complexity; an overview is shown in Table 3.

The approach that is arguably the least complex is not based on any machine learning technique, and was proposed by Huang et al. in 2018 [47], in the same paper that deals with grammar questions and reading comprehension questions. The authors claim that the difficulty of cloze items is determined by the difficulty of the correct answer. To estimate word difficulty, the authors use a graded word list made by an educational organization, the College Entrance Examination Center of Taiwan, which contains 6480 words in English divided into six levels of complexity. For QDET, the authors simply use the word difficulty from the aforementioned list, and observe that higher difficulty generally corresponds to lower correctness of students' answers, without evaluating any baseline.

Paper	Year	Uses sentence(s)	Uses gap word(s)	Approach	Question format
[47]	2018	-	✓	Considers word difficulty as question difficulty, and obtains it from a manually curated table containing the difficulty of 6480 words.	cloze
[44]	2019	✓	-	Linear regression model that uses as features mean token length and mean sentence length.	cloze
[104]	2017	✓	✓	Linear regression model that uses as features 25 linguistic variables at passage and item level.	cloze
[31]	2019	✓	-	Shannon's entropy is used to assign a score to each gap based on the number of candidate words that could fill the gap given the context; the score is used as a direct indication of question difficulty.	cloze
[6]	2015	✓	✓	SVM that uses 70 features from i) the difficulty of the passage, ii) the difficulty of the target word, and iii) test parameters.	prefix deletion, cloze, c-tests
[59]	2019	✓	✓	SVM that uses 59 features reduced from the 70 in [6].	c-tests
[88]	2020	✓	✓	Linear regression model that uses as features: average word length, sentence length, log-likelihood from a language model, and Fischer score.	cloze

Table 3. Overview of the approaches proposed for QDET in reduced redundancy testing.

Another approach was proposed in 2019 [44], in a paper that does not have QDET as the final goal, but still uses textual content for question calibration, specifically dealing with cloze items. The proposed approach is very simple, in that it does not use any information about the gap but only the reading complexity of the passage; in a sense, it might be considered as the opposite approach with respect to [47]. Specifically, it uses the mean token length and the mean sentence length of the textual passage to estimate question difficulty with a linear regression model. The ground truth difficulty is manually defined by human experts and there are two possible levels. Preliminary results presented in the paper show that, even though the chosen approach is arguably simple and cannot distinguish between different questions coming from the same textual passage, there is a positive correlation between the difficulty estimated with the proposed approach and the results observed in a test context; in this case as well, the authors do not perform a comparison with previous approaches.

Another paper addressing QDET of cloze items using only information from the text passage is [31], which performs a pilot study of an entropy based approach to estimate the difficulty. Specifically, the authors build on the assumption that the complexity of a gap is related to the number of possible answers determined by the surrounding context and the likelihood of each answer. In practice, they use Shannon's entropy [90] to assign a score to each gap based on the number of valid words that could fill in the slot given the surrounding context. As a result, gaps with many possible answers will yield higher entropy than those with fewer answers. The authors compute entropy using a 5-gram language model trained on the 1 Billion Word WMT 2011 News Crawl corpus[6] using KenLM [41], and considering only the 100 most probable words when computing the entropy of each gap (complete vocabulary has more than 82200 words). Using CEFR levels of the exams as difficulty gold standard, the authors study the correlation between the difficulty level and the entropy, and observe that indeed higher difficulty levels seem to correspond to greater entropy.

[6] https://www.statmt.org/lm-benchmark/

Trace et al. in [104] study which features affect the difficulty of cloze items, and perform a regression analysis to observe the correlation between item difficulty and such features. Specifically, the authors consider 25 linguistic variables at both passage level and item level (mostly related to the number of words, sentences and syllables, and to the word frequency). They find that both passage level and item level are helpful for QDET and observe that three features accounted for 24% of the total variance of item difficulty: i) the frequency of the item elsewhere in the items, ii) the number of syllables per word, and iii) the number of sentences per 100 words in the passage. Also, using PCA to reduce the original 25 features to 6 features, they observe that the reduced features do not lead to an improved accuracy and a majority of item difficulty remains unexplained.

The first paper addressing not only cloze tests but also c-tests and prefix deletion tests is [6], which extended previous work [5] and proposed a technique for QDET that is applicable to all three test types. Specifically, the proposed approach performs QDET with a SVM regression model using 70 features related to i) the reading difficulty of the text passage, ii) the difficulty of the target word (obtained from a predefined table), and iii) test parameters. The experiments showed a positive correlation between the selected features and the ground truth difficulty, but no baselines are evaluated.

Taking inspiration from [5, 6], in [59] the authors proposed a technique to modify the difficulty of c-tests by varying the number and position of the gaps. As for QDET, they evaluate a model similar to the one proposed in [6], extracting, from the original set of 70 features, 59 features related to: i) item dependency, ii) candidate ambiguity, iii) word difficulty, iv) and text difficulty. The regression is still performed with SVM. They evaluate the model both on the same data as [5, 6] and on a new private dataset, and obtain results in agreement with previous research. Additionally, the authors experiment with neural models for the regression component, but observe that they are outperformed by the SVM on both datasets, although the difference is not great. Specifically, the SVM reaches an RMSE of 0.24 and 0.21 on the two datasets, while a Fully Connected Neural Network (FCNN) reaches 0.25 and 0.22 and a Bidirectional LSTM reaches 0.24 and 0.24. The ground truth difficulties are real numbers in the range $[0; 1]$.

Lastly, [88] proposes a linear regression model for QDET of cloze tests. The proposed model is exactly the same that is used for elicited speech and dictation items (presented in Section 5.5), and very similar to the one presented in the same paper for single word vocabulary questions. Indeed, it uses as features i) the average word length, ii) the sentence length, iii) log-likelihood obtained from a word-level unigram language model, and iv) Fischer score features. The authors evaluate the model using AUC and the CEFR level of English cloze tests as gold standard, and observe that all features are helpful for difficulty estimation. Additionally, with an ablation study, they find that the Fischer score has the biggest impact on the estimation (as in the case of single word vocabulary questions).

5.4.2 *Grammar questions.*

An overview of the two approaches recently proposed is presented in Table 4.

Paper	Year	Uses sentence(s)	Uses gap word(s)	Approach	Question format
[47]	2018	-	✓	Uses a table containing 44 pre-evaluated grammar patterns of known difficulty; the difficulty of the question is the difficulty of the corresponding pattern.	CGFI
[74]	2019	✓	✓	Ridge regression, using 36 features from gap and context.	CGFI

Table 4. Overview of the approaches proposed for estimating the difficulty of grammar questions.

In [47], the authors assume that the difficulty of a grammar question is determined by the difficulty of the grammar pattern of the correct answer. They identify 44 grammar patterns and estimate the difficulty of each one of them observing their rate of occurrence in English textbooks of different grade levels (assuming that the difficulty of the grammar pattern depends on the grade level of the textbook in which it frequently appears). Then, difficulty is estimated by parsing the question to identify its grammar pattern and searching the table for the corresponding difficulty.

The other paper that focused on grammar questions [74] adopted a traditional machine learning approach. Specifically, the proposed approach consists in computing 99 features at i) gap-level (54), ii) item-level (18), and iii) context level (27), and using a ridge regression algorithm for difficulty estimation. Some features were directly extracted by the authors, while others were obtained from publicly available tools. The authors experiment with several configurations (i.e., different subsets of the 99 features) and observe that the best results are not obtained using all of them. Indeed, the best configuration – RMSE of 0.75 with a ZeroR baseline of 1.78 – leverages only 56 features. These features are selected via recursive feature elimination, which consists in recursively eliminating the least influential features. The full model (99 features) reaches a RMSE of 0.78. However, as final model, the authors propose an even smaller model, again obtained with recursive feature elimination, which is capable of 0.77 RMSE using only 36 features (26 gap-level features, 4 context features, and 6 item level features). Some of the most important features for the estimation are: i) the tense of the verb (e.g., simple present, simple past, etc.); ii) the presence of forms such as "used to" or "was going to" and adverbs; iii) the word order; iv) the word frequency (from [43] and [13]), v) the word length, vi) the age of acquisition [56], and vii) the concreteness [14] of the words appearing in the question.

5.4.3 Vocabulary. Three papers have dealt with CIM questions [97–99], an overview is presented in Table 5.

Paper	Year	Approach	Question format
[97]	2017	10 features from target word, reading passage, correct answer, and distractors.	CIM
[99]	2019	Features are reading passage difficulty, similarity between correct answer and distractors, and distractor word difficulty level. Two levels (low/high) for each of them, the number of "low" features represents the difficulty (from 0 to 3).	CIM
[98]	2020	Features are target word difficulty, similarity between correct answer and distractors, and distractor word difficulty level. Two levels (low/high) for each of them, the number of "low" features represents the difficulty (from 0 to 3).	CIM

Table 5. Overview of the approaches proposed for estimating the difficulty of closest-in-meaning questions. All proposed approaches leverage the text of the passage, the correct choice, and the distractors.

The first paper [97] is a study that investigates the relations between several factors of question items in English CIM tests and the corresponding item difficulty. Specifically, the authors consider 10 features obtained from four elements: i) the target word, ii) the reading passage, iii) the correct answer, and iv) the distractors. Most of these features (9 out of 10) are related to the word difficulty of the different elements (e.g., average word difficulty of the words in the reading passage), and the other feature is the number of word senses of the target word. The "word difficulty" is obtained from JACET 8000 [105], which is a list of 8000 words grouped in difficulty levels, specifically built for Japanese learners of English. The experimental results show that the number of word senses does not correlate well with the difficulty, probably because generally each word has one meaning that is much more frequent than the others. Considering the other features, the ones that correlate more with question difficulty are i) the difficulty of the target word, ii) the average word difficulty of the correct answer, and iii) the average word difficulty of the distractors.

In [99], the authors explore how three factors – related to the features mentioned above – can be leveraged to control the difficulty of CIM questions. The three factors are i) reading passage difficulty, ii) similarity between the correct answer and the distractors, and iii) distractor word difficulty level. For each of these factors, the authors only consider two levels (high and low), and the combination of levels is finally used for the task of QDET. For reading difficulty, the authors apply three well-established readability formulas to documents from two sources: *Times in Plain English* to represent lower complexity English, and the *New York Times* representing higher complexity English. The readability formulas used in this study are: Flesch-Kincaid Grade Level, Flesch-Kincaid Reading Ease [54], and Dale-Chall readability formula [16]. The average values obtained for the two levels of English are then considered as reference while performing QDET. For the similarity between correct answer and distractors, the authors use cosine similarity on the vectors representing the words; these vectors correspond to the frequency of the co-occurrence words within a certain window in the corpus. Finally, for the distractor word difficulty level, the authors use JACET 8000. As for the estimation of question difficulty, the authors use the aforementioned factors to obtain four level of difficulty (corresponding to the number of "high" factors in the question), from "LLL" to "HHH". To evaluate the approach for QDET, the authors observe the correctness of students' answers for questions of different difficulty levels, and note that indeed there is a positive correlation between the difficulty estimated by the model and the fraction of wrong answers.

The latest work [98] (2020) is an extension of [99]. Indeed, the authors have the same target of controlling item difficulty, but use slightly different features. The factors taken into consideration are: i) target word difficulty, ii) similarity between correct answer and distractors and iii) distractor word difficulty level. As before, for each factor two levels (high and low) are considered and the question difficulty is obtained from the combination of such levels (i.e., four levels). Again, for the target word difficulty and the distractor word difficulty level, the authors use JACET 8000, while the approach for computing the similarity is different from before and arguably more advanced. Indeed, the authors use GloVe [77] embeddings for calculating cosine similarity. The experimental results support this choice and show that this approach is capable of a more accurate QDET with respect to the previous one.

5.5 Elicited speech and dictation items

Elicited speech and dictation are two tasks that are very rare in the literature on QDET and, indeed, there is only one recent work of relevance here [88]. The elicited speech task taps into reading and speaking skills by requiring examinees to produce a sentence out loud, while the dictation task requires the examinees to transcribe an audio recording, thus it measures both listening and writing skills. The proposed approach is no different for these two tasks and it is the same that is used for c-tests as well (as presented in section 5.4). The difficulty ranking is performed with a linear model, and it uses features from different sources: i) average word length, ii) sentence length, and iii) log-likelihood and iv) Fischer score features from a word-level unigram language model. Therefore, the proposed approach does not really dig deep into the characteristics that are peculiar to this type of question, such as the audio in the dictation task or the number of vowels in each word in the elicited speech task, and these certainly are promising areas of focus for future research.

6 CONTENT KNOWLEDGE ASSESSMENT

All the approaches proposed in the domain of content knowledge assessment focus on knowledge questions, and can be categorised into i) *text only* questions, whose content is only text, and ii) *heterogeneous* questions, which contain information of other types such as images and tables. In Section 6.1 we present the approaches proposed for text only questions, which are by far the most numerous, and in Section 6.2 the approaches proposed for heterogeneous questions. Lastly, in Section 6.3 we describe two approaches which can be used only in very specific scenarios and therefore do not fit neatly into either of the previous categories.

Question:
A 55-year-old woman with small cell carcinoma of the lung is admitted to the hospital to undergo chemotherapy. Six days after treatment is started, she develops a temperature of 38C (100.4F). Physical examination shows no other abnormalities. Laboratory studies show a leukocyte count of 100/mm3 (5% segmented neutrophils and 95% lymphocytes). Which of the following is the most appropriate pharmacotherapy to increase this patient's leukocyte count?

Options: A. Darbepoetin B. Dexamethasone C. Filgrastim D. Interferon alfa E. Interleukin-2 (IL-2) F. Leucovorin

Fig. 5. Example of text only question – a MCQ of a medical exam – from [114].

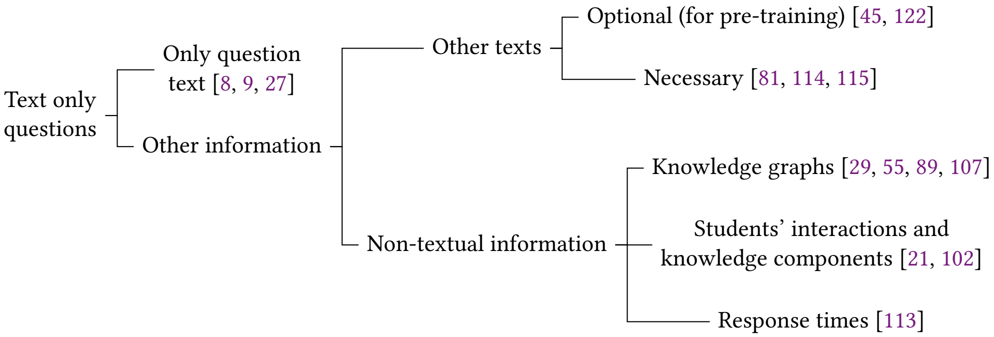

Text only questions
Only question text [8, 9, 27]
Other information
Other texts
Optional (for pre-training) [45, 122]
Necessary [81, 114, 115]
Non-textual information
Knowledge graphs [29, 55, 89, 107]
Students' interactions and knowledge components [21, 102]
Response times [113]

Fig. 6. Categorization of the approaches proposed for text only questions.

6.1 Text only questions

Text only questions (an example can be seen in Figure 5) have received the most attention in recent years, and in Figure 6 we show the categorization of the proposed approaches depending on the information they leverage and how they use it. Some approaches use only the text of the questions [8, 9, 27], while others also use some additional information, which is not part of the questions themselves. The approaches based on question text can be applied in the most scenarios and have the least constraints but, on the other hand, the fact that they cannot leverage additional information might be a limitation if that is available. The works that leverage some kind of additional information can be divided into models that leverage texts from other sources (e.g., books, lectures, etc.), and models that leverage non-textual information. As for the texts, they can be used either as data for pre-training a neural model [45, 122] or as data that is necessary for the implementation of the model [81, 114, 115]; this is an important difference since the need for additional corpora can limit the applicability of some models to scenarios other than the ones which they were built for. Lastly, the non-textual information leveraged by some models can come from different sources: knowledge graphs [29, 55, 89, 107], students' interactions and knowledge components [21, 102], and response times [113].

6.1.1 Models that leverage question text only. An overview of the proposed approaches is shown in Table 6.

Paper	Year	Features	ML model
[27]	2017	Features from *Coh-Metrix* grouped in narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion	Linear regression
[9]	2020	TF-IDF	Random forest regressor
[8]	2020	TF-IDF, linguistic features, readability measures	Random forest regressor

Table 6. Overview of the approaches proposed for QDET of textual questions that use only question text.

The first work [27] studied the correlation between question difficulty and several features obtained from the text of the questions, focusing on science tests from exams administered to 11 years old in England and Wales. Specifically, the authors experiment with over a hundred linguistic indicators generated using the *Coh-Metrix* software [35] and categorised into five dimensions: i) narrativity: extent to which the item uses language comparable to everyday language; ii) syntactic simplicity: the degree to which the item is concise and makes use of simple and familiar syntactic structures; iii) word concreteness: the degree to which the vocabulary use is concrete and meaningful; iv) referential cohesion: the degree of overlap of words and ideas across sentences forming explicit connections; v) deep cohesion: the extent to which the item contains causal and intentional connectives that help the reader build connections and understand relationships and processes in the text. Additionally, the authors experiment with sentence length (number of words per sentence) and paragraph length (number of sentences per paragraph) of descriptive statistics generated by *Coh-Metrix*. A linear regression model is used to estimate the difficulty of the 216 items of the experimental dataset from the aforementioned indicators, and the authors observe that the language variables do not correlate strongly with item difficulty, suggesting that the features from *Coh-Metrix* are not helpful for QDET.

A more recent approach is R2DE (Regressor for Difficulty and Discrimination Estimation) [9], which estimates both the difficulty and the discrimination, as defined in IRT, of MCQ using as input only the text of the questions and the text of the possible choices (both the correct answer and distractors). Experimenting with different input configurations, the authors observed that using the possible answer choices (both the correct answer and the distractors) improves the accuracy of QDET, which is in agreement with other research [45]. From a high-level perspective, R2DE is made of two parallel pipelines, one for QDET and the other for discrimination estimation from text. The two pipelines are architecturally the same but the learned parameters are different; thus, here we focus only on the one that is used for QDET. R2DE first encodes the questions into feature arrays using TF-IDF, removing the tokens that are either too frequent or too infrequent in the corpus; then, it uses these feature arrays as input to a Random Forest regression model, that performs the actual estimation of question difficulty. The authors experiment with other regressors as well, and the Random Forest seemingly outperforms Decision Trees, Linear Regression, and SVM (which is a close second).

An improvement over R2DE was proposed in [8]: in addition to TF-IDF, the authors use readability indexes and linguistic features for the task of QDET of MCQ. Specifically, the authors use the following readability indexes: Flesch reading ease [32], Flesch-Kincaid Grade level [54], Automated Readability Index [86], Gunning FOG Index [37], Coleman-Liau Index [22] and the SMOG Index [68], which are all computed with deterministic formulas from measures such as the number of words and the average word length. As for the linguistic features, the authors compute 9 features related to the number and length of words and sentences in the question and in the answer choices. This improved version still relies on Random Forests for the regression and outperforms the original R2DE model, thus showing that different kinds of features can bring different perspectives to QDET and therefore improve its accuracy. Specifically, the authors present a reduction in RMSE from 0.807 to 0.753, with ZeroR of 0.820. Also, by presenting the results of an ablation study, they show that all groups of features are helpful for the task of QDET, and the ones that bring the most information are the TF-IDF based features.

6.1.2 *Models that leverage additional texts.*

An alternative to using only the question text consists of leveraging additional resources such as books or lecture transcripts, with the constraint that such resources must deal with the same topics that are assessed by the questions. This is a crucial difference from language assessment, as in that case the additional resources can be general domain corpora. Six works explored this area, along two different directions. On the one hand, the models proposed in [7, 45, 122] make use of publicly available pre-trained models and leverage the additional texts only for an additional pre-training, therefore they can be used for QDET even if such additional texts are not available (although in this case they generally lead to worse performance due to the lack of the additional pre-training). On the other hand, the models proposed in [81, 114, 115] have internal components that require such additional data and, if that is missing, cannot be implemented without major modifications to the architecture. This may be seen as a limitation, but it sometimes enables such models to extract more information from the additional texts leading to more accurate QDET. An overview of the proposed approaches is shown in Table 7.

Paper	Year	Other texts necessary	Approach
[45]	2018	-	SVM that uses as features the cosine similarity between the word2vec embeddings of stem, correct choice, and distractors. The additional texts are use to further pre-train the word2vec embeddings.
[122]	2020	-	BERT, the additional texts are used to further pre-train the language model.
[7]	2021	-	BERT and DistilBERT, the additional texts are used to further pre-train the language models.
[114]	2019	✓	Random Forest that uses as features: word embeddings (word2vec, ELMo), linguistic features, Information Retrieval-based features. The additional texts are required to compute some of the features.
[115]	2020	✓	Same as [114]
[81]	2019	✓	Two neural networks, which estimate two components of question difficulty (recall difficulty and confusion difficulty); their estimations are then averaged. The additional texts are used by the "recall" component of the model.

Table 7. Overview of the approaches that use additional texts proposed for QDET of textual questions.

The first paper that leverages additional textual corpora for the task of QDET is [45], in which the authors propose an approach built for MCQ on social sciences in Chinese. It is made of two steps: first, i) a word2vec model is used to obtain semantic vectors representing the question, the correct choice, and the distractors, then ii) the cosine similarities between these vectors are used as input to an SVM classifier that outputs the estimated difficulty. The additional dataset is used to pre-train the word2vec embeddings and, if it is not available, a pretrained word2vec model can be used (likely compromising the accuracy of the model, though). The authors observe that i) there is a negative correlation between the item difficulty and the similarity between stem and answer (i.e., if a stem is similar to the answer, the question is easier) and ii) there is a positive correlation between the item difficulty and the similarity between the correct answer and the distractors (i.e., if the correct answer is similar to the distractors, then the question is more difficult). As for the classification experiments, the proposed model reaches 78% accuracy, with a random baseline of 20% (categorical difficulty on five levels).

Another approach that leverages additional corpora for pre-training is proposed in [122], which is a preliminary work about the effects of using multi-task BERT [24] for performing QDET; specifically dealing with English programming questions. The proposed approach i) starts from the pre-trained BERT model and ii) further pre-train it on a corpus of related documents, finally iii) it fine-tunes it on the task of QDET. The authors model QDET as a binary classification task, and the experimental results show about 76% accuracy (with a random baseline of 50%). A similar approach is evaluated in [7], which experiments with both BERT and DistilBERT [84], observing that the additional pre-training is indeed helpful for the task of QDET and BERT consistently outperforms DistilBERT. The authors also observe that the two Transformers outperform R2DE, thanks to the additional pre-training.

Three papers have presented approaches that leverage additional textual information and cannot be implemented without it. Two of them (from the same team of researchers) focus on the task of QDET for MCQ in high stakes medical exams [114, 115], and evaluate the same model. The proposed approach is divided into two steps: i) first, there is a feature engineering phase, when the input text is converted into feature arrays, then ii) the feature arrays are used as input to a regression model that performs the actual estimation of difficulty. The features can be categorized into three groups: i) word embeddings, ii) other linguistic features, and iii) Information Retrieval (IR) features. As for the embeddings, the authors use word2vec (300 dimensions) and ELMo (1024 dimensions), both pretrained on a corpus of about 22M MEDLINE abstracts[7]. The linguistic features are a set of about 60 values coming from different sources: lexical features, syntactic features, semantic ambiguity features, readability formulae, cognitively-motivated features, word frequency features, and text cohesion features. Lastly, the IR features are obtained from an automated Question Answering system that is trained to respond to the items by retrieving relevant documents from the MEDLINE corpus, and this is the group of features that cannot be implemented without the additional dataset. The authors experiment with different regression models (Random Forest, Linear Regression, SVM, Gaussian processes, Fully Connected Neural Networks), and observe that random forests are the best performing. The authors perform an ablation study and find that all the features are helpful for the estimation, leading to an RMSE of 22.45 (the ZeroR baseline is 23.65). They also find that the IR features are, on their own, the most useful for the estimation. On the other hand, it is apparent that embeddings and linguistic features lead to comparable performance when used singularly, which is somewhat surprising given that the embeddings are obtained with models which model the word semantics.

The other approach to QDET using additional texts was proposed in [81], again targeting MCQs in medical exams. The proposed approach is composed of two neural networks: these are used in parallel to compute different components of question difficulty, which are later averaged to obtain a final difficulty score. The first of them, referred to as Recall Difficulty Module, receives as input i) a corpus of related medical documents, ii) the text of the questions and iii) of the correct choices, and it has the goal of estimating how difficult it is to recall the knowledge assessed by the question; this is the component that cannot be implemented if the additional dataset of related documents is not available. The other component, named the Confusion Difficulty Module, receives as input the stem of the question and the possible choices (both the correct one and the distractors) and has the target of estimating how difficult it is to distinguish between the different choices. Finally, the two components of the difficulty are combined with a weighted average (the weight is learned). The authors observe that, when using only the Recall Difficulty Module or the Confusion Difficulty Module, the error is higher than when using the complete network, although the difference is not great. Specifically, the complete model leads to an RMSE or 0.1311 (difficulty range is $[0; 1]$), the recall module only leads to an RMSE of 0.1319, and the confusion module only leads to an RMSE of 0.1321. The authors also compare the proposed model with a baseline similar to R2DE, building a model that creates feature vectors using TF-IDF and performs the regression with a SVM, which achieves an RMSE of 0.1716 and is clearly outperformed.

6.1.3 Models that leverage knowledge graphs as additional information. Four works perform QDET using a Knowledge Graph (KG, e.g., YAGO2s [96]) as additional source of information [29, 55, 89, 107]; an overview is presented in Table 8.

An important aspect to note is that, in all these papers, the text of the questions (and possibly the text of the choices) is used for Named Entity Recognition (NER) only, to identify the nodes of the KG that are involved in the question. This is a crucial difference from all the other approaches presented in this survey since the difficulty does not really depend on the verbalization of the question but only on the nodes of the graph and the links between them.

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

Paper	Year	Approach
[107]	2015	Difficulty is defined as the similarity between the correct choice and the distractors.
[89]	2017	Logistic regression model that uses 15 features related to i) entity salience (a proxy of entity popularity) and ii) coherence of entity pairs (i.e., their tendency to appear in the same context).
[29]	2018	Defines difficulty as the inverse of the average popularity of the entities in the question.
[55]	2019	Difficulty obtained from the confidence and selectivity of the question.

Table 8. Overview of the approaches to QDET that use knowledge graphs as additional information.

The first of these works [107] estimates the difficulty of an MCQ by observing the similarity between the correct choice and the distractors, assuming that if the distractors are very similar to the correct choice students may find it very difficult to answer the question. Named entities are identified in the texts and linked to corresponding nodes in the KG. The authors define a similarity measure – named Label-set Similarity Ratio (LSR) – to represent the similarity between two nodes depending on their position in the KG. The LSR is not symmetric, therefore the authors define the Closeness between two nodes as the average of the LSR in each direction. Finally, the question difficulty is computed as the average of the Closeness values between the correct choices and the distractors, and later converted into one of three classes (high, medium, and low difficulty). This means that the difficulty obtained as a final result is not based on a testing theory. For the evaluation, the authors compare the difficulty estimated by the proposed model with the difficulty selected by human domain experts and observe that the proposed model reaches an accuracy of 65.3%, with a random baseline of 33.3%.

In [89], QDET is modeled as a binary classification task and a logistic regression model is used for difficulty estimation. The ground truth difficulty is set by human experts. The model leverages 15 features related to two concepts: i) entity salience and ii) coherence of entity pairs. Entity salience is a normalized score that is used as a proxy for an entity's popularity. The entities come from Wikipedia and the authors make use of the link structure within Wikipedia to compute entity salience as the relative frequency with which an entry for an entity is linked to from all other entries. The coherence of entity pairs captures the relative tendency of two entities to appear in the same context. The linear regression model proposed in this paper is capable of reaching a validation accuracy of 66.4% (whereas a random baseline achieves 50%) on a set of 500 questions. The authors also perform an ablation study to demonstrate that both the salience-related features and the coherence-related features are useful for difficulty estimation.

The model proposed in [29] makes use of DBpedia [2] for QDET of MCQ. The proposed model works only for questions which can be modeled as *triples* using entities and relationships in the knowledge base; an example of a triple is: *<London> <capital> <United Kingdom>*. It makes use of the DBpedia PageRank value [101] as a popularity measure, and defines question difficulty as the inverse of the popularity of the triple representing the question (averaging the popularity of the entities in the triple). This difficulty score – which is in the range $[0; 1]$ – is converted to a binary value using a threshold of 0.5. This approach is arguably simpler than the previously proposed ones and this is also visible in the experimental results. The authors recruited 50 participants to evaluate the accuracy of the QDET model, and asked them to rate the difficulty of the questions marked either high or low by the model. While 84.7% of the participants agreed that the easy questions were in fact easy, only 38.5% agreed that hard questions were difficult (average 61.6%).

The latest model in this category is [55], in which the authors address the problem of automatic generation of complex, multi-hop questions over knowledge graphs. As part of the generation process, the authors implement a model for QDET, and here we focus only on that component. The proposed model estimates the difficulty from two measures: i) the confidence and ii) the selectivity of the question. As for the confidence (Con), the authors use that of a

NER model, assuming that higher confidence corresponds to lower difficulty. To measure selectivity (Sel), the authors query Wikipedia with each mention, and use the number of returned hits as an estimation of its selectivity. Finally, question difficulty is computed as $\text{diff} = (1 + \text{Sel})/(1 + \text{Con})$, normalized in the closed interval $[0; 1]$ and converted into a binary vector by thresholding. Evaluation was performed on three publicly available datasets (WebQuestionsSP [117], ComplexWebQuestions [100], and PathQuestion [121]), by asking four participants to judge the accuracy of the difficulty level selected by the model and the results are in line with previous research, with an overall accuracy between 60% and 68%; no baselines are evaluated.

6.1.4 Models that leverage students' interactions and knowledge components. Two papers [21, 102] proposed approaches for QDET using, as additional information, the knowledge components (i.e., topics) associated to each question and the results of students' answers; an overview is presented in Table 9. The fact that such models leverage students' interactions make them unusable for QDET on new items, which have no log of interactions available: indeed, both papers have students' performance prediction as final target, which motivates the need for a history of previous answers.

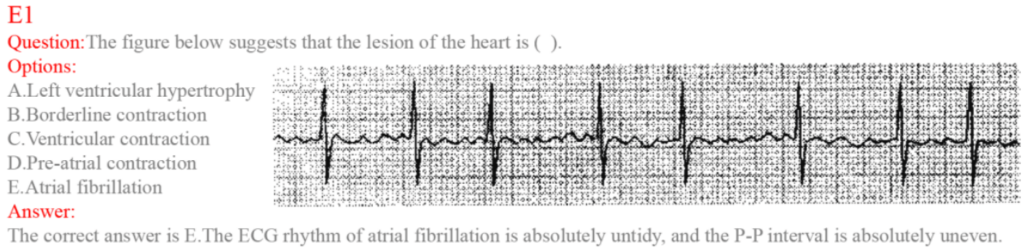
Paper	Year	Approach
[21]	2019	Two components: i) LSTM that receives the text of the question, ii) attention based model that captures relevance between texts and knowledge components. Then, average pooling.
[102]	2020	Pre-trained BERT to embed questions, and TextCNN to perform QDE.

Table 9. Overview of the approaches that use knowledge components and students' interactions as additional information.

The first of these papers [21] proposed DIRT (Deep Item Response Theory), a model that takes inspiration from IRT for estimating the probability that a given student correctly or wrongly answers a question, but relies on neural networks for the estimation of the IRT latent traits (skill level θ, difficulty b, and discrimination a). DIRT is made of three modules: i) an input module, ii) a deep diagnosis module, and iii) a prediction module. For the scope of the current survey, we are interested only in the deep diagnosis module and, specifically, in the component that performs question estimation, therefore we will not present the other components of DIRT. The model used for QDET is made of two parts, which estimate the difficulty from two different perspectives. The first one exploits semantics of question texts for the estimation, which is performed with an LSTM network that receives as input the text of the questions. The second perspective considers the width and depth of knowledge concepts, which is reflected by the relevance between question texts and knowledge concepts. In practice this is done with an attention mechanism, that captures the relationship between question texts and knowledge concepts. Lastly, an average pooling operation is performed to obtain the difficulty. Since the final target of the paper is student answer prediction, there are no experiments to directly compare the estimated difficulty with a ground truth value. However, the authors perform an analysis of the correlation between the estimated difficulty and the correctness of students' answers and observe that the correctness is higher for questions with lower difficulty.

The other approach that leverages students' interactions and knowledge components for the task of QDET was proposed in [102], which has students' answers prediction as final target. The model proposed for QDET is fairly simple: indeed, the authors employ a pre-trained BERT model (without the fine-tuning seen in [7, 122]) for embedding the questions and apply a TextCNN [53] model for QDET. No experiments are performed to directly evaluate QDET.

6.1.5 Models that leverage response times. One paper [113] has proposed a transfer learning based model for QDET of MCQ in medical exams, using question text and response times as features. The proposed model is made of an



E1

Question: The figure below suggests that the lesion of the heart is ( ).

Options:

A.Left ventricular hypertrophy

B.Borderline contraction

C.Ventricular contraction

D.Pre-atrial contraction

E.Atrial fibrillation

Answer:

The correct answer is E.The ECG rhythm of atrial fibrillation is absolutely untidy, and the P-P interval is absolutely uneven.

Fig. 7. Example of a medicine question from [30].

ELMo network, pretrained on the One Billion Word Benchmark [17], followed by an encoding layer added to learn the sequential information from the ELMo embeddings; the encoding layer is made of a BiLSTM. A dense layer then follows the encoding layer to convert the feature vectors to the targets through a non-linear combination of the feature vectors' elements. Considering the target, the model is first trained for response time prediction, and later fine-tuned for the task of QDET. The authors experimented with three different ELMo configurations (*small*, *middle*, and *original*) and various input configurations (*stem only*, *options only*, and *stem and options*). The results indicate that transfer learning can be applied to improve the prediction of question difficulty when response time is used as pretraining and the difficulty is best predicted when using only the item stem. Specifically, the authors target the difficulty obtained with CTT (in the range [0; 100]) and obtain, with the best configuration (i.e., ELMo original, only item stem), an RMSE of 23.32. Contrary to the findings from [8, 9, 45], using the answer options does not increase the performance of the model and actually hinders it, even though the difference is not great (RMSE of 23.43 when using the full item). Also, it is interesting to observe that, even though the best results are obtained with the larger model (i.e., ELMo original, 93.6M parameters), there is not clear correlation between the size of the models and the accuracy of the estimation: indeed, the errors obtained with ELMo middle (20.8M parameters) are generally larger than the ones obtained with ELMo small (13.6M parameters). This is definitely worth some further exploration, especially considering that the model which leads to the best results (ELMo original) is the only for which the inner parameters of ELMo are not updated during QDET training.

6.2 Heterogeneous Questions

Considering the publications that deal with heterogeneous questions (an example is shown in Figure 7), they all focus on questions with accompanying images [30, 94, 119] but one of them [94] also focuses on the effects that tables have on question difficulty. An overview is shown in Table 10.

Paper	Year	Approach
[94]	2016	Studies how the presence of images, tables, formulas, and some textual features (text length, presence of specialist terms and abstract concepts) affect item difficulty.
[30]	2019	ResNet for extracting image representations, BERT for embedding textual content. Capsule Neural network to obtain a fixed-length vector which represents the exercise. Bayesian inference-based softmax regression classifier to perform estimation.
[119]	2019	i) embedding of heterogeneous content (word2vec for texts, convolutional layers for images, fully connected layers for metadata), ii) BiLSTM, iii) self-attention, and iv) max pooling to obtain pre-trained question representations. Fine-tuning on QDE with an FCNN.

Table 10. Overview of the approaches proposed for QDET of heterogeneous questions.

The first work to focus on question images and their effects on the difficulty was [94], which performed a study of how some textual features and the presence of images, tables, and formulas (not their content) affects the IRT difficulty of MCQ in a scientific reasoning exam. Considering textual features, the authors take into consideration text length, and the presence of specialist terms and abstract concepts. By studying the correlation between the aforementioned features and the IRT difficulty of the items, the authors observed that the difficulty is significantly increased by the presence of abstract concepts and specialist terms, suggesting that they might be a good predictor of question difficulty. The presence of images as well has a positive effect on item difficulty, meaning that items that contain visual images tend to be harder to solve. This result is in contrast with previous research [60], and the authors claim that this might happen because the images in the experimental dataset are generally used to show complex scientific models and therefore the increase in difficulty might come from that, not from the images.

Being able to model the content of the images and not only their presence might be very helpful for improving the accuracy of QDET for questions containing images, which is the focus of two papers from 2019. The first of them [30] proposed an approach for predicting the difficulty of visual-textual exercises, using as input both the text and the image of each question. The authors experiment on two datasets, one containing maths questions and the other containing medicine questions. The proposed model is made of two modules: i) a feature extraction module and ii) a difficulty classifier module. The feature extraction module contains two components: a Residual Network [40] for extracting the representation of the images, and a BERT model [24] for embedding the textual content. Since the two vector representations can have different lengths, they are then fed into a Capsule Neural Network [83] to obtain a fixed-length vector which contains the unified representation of the exercise. The fixed-length representation of each exercise is then used as input to the difficulty classifier module, that is a Bayesian inference-based softmax regression classifier and performs the actual estimation of difficulty.

In [119], the authors introduced a general pre-training method – namely QuesNet – to learn question representations that could be fine-tuned for several downstream tasks, one of them being difficulty estimation, similarly to what is done in general purpose pre-trained language models (e.g., BERT). Specifically, the paper focuses on maths MCQ containing images. At a high level, QuesNet is made of three components: i) an embedding module, ii) a content module, and iii) a sentence module. The embedding module projects heterogeneous input content into a unified space, which enables the model to work on inputs from different sources. Specifically, the input can be i) text from the body of the question, ii) an image which is part of the question, and iii) question metadata (e.g., the knowledge components associated with a question). Text embedding is performed with word2vec, image embedding is done with a convolutional neural network, and metadata embedding is performed with a fully connected network. The content module is made of a BiLSTM which receives the concatenation of the vectors produced by the embedding module. Then, the sentence module leverages self-attention for aggregating the item representation vector into a sentence representation; this is done with a multi-head attention module to perform global self attention [106]. Finally, there is a max pooling layer to produce a single vector representing the heterogeneous input image. The proposed architecture is pre-trained with a two level hierarchical approach: first, a masked language model is used as objective for learning low level linguistic features; then, a domain oriented objective is used for learning high level domain logic and knowledge. The embedding modules are pre-trained separately: the text embedding is a word2vec model trained on the specific corpus, the image and metadata embeddings are fully connected neural networks pre-trained using an encoder-decoder architecture and an auto-encoder loss. Once pre-trained, the model can be fine-tuned for specific downstream tasks. Among other tasks, the authors experiment with QDET, and do so by adding a fully connected layer on top of the question embeddings.

6.3 Others

There are two works that do not really fall in any of the previous categories: one of them [78] proposed a model for QDET in the case of questions whose answer is a First Order Logic formulas, the other [72] performs QDET leveraging the variance of students' answers, without looking at the text of the questions.

First Order Logic (FOL) can be defined as a collection of objects, their attributes, and relations among them to represent knowledge. In [78], the authors focus on exercises that ask students to convert a natural language sentence into a FOL (i.e., the answer must be a FOL), and propose an approach that can only be used in this niche. They assume that the overall difficulty of an exercise depends both on the natural language sentence and the FOL formula, therefore the proposed approach uses features from both components. Specifically, the textual features are i) word order matching, ii) anaphoric connectives, iii) negating keywords, iv) special words/phrases, v) quantifier mismatch, and vi) connective mismatch; the FOL features are i) number of quantifiers, ii) type of quantifiers, iii) order of quantifiers, iv) number of implication symbols, and v) number of different connectives. In practice, for the estimation of the difficulty there is no machine learning involved, and the model simply performs a deterministic mapping from the input features to the output difficulty, according to two predefined table curated by the authors: an initial estimation is performed using the FOL features only, then this is further modified according to the NLP features.

In [72], the author observed that the question difficulty can be approximated by the amount of variation in students' answers, and this amount of variation can be computed before grading. The paper deals with questions from the computer science short answers in German corpus [73], which require the student to freely formulate one to three sentence answers. The author models the variance of student answers through the Greedy String Tiling similarity measure [111] (which ranges between 0 and 1), and measures both the variance between students answers and the variation with regard to a reference answer created by a domain expert. It is observed that the variation of answers among themselves is a much stronger prediction than variation with regard to the reference answer. The author also experiments with predicting the question difficulty using the levels from the Bloom's taxonomy [11] and observes that, when available, they are better predictors of question difficulty; however, such levels are not always available and have to be manually selected, therefore the answer variation is a good alternative. The results obtained in this paper are very relevant but are obtained experimenting on 25 questions only, therefore further research is needed to better evaluate the possibility of performing QDET using student answer variation. Also, it would be interesting to evaluate a similar approach on questions of different nature, where the variation resides, for instance, in the choices picked in a MCQ.

7 FURTHER ANALYSIS

In this section, we perform an additional analysis of the papers presented in this survey, focusing on different dimensions with respect to the ones discussed so far. Specifically, we focus on a categorization by features and algorithm (§ 7.1), on evaluation and reproducibility (§7.2), on learning theories and difficulty format (§7.3), on natural language (§7.4), and on the number of publications per year and venue (§7.5). Finally, we collect all the aforementioned information and present it in a single table (Table 13).

7.1 Features and machine learning algorithm

The approaches proposed in previous research are very diverse, and they use different features and machine learning algorithms. In order to provide an overview of their usage in the literature and their popularity, in Table 11 we list several features and algorithms and the papers that use them. This list of features and algorithms is not comprehensive, as we report here only the ones that are shared across different papers; for a recap of all the models we refer to Section 8 and, specifically, to Table 14 and Table 15.

Algorithm/Feature	Papers
Word difficulty	[6, 59, 98, 99]
Reading difficulty or readability indexes	[6, 8, 47, 59, 99]
Word length, passage length, or similar linguistic features	[8, 10, 23, 44, 88, 94, 104, 114–116]
Word frequency	[23, 116]
TF-IDF	[8, 9]
Word2vec embedding	[26, 45, 48, 61, 114–116, 119]
ELMo embedding	[113–115]
Attention-based Neural Network	[21, 30, 48, 102, 119, 122]
Fully Connected Neural Network	[48, 61, 113, 119]
LSTM or BiLSTM	[21, 61, 113, 119]
BERT	[30, 102, 122]
Convolutional Neural Network	[48, 102, 119]
Random Forest	[8, 9, 67, 114, 115]
SVM	[6, 26, 45, 59, 116]
Linear Regression	[27, 44, 104]
Similarity measures	[45, 98, 99, 107]

Table 11. List of papers that use specific features or machine learning algorithms for QDET. The list is not comprehensive as it does not include approaches (e.g., entropy, ResNet) which are used in single papers and not shared between them.

7.2 Evaluation and reproducibility

Performing a quantitative comparison of the models proposed for QDET is extremely challenging, as almost every paper works on its *silo*. First of all, they deal with different educational domains (i.e., Language Assessment, Content Knowledge Assessment, and various subdomains within the two). Secondly, the question format is diverse in different studies (e.g., MCQ and open questions, text only questions and heterogeneous questions, etc.), meaning that the models cannot be directly transferred between scenarios. Also, the learning theories used in each paper and therefore the format and definition of the difficulty itself can be diverse.

Moreover, due to the significant value held by educational data and exam-security concerns, most of the papers do not publicly share the experimental dataset nor the code used for the implementation, which makes a quantitative comparison even more challenging. Only [26] shared the complete dataset used in the paper, and [10, 89, 107] shared a portion of the experimental dataset. As for the code, six papers publicly shared the implementation of their models [8, 9, 55, 59, 115, 119], but none made their trained models available, thus there will inevitably be some degrees of freedom in reproducing the work due to differences in data processing, model parameters and computing resources.

As for the experimental setup, QDET is generally evaluated by comparing the estimated difficulty with a target value, as is common practice in supervised machine learning. However, some papers evaluate QDET using Students' Answers Prediction (SAP): in practice, this consists of using question difficulty to predict the correctness of a student's answer and comparing the predicted correctness with the observed outcome. Table 12 lists the metrics that have been used for evaluation. It is clear that in the papers that model QDET as a regression task Root Mean Squared Error and Pearson's Correlation Coefficient are by far the most common ones, but the choice is not agreed in the research community, as many papers use different metrics. Considering the papers that model difficulty as a discrete variable, Accuracy is by far the most commonly used metric, but again a variety of methods are used. In particular, it is interesting to note that only a few papers use metrics that are not affected by the imbalance classes often found in the evaluation datasets, which is a problem worth addressing, for instance by avoiding the use of accuracy and using more robust metrics.

Metric	Papers
Pearson's Correlation Coefficient	[6, 23, 31, 44, 48, 59, 67, 74, 88, 97, 104, 119]
Root Mean Squared Error	[8, 9, 26, 48, 59, 74, 81, 113–115, 119]
Mean Absolute Error	[8, 9, 114, 119]
R squared (R2)	[72, 97]
Degree Of Agreement	[48, 119]
F1 score	[115]
Mean Squared Error	[9]
Quadratic Weighted Kappa	[59]
Passing Rate	[48]
Spearman rank Correlation Coefficient	[81]
Kendall rank Correlation Coefficient	[81]
Area Under the ROC Curve on SAP	[21, 102]
Accuracy on SAP	[21]
Accuracy	[29, 30, 45, 55, 89, 116, 122]
F1 score	[10, 78, 122]
Pearson's Correlation Coefficient	[98, 99]
Confusion Matrix	[45, 61]
Accuracy on SAP	[47]
Precision	[78]
Recall	[78]
Fleiss Kappa	[89]

Table 12. List of evaluation metrics used in the papers presented in this survey; the top half of the table considers the papers that model difficulty as a continuous variable, the bottom half the ones that model difficulty as a discrete variable.

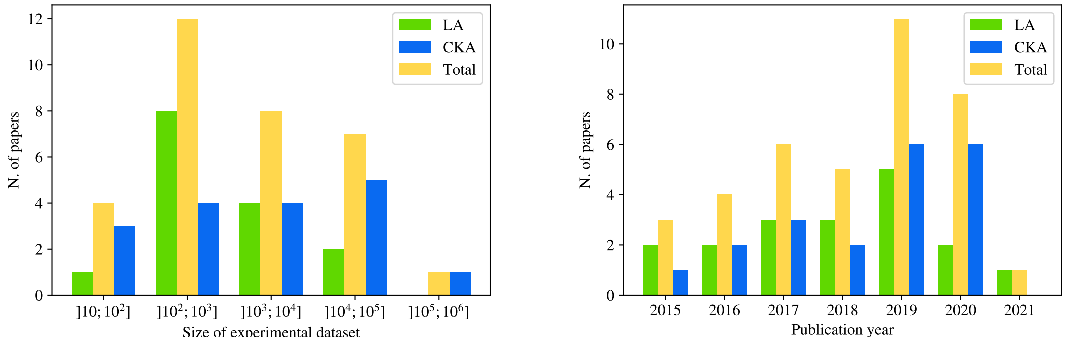
Another aspect to consider is the size of the experimental datasets (i.e., number of questions) used to evaluate model performance; indeed, it is very diverse across the different works. Figure 8a shows the distribution of papers with respect to dataset size, displaying the overall distribution and the distributions for LA and CKA. It is clearly visible that dataset sizes vary widely in these domains: indeed, the average dataset size is about 16,000 across all papers, 28,700 considering the papers working on CKA, and only 3975 considering the papers dealing with LA.

7.3 Learning theories and difficulty format

The majority of papers – 25 – model QDET as a regression task, considering difficulty as a continuous value; only 13 papers consider discrete difficulties, thus modeling QDET as a classification task. Considering the learning theories chosen in each paper, IRT and CTT are by far the most common choices (12 and 14 papers, respectively), but some papers use other approaches: 7 papers work on manually crafted difficulty levels, and 6 papers use other definitions (e.g., CEFR levels, EQDelta).

7.4 Natural language

Considering the natural language of the questions featuring in the task of QDET, there is an overwhelming majority (33) of models built for questions written in English. Other languages occurring in the works presented here are German



(a) Distribution of publications per dataset size.

(b) Distribution of publications per year.

Fig. 8. Distribution of publications per year and dataset size, showing the overall distribution and separately for Language Assessment (LA) and Content Knowledge Assessment (CKA).

(3), Chinese (2), French (1), and Russian (1). It is also interesting to observe that the papers dealing with LA overall deal with more diverse languages, and the papers on CKA have so far worked on English, German (2), and Chinese (1).

7.5 Publications per year and venue

Figure 8b shows the number of published papers per year, and we can see that in recent years there has been an increased interest towards QDET, reflected in the increase in the number of publications. However, if we consider separately the number of publications for the LA domain and the CKA domain, we can see that this increase has not been equally paced. Indeed, while the number of works in LA seems to be fairly constant (except one small spike in 2019), the number of published papers for CKA rapidly increased in the last two years. Most likely, this is due to the rapid improvements in NLP techniques, such as word embeddings and neural network language modelling, which are now being used more frequently for the task of QDET.

In terms of publication venues, the majority of papers have been published at conferences and workshops, and only ten in journals. No single conference or workshop attracts a majority of papers: the venue where the largest number of papers (5) was published is BEA (The Workshop on Innovative Use of NLP for Building Educational Applications). Other venues where more than one of the papers featured have been published are ACL, AIED, CIKM, and LREC.

8 DISCUSSION AND CONCLUSIONS

In this survey, we have investigated the state-of-the-art in research on Question Difficulty Estimation from Text (QDET). We have observed that recent years have witnessed an increased research interest in this domain, which is at least partially due to the concurrent advancements in NLP. Indeed, the research on QDET is shifting from techniques grounded in linguistic and education theory towards approaches that are solely based on recent machine learning models. A very brief recap of all the models evaluated in this survey is provided in Table 14 and Table 15.

Recent work can be separated in two broad categories i) Language Assessment (LA) and ii) Content Knowledge Assessment (CKA), and the task of QDET is often approached with different techniques in the two domains. While research on LA still relies heavily on theoretically grounded techniques (possibly supported by recent machine learning models), work on CKA has shifted towards using almost exclusively learned features and machine learning models.

Paper	Year	Educational Domain	Difficulty format	Testing theory	Only features from text?	Additional features	Additional feat. necessary?	Only Q text	Additional texts domain specific	Additional texts necessary	Natural language	QDE final target
[6]	2015	LA	C	CTT	✗	Predefined tables	✓	✓	-	-	En, Fr, De	✓
[8]	2020	CKA	C	IRT	✓	-	-	✓	-	-	En	✓
[9]	2020	CKA	C	IRT	✓	-	-	✓	-	-	En	✓
[10]	2021	LA	D	Oth.	✓	-	-	✓	-	-	En	✗
[21]	2019	CKA	C	Oth.	✗	Q-matrix, test logs	✓	✓	-	-	En	✗
[23]	2015	LA	C	IRT	✓	-	-	✗	✗	✓	En	✓
[26]	2018	LA	C	IRT	✓	-	-	✗	✗	✗	En	✓
[27]	2017	CKA	C	IRT	✓	-	-	✓	-	-	En	✓
[29]	2018	CKA	D	Oth.	✗	Knowledge graph	✓	✓	-	-	En	✗
[30]	2019	CKA	D	Oth.	✗	Images	✓	✓	-	-	En	✓
[31]	2019	LA	C	Oth.	✓	-	-	✗	✗	✗	En	✓
[44]	2019	LA	C	Oth.	✓	-	-	✓	-	-	Ru	✗
[45]	2018	CKA	D	IRT	✓	-	-	✗	✓	✗	Ch	✓
[48]	2017	LA	C	CTT	✓	-	-	✓	-	-	En	✓
[47]	2018	LA	D	Oth.	✓	-	-	✗	✗	✓	En	✗
[55]	2019	CKA	D	Oth.	✗	Knowledge graph	✓	✓	-	-	En	✗
[59]	2019	LA	C	CTT	✓	-	-	✓	-	-	En	✗
[61]	2019	LA	D	IRT, Oth.	✓	-	-	✗	✗	✓	Ch	✓
[67]	2016	LA	C	Oth.	✗	Audio	✓	✗	✓	✓	En	✓
[72]	2017	CKA	C	IRT	✗	Bloom's taxonomy	✗	✗	✓	✓	De	✗
[74]	2019	LA	C	IRT	✗	Predefined tables	✓	✗	✗	✓	En	✓
[78]	2016	CKA	D	Oth.	✗	FOL	✓	✓	-	-	En	✓
[81]	2019	CKA	C	CTT	✓	-	-	✗	✓	✓	En	✓
[88]	2020	LA	C	Oth.	✓	-	-	✗	✗	✓	En	✓
[89]	2017	CKA	D	Oth.	✗	Knowledge graph	✓	✓	-	-	En	✗
[94]	2016	CKA	C	IRT	✗	Images, tables	✓	✓	-	-	De	✓
[97]	2016	LA	C	CTT, Oth.	✗	JACET8000	✓	✓	-	-	En	✗
[99]	2017	LA	D	CTT	✗	JACET8000	✓	✓	-	-	En	✗
[98]	2020	LA	D	CTT	✗	JACET8000	✓	✓	-	-	En	✓
[102]	2020	CKA	C	CTT	✗	Q-matrix, test logs	✓	✓	-	-	En	✗
[104]	2017	LA	C	IRT	✗	Brown corpus	✓	✓	-	-	En	✓
[107]	2015	CKA	C	Oth.	✗	Knowledge graph	✓	✓	-	-	En	✗
[113]	2020	CKA	C	CTT	✗	Response times	✗	✓	-	-	En	✓
[114]	2019	CKA	C	CTT	✓	-	-	✗	✓	✓	En	✓
[115]	2020	CKA	C	CTT	✓	-	-	✗	✓	✓	En	✗
[116]	2018	LA	D	Oth.	✓	Predefined tables	✓	✗	✗	✓	En	✓
[119]	2019	CKA	C	CTT	✗	Q-matrix, images	✗	✓	-	-	En	✗
[122]	2020	CKA	D	CTT	✓	-	-	✗	✓	✗	En	✓

Table 13. An overview of all the approaches proposed in recent years, showing for each of them: the educational domain (Content Knowledge Assessment or Language Assessment), the format of the difficulty (Continuous or Discrete), the testing theory, the type and source of features leveraged by the model, the natural language under study, and whether QDET was a final or intermediate target in the paper.

Paper	Brief description of the approach to QDET
[6]	SVM that uses 70 features related to i) the difficulty of the text passage, ii) the difficulty of the target word, and iii) test parameters.
[10]	Five features are computed from the question, the passage, and the relation between the two. They are normalized, averaged, and compared to a threshold.
[23]	Found that character length and corpus frequency significantly correlate with vocabulary difficulty.
[26]	Word2vec embeddings given as input to a SVM regressor.
[31]	Built for cloze items; Shannon's entropy is used to assign a score to each gap based on the number of valid words that could fill the gap given the context (candidates obtained with a 5-gram language model); the score is considered as question difficulty.
[44]	Linear regression model that uses as features mean token length and mean sentence length.
[47]	Reading difficulty directly considered as an indication of question difficulty.
[48]	Words are converted to word embeddings with word2vec; sentences are then embedded with a sentence CNN; an attention mechanism is used to detect the relevant parts of the reading passage; lastly the estimation is done with an FCNN. Built for reading comprehension questions.
[59]	SVM that uses 59 features (reduced from the 70 in [6]).
[61]	Input words are converted to word embeddings with word2vec, the sequences of word embeddings are then embedded with LSTM, and the final estimation is done with an FCNN.
[67]	Built for listening comprehension questions; uses 339 raw features obtained from the text (written and spoken) using *TextEvaluator* as input to a random forest.
[74]	Ridge regression, using 36 features from the gap and the context (works on Cued Gap-Filling Items).
[88]	Wighted softmax model that uses as features: word length, log-likelihood from a character-level language model, and Fisher score.
[97]	10 features from target word, reading passage, correct answer, and distractors.
[99]	Features are reading passage difficulty, similarity between correct answer and distractors, and distractor word difficulty. Features can be *low* or *high* and the number of "low" features is the difficulty.
[98]	Features are target word difficulty, similarity between correct answer and distractors, and distractor word difficulty. Features can be *low* or *high* and the number of "low" features is the difficulty.
[104]	Linear regression model that uses as features 25 linguistic variables at passage and item level.
[116]	SVM that uses as features: word length, word frequency, utilization on the web, Age-of acquisition, concreteness rating, number of POS tags, most frequent POS tag, word2vec embeddings, number of double consonants, number of vowels, presence of shorter homophones.

Table 14. The table presents a very brief recap of the models that have been proposed in recent years for QDET in the Language Assessment domain. For a detailed description of all the models, we refer to Section 5.

This difference is not only due to the long history of research in the LA domain, but also to the fact that the "source" of question difficulty is diverse in the two domains. While in LA the difficulty comes from the language itself and the wording, in CKA the majority of the difficulty is related to the topics that are assessed by each question, and only influenced by the wording. Therefore, accurate models in CKA need to be able to model the semantics of the question in order to perform QDET, which is not as important in LA. Although recent neural models are not a panacea for all the problems related to QDET, they can likely bring advantages which have yet to be fully explored, especially in LA.

In this survey we have not explored approaches to QDE which infer question difficulty based on large datasets of interactions between students and items. For instance, by learning from dichotomous pass/fail outcomes we can learn high-dimensional abstract item representations – so-called "skills embeddings" which indicate how items relate to and depend on each other [70]. Or similarly, we can learn how tasks prompt different types of grammatical error in a language learning context, and thus estimate question difficulty based on large datasets of learner essays [120]. Such

Paper	Brief description of the approach to QDET
[8]	TF-IDF, linguistic features, and readability measures as input to a random forest.
[9]	TF-IDF features as input to a random forest.
[21]	End to end neural network made of an LSTM component and an attention-based component.
[27]	Features from *Coh-Metrix* (grouped in narrativity, syntactic symplicity, word concreteness, referential cohesion, and deep cohesion) as input to a linear regression model.
[29]	Difficulty defined as the average popularity of the entities in the question, which is computed using a (required) additional knowledge graph. Text is used only for Named Entity Recognition.
[30]	Works on questions with images. It uses i) ResNet to extract image representations, ii) BERT for embedding textual content, iii) a capsule neural network to obtain a fixed-length array which represents each question, and iv) a Bayesian inference-based softmax regression classifier to perform the numerical estimation.
[45]	SVM that uses as features the cosine smilarity between the word2vec embeddings of the stem, the correct choice, and the distractors.
[55]	Difficulty obtained from the *confidence* and *selectivity* of the question, which are computed from a (required) additional knowledge graph. Text is used only for Named Entity Recognition.
[72]	Difficulty is defined as the variance in the text of students' answers.
[78]	Manually curated table that maps from specific feature values to question difficulty, using features from text and FOL formulas.
[81]	Two neural networks estimate two components of question difficulty (*recall difficulty* and *confusion difficulty*), that are averaged to obtain the difficulty.
[89]	Logistic regression model that uses 15 features related to i) entity salience (a proxy of entity popularity) and ii) coherence of entity pairs (captures their tendency to appear in the same context).
[94]	Studies how the presence of images, tables, formulas, and some textual features (text length, presence of specialist terms and abstract concepts) affect item difficulty.
[102]	Pre-trained BERT to embed questions, and TextCNN to perform QDE.
[107]	Difficulty is defined as the similarity between the correct choice and the distractors.
[113]	Uses pre-trained ELMo embeddings, followed by an encoding layer made of a BiLSTM and a dense layer to convert the feature vectors to the target values. It is trained firstly on response time prediction and subsequently on QDET.
[114]	Random Forest that uses as features: word embeddings (word2vec, ELMo), linguistic features, Information Retrieval-based features.
[115]	Same as [114].
[119]	End-to-end neural network built for heterogeneous questions. It is made of i) an embedding layer for heterogeneous content (word2vec for texts, convolutional layers for images, fully connected layers for metadata), ii) BiLSTM, iii) self-attention, and iv) max pooling to obtain pre-trained question representations. Fine-tuning on QDE with a fully-connected neural network.
[122]	Multi-task BERT.

Table 15. The table presents a very brief recap of the models that have been proposed in recent years for QDET in the Content Knowledge Assessment domain. For a detailed description of all the models, we refer to Section 6.

models can be used to indicate which items students should be shown next, in a recommender system type setting. However, these approaches rely on the availability of plentiful student data, which is not available for new items. A solution could be for these data-driven models to be combined with features extracted from the question text in hybrid approaches which allow continuous QDET updating as student data comes in for new items.

In this survey, we have highlighted one important issue in this domain: the lack of a well established framework for the evaluation of the proposed architectures. Exam questions are generally seen as confidential data that must be kept out of reach from all outsiders, and therefore it is not easy to find publicly available datasets for this task. For similar reasons, there is also a lack of publicly available code. Both these aspects make the evaluation and the comparison of competing algorithms very difficult and thus it is not immediately clear which approaches are best performing in which scenarios, nor the strengths and the weaknesses of each one. We understand that it is generally unfeasible to publicly share exam content, but we believe that the community should work towards publicly sharing the code for re-implementing the proposed models on available datasets such as Eedi's contribution to NeurIPS 2020 Education Challenge [109], Duolingo SLAM dataset [87], or Cambridge English Readability Dataset [112].

Even though there are differences between the models proposed for each category of our taxonomy, there are some observations that hold true across the spectrum of possible domains and might be helpful as guidelines for future research on QDET. First of all, the conclusion from several papers is that that question difficulty is influenced by all the components of the question and, therefore, better performance is possible when all this information is leveraged by the model. For instance, in Multiple Choice Questions both the correct choice and the distractors influence question difficulty, and in reading comprehension questions the reading passage (and its similarity with the question) has a crucial role in determining question difficulty.

Another observation is that additional datasets can be very useful for improving the domain knowledge of the QDE model and thus its accuracy. While this is especially true for CKA – as the model needs to learn the question semantics – and can be done with textual content related to the same topics that are assessed by the questions (e.g., lecture transcripts, books, etc.), several works observed that it is also relevant for LA.

As for the architectures, most of research seems to agree that in CKA the models that lead to the most accurate estimations are end-to-end neural networks (especially when using an attention mechanism, as in Transformers), but they can be limited by the size of the experimental dataset; indeed, they perform best when trained on large amount of data. A promising alternative, which can be trained on less data, are keyword-based approaches such as the ones based on TF-IDF. The observations for LA, on the other hand, are different: indeed, even though some papers successfully leveraged ens-to-end neural networks (especially for reading comprehension questions), others obtain good results using simpler models based on word complexity measures, readability indexes, and other linguistic features. In this sense, we argue that in such cases the computational cost of large neural models is probably not motivated by the reduced improvement (if any) in accuracy.

REFERENCES

[1] Yigal Attali, Luis Saldivia, Carol Jackson, Fred Schuppan, and Wilbur Wanamaker. 2014. Estimating item difficulty with comparative judgments. *ETS Research Report Series* 2014, 2 (2014), 1–8.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[3] Lyle F Bachman et al. 1990. *Fundamental considerations in language testing*. Oxford university press.

[4] David Beglar and Alan Hunt. 1999. Revising and validating the 2000 word level and university word level vocabulary tests. *Language testing* 16, 2 (1999), 131–162.

[5] Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics* 2 (2014), 517–530.

[6] Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. 1–11.

[7] Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of Transformers for estimating the difficulty of Multiple-Choice Questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. 147–157.

[8] Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. Introducing a framework to assess newly created questions with Natural Language Processing. In *International Conference on Artificial Intelligence in Education*. Springer, 43–54.

[9] Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. 2020. R2DE: a NLP approach to estimating IRT parameters of newly generated questions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 412–421.

[10] Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or Complex? Complexity-controllable Question Generation with Soft Templates and Deep Mixture of Experts Model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4645–4654.

[11] Benjamin Samuel Bloom. 1956. Taxonomy of educational objectives: The classification of educational goals. *Cognitive domain* (1956).

[12] Robert F Boldt and Roy Freedle. 1996. Using a neural net to predict item difficulty. *ETS Research Report Series* 1996, 2 (1996), i–19.

[13] Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods* 41, 4 (2009), 977–990.

[14] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46, 3 (2014), 904–911.

[15] Dhawaleswar Rao Ch and Sujan Kumar Saha. 2018. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies* 13, 1 (2018), 14–25.

[16] Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula.* Brookline Books.

[17] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *Fifteenth Annual Conference of the International Speech Communication Association.*

[18] Chih-Ming Chen, Hahn-Ming Lee, and Ya-Hui Chen. 2005. Personalized e-learning system using item response theory. *Computers & Education* 44, 3 (2005), 237–255.

[19] Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*. 167–176.

[20] Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. 84–94.

[21] Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yiying Chen, Haiping Ma, and Guoping Hu. 2019. DIRT: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2397–2400.

[22] Edmund B Coleman. 1965. On understanding prose: some determiners of its complexity. *NSF Final Report GB-2604. Washington, DC: National Science Foundation* (1965).

[23] Brent Culligan. 2015. A comparison of three test formats to assess word difficulty. *Language Testing* 32, 4 (2015), 503–520.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[25] Thomas Eckes and Rüdiger Grotjahn. 2006. A closer look at the construct validity of C-tests. *Language Testing* 23, 3 (2006), 290–325.

[26] Yo Ehara. 2018. Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

[27] Yasmine H El Masri, Steve Ferrara, Peter W Foltz, and Jo-Anne Baird. 2017. Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal* 28, 1 (2017), 59–82.

[28] Charles Elkan. 2005. Deriving tf-idf as a fisher kernel. In *International Symposium on String Processing and Information Retrieval*. Springer, 295–300.

[29] Ainuddin Faizan and Steffen Lohmann. 2018. Automatic generation of multiple choice questions from slide content using linked data. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. 1–8.

[30] Jiansheng Fang, Wei Zhao, and Dongya Jia. 2019. Exercise Difficulty Prediction in Online Education Systems. In *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 311–317.

[31] Mariano Felice and Paula Buttery. 2019. Entropy as a proxy for gap complexity in open cloze tests. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 323–327.

[32] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[33] Yifan Gao, Lidong Bing, Wang Chen, Michael Lyu, and Irwin King. 2019. Difficulty Controllable Generation of Reading Comprehension Questions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 4968–4974. https://doi.org/10.24963/ijcai.2019/690

[34] Siddhant Garg and Alessandro Moschitti. 2021. Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7329–7346.

[35] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.

[36] David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia* 4, 1 (2003), 34.

[37] Robert Gunning et al. 1952. Technique of clear writing. (1952).

[38] Ronald K Hambleton and Russell W Jones. 1993. Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice* 12, 3 (1993), 38–47.

[39] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory.* Sage.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[41] Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*. 187–197.

[42] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[43] Sebastian Hoffmann, Stefan Evert, Nicholas Smith, David Lee, Ylva Berglund-Prytz, et al. 2008. *Corpus linguistics with BNCweb-a practical guide*. Vol. 6. Peter Lang.

[44] Jue Hou, Koppatz Maximilian, José María Hoya Quecedo, Nataliya Stoyanova, and Roman Yangarber. 2019. Modeling language learning using specialized Elo rating. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 494–506.

[45] Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management* 54, 6 (2018), 969–984.

[46] Yi-Ting Huang, Hsiao-Pei Chang, Yeali Sun, and Meng Chang Chen. 2011. A robust estimation scheme of reading difficulty for second language learners. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*. IEEE, 58–62.

[47] Yi-Ting Huang, Meng Chang Chen, and Yeali S Sun. 2018. Development and Evaluation of a Personalized Computer-aided Question Generation for English Learners to Improve Proficiency and Correct Mistakes. *arXiv preprint arXiv:1808.09732* (2018).

[48] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[49] Zhenya Huang, Qi Liu, Chengxiang Zhai, Yu Yin, Enhong Chen, Weibo Gao, and Guoping Hu. 2019. Exploring multi-objective exercise recommendations in online education systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1261–1270.

[50] Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, Guoping Hu, et al. 2019. EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[51] Syeda Asima Iqbal and Syeda Anila Komal. 2017. Analyzing the Effectiveness of Vocabulary Knowledge Scale on Learning and Enhancing Vocabulary through Extensive Reading. *English Language Teaching* 10, 9 (2017), 36–48.

[52] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* (1972).

[53] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[54] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Technical Report. Naval Technical Training Command Millington TN Research Branch.

[55] Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *International Semantic Web Conference*. Springer, 382–398.

[56] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior research methods* 44, 4 (2012), 978–990.

[57] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30, 1 (2020), 121–204.

[58] Suzanne Lane, Mark R Raymond, and Thomas M Haladyna. 2015. *Handbook of test development*. Routledge.

[59] Ji-Ung Lee, Erik Schwan, and Christian M Meyer. 2019. Manipulating the Difficulty of C-Tests. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 360–370.

[60] SC Leong. 2006. On varying the difficulty of test items. In *32nd Annual Conference of the International Association for Educational Assessment, Singapore*. 21–26.

[61] Li-Huai Lin, Tao-Hsing Chang, and Fu-Yuan Hsu. 2019. Automated Prediction of Item Difficulty in Reading Comprehension Using Long Short-Term Memory. In *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 132–135.

[62] Wim J Linden, Wim J van der Linden, and Cees AW Glas. 2000. *Computerized adaptive testing: Theory and practice*. Springer.

[63] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 923–929.

[64] Jing Liu, Quan Wang, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. Question difficulty estimation in community question answering services. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 85–90.

[65] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. 2018. Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1821–1830.

[66] Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. 63–70.

[67] Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 3245–3253.

[68] G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading* 12, 8 (1969), 639–646.

[69] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*. 3111–3119.

[70] Russell Moore, Andrew Caines, Mark Elliott, Ahmed Zaidi, Andrew Rice, and Paula Buttery. 2019. Skills Embeddings: a Neural Approach to Multicomponent Representations of Students and Tasks. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM)*.

[71] ISP Nation and P Teaching. 1983. Learning vocabulary. *New Zealand Language Teacher* 9, 1 (1983), 10–11.

[72] Ulrike Padó. 2017. Question difficulty–How to estimate without norming, how to use for automated grading. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*. 1–10.

[73] Ulrike Pado and Cornelia Kiefer. 2015. Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the fourth workshop on NLP for computer-assisted language learning*. 42–50.

[74] Irina Pandarova, Torben Schmidt, Johannes Hartig, Ahcène Boubekki, Roger Dale Jones, and Ulf Brefeld. 2019. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education* 29, 3 (2019), 342–367.

[75] Shalini Pandey and Jaideep Srivastava. 2020. RKT: Relation-Aware Self-Attention for Knowledge Tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1205–1214.

[76] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, 11 (1901), 559–572.

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

[78] Isidoros Perikos, Foteini Grivokostopoulou, Konstantinos Kovas, and Ioannis Hatzilygeroudis. 2016. Automatic estimation of exercises' difficulty levels in a tutoring system for teaching the conversion of natural language into first-order logic. *Expert Systems: The Journal of Knowledge Engineering* 33, 6 (2016), 569–580.

[79] Kyle Perkins, Lalit Gupta, and Ravi Tammana. 1995. Predicting item difficulty in a reading comprehension test with an artificial neural network. *Language testing* 12, 1 (1995), 34–53.

[80] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2227–2237.

[81] Zhaopeng Qiu, Xian Wu, and Wei Fan. 2019. Question difficulty prediction for multiple choice problems in medical exams. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 139–148.

[82] Georg Rasch. 1960. *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

[83] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3859–3869.

[84] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[85] Norbert Schmitt, Diane Schmitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing* 18, 1 (2001), 55–88.

[86] RJ Senter and Edgar A Smith. 1967. *Automated readability index.* Technical Report. CINCINNATI UNIV OH.

[87] Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*. 56–65.

[88] Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics* 8 (2020), 247–263.

[89] Dominic Seyler, Mohamed Yahya, and Klaus Berberich. 2017. Knowledge questions from knowledge graphs. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 11–18.

[90] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* 5, 1 (2001), 3–55.

[91] Kathleen M Sheehan, Michael Flor, and Diane Napolitano. 2013. A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*. 49–58.

[92] Kathleen M Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The TextEvaluator Tool: Helping Teachers and Test Developers Select Texts for Use in Instruction and Assessment. *The Elementary School Journal* 115, 2 (2014), 184–209.

[93] Bernard Spolsky. 1969. Reduced redundancy as a language testing tool. (1969).

[94] Jurik Stiller, Stefan Hartmann, Sabrina Mathesius, Philipp Straube, Rüdiger Tiemann, Volkhard Nordmeier, Dirk Krüger, and Annette Upmeier zu Belzen. 2016. Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education* 41, 5 (2016), 721–732.

[95] Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. 2018. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[96] Fabian M Suchanek, Johannes Hoffart, Erdal Kuzey, and Edwin Lewis-Kelham. 2013. Yago2s: Modular high-quality information extraction with an application to flight planning. *Datenbanksysteme für Business, Technologie und Web (BTW) 2048* (2013).

[97] Yuni Susanti, Hitoshi Nishikawa, Takenobu Tokunaga, and Obari Hiroyuki. 2016. Item Difficulty Analysis of English Vocabulary Questions. In *Proceedings of the 8th International Conference on Computer Supported Education*. 267–274.

[98] Yuni Susanti, Takenobu Tokunaga, and Hitoshi Nishikawa. 2020. Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning* 15 (2020), 1–22.

[99] Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2017. Controlling item difficulty for automatic vocabulary question generation. *Research and practice in technology enhanced learning* 12, 1 (2017), 1–16.

[100] Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 641–651.

[101] Andreas Thalhammer and Achim Rettinger. 2016. PageRank on Wikipedia: towards general importance scores for entities. In *European Semantic Web Conference*. Springer, 227–240.

[102] Hanshuang Tong, Yun Zhou, and Zhen Wang. 2020. Exercise Hierarchical Feature Enhanced Knowledge Tracing. In *International Conference on Artificial Intelligence in Education*. Springer, 324–328.

[103] Hanshuang Tong, Yun Zhou, and Zhen Wang. 2020. HGKT: Introducing Problem Schema with Hierarchical Exercise Graph for Knowledge Tracing. *arXiv preprint arXiv:2006.16915* (2020).

[104] Jonathan Trace, James Dean Brown, Gerriet Janssen, and Liudmila Kozhevnikova. 2017. Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing* 34, 2 (2017), 151–174.

[105] Toshihiko Uemura and Shinichiro Ishikawa. 2004. JACET 8000 and Asia TEFL vocabulary initiative. *Journal of Asia TEFL* 1, 1 (2004), 333–347.

[106] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[107] Ellampallil Venugopal Vinu et al. 2015. A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Journal of Web Semantics* 34 (2015), 40–54.

[108] Tzu-Hua Wang. 2014. Developing an assessment-centered e-Learning system for improving student learning effectiveness. *Computers & Education* 73 (2014), 189–203.

[109] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. 2020. Diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061* (2020).

[110] Walter D Way. 1998. Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice* 17, 4 (1998), 17–27.

[111] Michael J Wise. 1996. YAP3: Improved detection of similarities in computer program and other texts. In *Proceedings of the twenty-seventh SIGCSE technical symposium on Computer science education*. 130–134.

[112] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, San Diego, CA, 12–22. https://doi.org/10.18653/v1/W16-0502

[113] Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 193–197.

[114] Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 11–20.

[115] Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2020. Predicting Item Survival for Multiple Choice Questions in a High-stakes Medical Exam. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 6812–6818.

[116] Hua Yang and EUM Suyong. 2018. Feature Analysis on English word difficulty by Gaussian Mixture Model. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 191–194.

[117] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 201–206.

[118] Yu Yin, Zhenya Huang, Enhong Chen, Qi Liu, Fuzheng Zhang, Xing Xie, and Guoping Hu. 2018. Transcribing content from structural images with spotlight mechanism. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2643–2652.

[119] Yu Yin, Qi Liu, Zhenya Huang, Enhong Chen, Wei Tong, Shijin Wang, and Yu Su. 2019. Quesnet: A unified representation for heterogeneous test questions. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1328–1336.

[120] Ahmed H Zaidi, Andrew Caines, Christopher Davis, Russell Moore, Paula Buttery, and Andrew Rice. 2019. Accurate Modelling of Language Learning Tasks and Students Using Representations of Grammatical Proficiency.. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*.

[121] Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An Interpretable Reasoning Network for Multi-Relation Question Answering. In *Proceedings of the 27th International Conference on Computational Linguistics*. 2010–2022.

[122] Ya Zhou and Can Tao. 2020. Multi-task BERT for problem difficulty prediction. In *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*. IEEE, 213–216.