# A Systematic Review of Data-driven Approaches to Item Difficulty Prediction

Samah AlKhuzaey[0000−0001−8883−1172], Floriana Grasso[0000−0001−8419−6554], Terry R. Payne[0000−0002−0106−8731], and Valentina Tamma[0000−0002−1320−610X]

University of Liverpool, Liverpool, UK, L69 3BX
{S.Alkhuzaey, F.Grasso, T.R.Payne, V.Tamma}@liverpool.ac.uk

**Abstract.** Assessment quality and validity is heavily reliant on the quality of items included in an assessment or test. Difficulty is an essential factor that can determine items and tests' overall quality. Therefore, *item difficulty prediction* is extremely important in any pedagogical learning environment. Data-driven approaches to item difficulty prediction are gaining more and more prominence, as demonstrated by the recent literature. In this paper, we provide a systematic review of data-driven approaches to item difficulty prediction. Of the 148 papers that were identified that cover item difficulty prediction, 38 papers were selected for the final analysis. A classification of the different approaches used to predict item difficulty is presented, together with the current practices for item difficulty prediction with respect to the learning algorithms used, and the most influential difficulty features that were investigated.

**Keywords:** Difficulty prediction · Item difficulty · Question difficulty · Systematic review · Difficulty modelling · Difficulty estimation.

## 1 Introduction

Student assessments are a fundamental component of any pedagogical learning environment. Assembling tests that contain items (i.e. questions) which measure the various types of skills of different levels of learners in a fair way is a challenging task. Teachers and item writers must ensure the consistent quality of assessment materials to provide objective and effective evaluation.

Assessment quality and validity is heavily reliant on the quality of items included in the test; therefore, significant effort and resources have been devoted to *item analysis tasks*. For item writers, item analysis is of great importance as it allows them to improve items' overall quality by eliminating non-functional items [30]. *Difficulty* is an essential factor that can determine the overall quality of items and tests, whereas *item difficulty* refers to the estimation of the skill or knowledge level needed by students to answer an item [13]. Thus, difficulty calibration is crucial in the assessment construction process; to provide equitable opportunities to all test takers in any assessment, the item selection process must be conducted according to the difficulty level of each item [34]. Designing unbalanced tests which contain arbitrary numbers of easy and difficult items can result in significant disadvantages to test takers who are affected by assessment-based decisions. For example, assessments that consist of mostly *easy* items will result in wrongly qualifying and certifying those less-than-competent test takers.

Traditional methods for obtaining an a priori estimation of difficulty rely on two methods [10, 39]: i) pretesting and ii) experts' judgement. However, such approaches are frequently criticised in the literature for being costly, time-consuming, subjective and difficult to scale [6, 20, 29]. Therefore, a number of alternative methods have been considered to overcome these limitations.

In this paper, we will examine the item difficulty prediction literature with a special focus on data-driven approaches. To the best of our knowledge, there has been no such review, nor a summary of the empirical evidence established so far in this emerging research area. More specifically, the following research questions will be addressed:

**RQ1:** What AI-based computational models are currently developed to offer a priori difficulty prediction?
**RQ2:** What are the most investigated domains and item types?
**RQ3:** What are the influential features that were found to affect difficulty?

We provide a overview of the research on item difficulty estimation in Section 2. We then present the method by which the systematic review was conducted (Section 3), before discussing research questions and how they fit the literature within the review (Section 4). We then conclude in Section 5.

## 2 Background

The research on item difficulty estimation is extensive and well-established. Psychometricians, educational psychologists and linguists have long been studying the potential sources of difficulty in educational items. These fields have provided theoretical frameworks of cognitive processes involved in assessments. Furthermore, statistical methods and manual coding practices have been applied to extract features and explore the relationship between different variables. More recently however, AI techniques such as neural networks, natural language processing (NLP), expert systems and machine learning algorithms have transformed the field by applying unconventional concepts of non-linear modelling, linguistic pattern recognition and advanced predictive power. We present a classification of two opposing approaches to item difficulty prediction based on a comprehensive survey of the literature; that of *cognitive* and *systematic* approaches (Figure 1).

*Cognitive approaches* include methods that address difficulty on the cognitive level by examining what cognitive abilities are required to answer an item correctly. These approaches are qualitative in nature and rely on pre-defined notions of difficulty, based on educational taxonomies or heuristic methods which define difficulty according to the perceptions of educators, item writers and/or learners. In contrast, *systematic approaches* focus on quantifying the concept of difficulty by employing more objective techniques found in statistical or data-driven prediction models. Some of the most employed statistical methods are

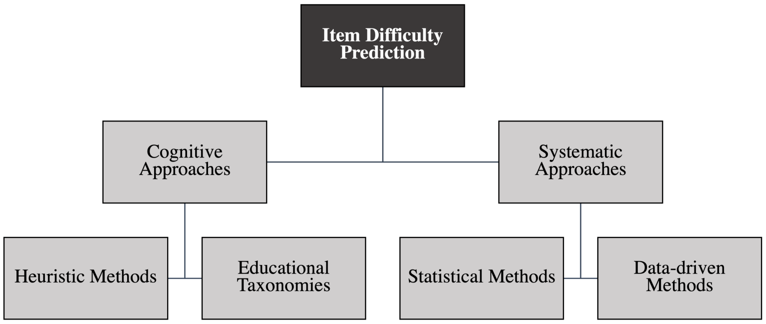

**Fig. 1.** Item Difficulty Prediction Approaches

psychometric statistical models that analyse the relationship between difficulty and examinees' latent traits. Furthermore, basic statistical models (e.g. regression) have also been used to examine the relationship between difficulty and various variables [11, 24, 32, 42]. Despite the fact that, in this approach, researchers were using data to draw conclusions, it is nonetheless heavily theory driven. Difficulty variables were either produced by experts or identified from previous theories in the literature. Moreover, feature extraction processes are typically conducted manually by domain experts in this type of investigation.

More recently, there has been a focus on employing data-driven approaches that represent an array of methods and techniques used to quantify and objectify the process of difficulty prediction. This line of investigation strives to eliminate or at least reduce any subjectivity caused by human intervention [21], and do not necessarily require domain experts to label or define difficulty features. Moreover, pre-testing the items to an appropriate sample will not be needed if automatic methods prove its validity. Hence, data-driven approaches (which include techniques such as NLP, rule-based and machine learning algorithms) are gaining more and more prominence [3, 5, 10, 17, 20, 21, 28].

In this paper, we provide a general overview of the broader field of item difficulty prediction in order to gain a full understanding of the research area. However, the scope of this review will only include data-driven approaches which incorporate computational models to model difficulty.

## 3 Review Method

This review's protocol is informed by the guidelines provided in [25], and is illustrated in Figure 2. The search process was conducted manually using the following paper archives: IEEE[1], ACM Digital Library[2], ScienceDirect[3], Springer[4],

[1] https://ieeexplore.ieee.org/Xplore/home.jsp
[2] https://dl.acm.org/
[3] https://www.sciencedirect.com/
[4] https://www.springer.com/

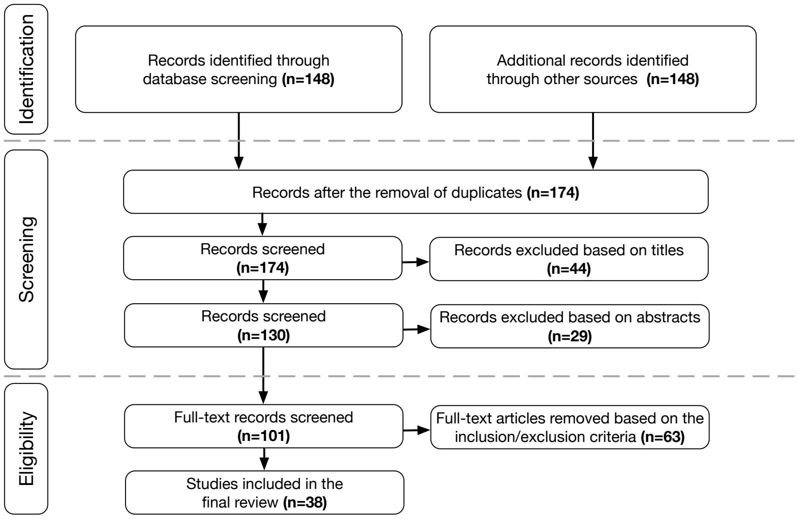
Identification
Screening
Eligibility
Records identified through database screening (n=148)
Additional records identified through other sources (n=148)
Records after the removal of duplicates (n=174)
Records screened (n=174)
Records excluded based on titles (n=44)
Records screened (n=130)
Records excluded based on abstracts (n=29)
Full-text records screened (n=101)
Full-text articles removed based on the inclusion/exclusion criteria (n=63)
Studies included in the final review (n=38)

**Fig. 2.** The study selection process

and Elsevier[5]. Additional papers were included in the search by examining the 'related work' and the 'reference list' sections of each identified paper. Also, general and academic search engines such as Google search and Google Scholar were included to identify relevant papers. We also considered the citations to certain papers by using the 'cited by' option in Google Scholar to include papers which were not identified by the previous methods. The search process identified 148 papers which were screened in three stages: 1) title and abstract screening, 2) full-text screening, and 3) inclusion and exclusion criteria-based filtering. As a result, 38 papers were included for the final analysis.

Papers focusing on data-driven approaches to item difficulty prediction were included without constraints on publication year, paper type, domain or item type. Papers were excluded if they violated one or more of the following criteria:

- The paper is not written in English.
- The full text of the paper is not available.
- The proposed prediction model is not evaluated.
- The difficulty model is not data-driven. We exclude papers that predict difficulty based on heuristic, statistical or educational taxonomies approaches.
- The paper estimates difficulty *after* administrating the test. We only focus on methods which offer a priori prediction of difficulty in order to overcome limitations of traditional prediction methods.
- The items are not textual (i.e. containing images, graphs or formulas). We exclude these types of items as they require different analytical techniques compared to textual items.

[5] https://www.elsevier.com/

- The paper does not address assessment items. For example, we exclude studies that predict the difficulty of questions in question answering communities such as Stack Overflow as this type of question differs completely from assessment questions with regard to their purpose, style and structure.
- The difficulty features are not extracted from items. We focus on difficulty features that are derived from items' structure, hence, we exclude features which are extracted from other data sources such as eLearning environments or sensors.
- The paper focuses on item classification based on features other than difficulty. For instance, we exclude papers that classify items based on question type.

The field of item difficulty estimation is an interdisciplinary one. Relevant fields such as educational assessment, psychology and computer science use different-yet synonymous terms to address the same tasks. Therefore, different combinations of search terms were assembled. As a result, the following combinations of keywords and operators were used:

*Item difficulty prediction, Item difficulty estimation, Item difficulty modelling, Difficulty modelling, (item OR question) AND difficulty AND (estimation OR prediction OR modelling)*

A specific form was designed for the data extraction process given the objectives of this review, which included: *title, year of publication, method/approach, domain, item type, number of items, data, evaluation, participant, metrics, difficulty feature, results, paper type, publication venue and quality score.* Eight quality assessment criteria were adopted from [50], where *reporting quality*, *rigour* and *credibility* were the most frequently assessed dimensions in software engineering systematic reviews. The quality assessment process was conducted after reading the full text and after completing the data extraction with values assigned as Yes= 1, No= 0 and Partly= 0.5.

**Included Papers:** [1] [2] [3] [4] [5] [6] [8] [9] [10] [12] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [26] [27] [28] [29] [33] [35] [36] [37] [38] [40] [41] [43] [44] [45] [46] [47] [48] [49]

## 4 Results and Discussion

### 4.1 RQ1: Data-driven Item Difficulty Prediction

In this section, we address the question: *What AI-based computational models are currently developed to offer a priori difficulty prediction?* The computational models used in the prediction process could be discussed under two headings: *machine learning* and *rule-based modelling.* The majority of papers considered utilise machine learning algorithms such as neural networks and support vector machine (SVM) [14, 19–21, 38], with NLP being used to perform automatic extraction of difficulty features. Neural networks were some of the first data-driven

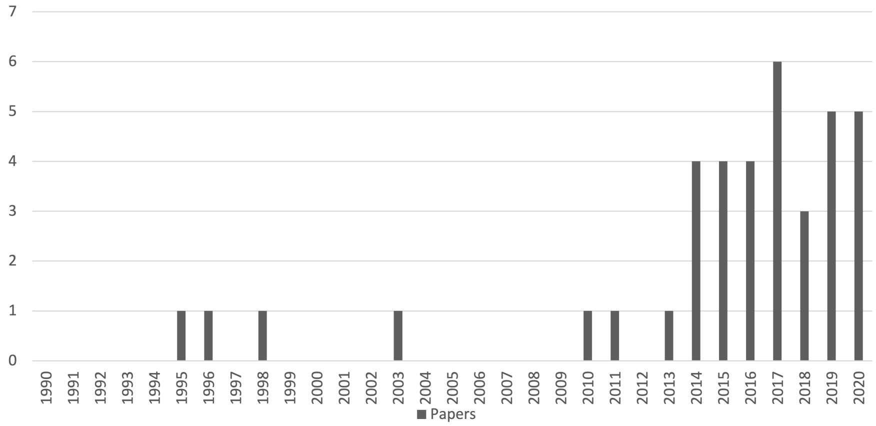

**Fig. 3.** Number of publications distributed by year

methods to be implemented in the item difficulty prediction field. In 1995 they were used to predict the difficulty of reading comprehension items taking from a TOEFL test [37], with the aim of exploring an unconventional approach that could outperform statistical approaches. Rule-based algorithms were also used, but relied on hard-coded instructions which do not follow a pre-defined algorithm [15, 36]. For this type of modelling, difficulty features were manually identified and extracted by experts, which were represented in the form of rules.

It is clear from Figure 3, which depicts the number of publications over time, that publishing in this research area progressed through two different stages. The first wave of publications started in the mid-1990s by employing neural networks which, at the time, represented a novel approach for exploring non-linear relationships between item parameters and difficulty. Previous research had to this point only employed statistical approaches, which explains the relationships in a linear manner [8, 9, 12, 37]. The second wave of studies started in 2010 as researchers began to explore different data-driven approaches to this problem, such as rule-based expert systems, support vector machine (SVM) and Naïve Bayesian models [4, 22, 35]. A steep increase in publications is noticeable from 2014 to 2020, especially in 2017, 2019 and 2020, suggesting a growing relevance of machine learning in the item difficulty research community.

In [21], a data-driven approach was employed to predict the difficulty of 30,000 reading comprehension items collected from a standard English test. The item, options and the reading passages were analysed for each item. Sentence representations were then extracted from the item components using a CNN-based architecture. Finally, the difficulty level was determined by aggregating the semantic representation of all items' components. In a different study, the authors investigated whether item difficulty correlates with the semantic similarity between item components [20]. To achieve this, they utilised word embeddings to construct the semantic space of learning materials and obtained the semantic vectors of item elements. The semantic similarity scores were used as an input to a SVM model to predict the difficulty of items collected from Entrance Exams in the social studies domain. This contrasts with [35], where a difficulty estimation approach was presented which attempted to estimate the difficulty of converting natural language sentences into First Order Logic (FOL) formulae. An expert system was then employed to estimate the difficulty level of exercises based on several parameters for measuring the complexity of the conversion process, such as the connectives of the FOL expressions.

In general, there are four key architectural components that item difficulty prediction models have in common, which represent four fundamental tasks:

**Observed difficulty measurement:** where the ground-truth difficulty is measured using psychometric models or labeled by experts for later comparison with the predicted difficulty.

**Pre-processing:** where textual data is prepared for use by removing irrelevant words and producing well-defined pieces of text.

**Feature extraction:** this is used to transform text into machine-processable representations. Various NLP techniques are used in this step such as Bag-of-Words, Word2vec and TF-IDF.

**Prediction Model:** the specified machine learning algorithm is used to analyse the data.

### 4.2 RQ2: Domains and Item Types

With respect to the question: *What are the most investigated domains and item types?* we found that the majority of papers on data-driven difficulty prediction are domain specific (Figure 4). Language learning is the most frequently investigated domain [3, 14, 21, 33], followed by Computer Science [17, 35] and Medicine [18, 26, 38]. This contrasted with other domains, such as Mathematics and Social Studies, which appeared in a minority of cases[20, 23]. The popularity of the language learning and medical domains could be explained by the existence of several standardised test-organisations that offer international and national language proficiency tests (e.g. TOELF or IELTS), and medical licensing examinations which require a massive number of frequently updated items. Difficulty calibration is considered a fundamental process in these types of tests as it ensures fairness and comparability of high-stakes exams, which are used to inform important decisions regarding certification and employment.

Domain-independent (i.e. *generic*) studies accounted for almost 27% of the publications that we examined. The main rationale for investigating domain-independent studies is the possibility of producing difficulty prediction frameworks that are generalisable, and that could be applied to other domains.

The types of item formats investigated included Multiple Choice Questions (MCQs), true/false questions, gap-filling, and factual items in addition to other types (Figure 5). MCQs represented the majority of item types studied; due to the ability to explore different sources of difficulty by analysing the relationship between item components such as item stem, distractor and correct responses.

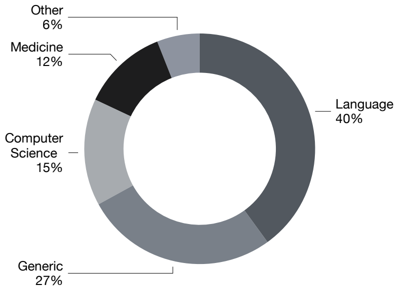

**Fig. 4.** Most investigated domains

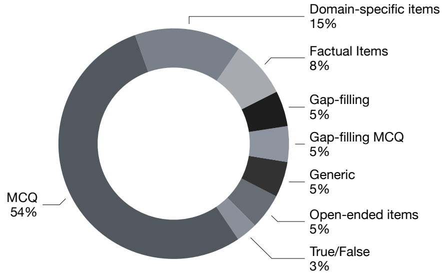

**Fig. 5.** Most investigated item types

### 4.3 RQ3: Difficulty Features

The third question we investigated poses the question *What are the influential features that were found to affect difficulty?* Educational items are natural language phrases constructed by experts to assess a certain skill. When investigating the sources of difficulty in textual items, textual complexity plays an important role. The underlying theory is that more textually complex items require more advanced language proficiency skills in order to read, comprehend and correctly answer items. Therefore, linguistic features are considered the most obvious sources of difficulty when studying textual items. Recent studies on item difficulty prediction use NLP and text mining techniques to automatically extract syntactic and semantic features of items [4, 6, 7, 31, 40].

Linguistic features provide information regarding two levels of language: *syntactic* and *semantic*. The relationship between difficulty and linguistic variables have previously been extensively studied and focus mainly on syntactic features [9, 19, 35, 37]. More recently, researchers have started examining semantic-related factors by exploring semantic relevance and semantic similarity between item components (stem, distractor and correct answers) [20, 28, 38, 46].

Before discussing linguistic features in depth, it is worth mentioning a type of feature that was observed in four other studies [18, 22, 33, 37]. Psycholinguistic variables were examined to explore the affect of cognitive aspects of language on item difficulty. Such features are concerned with how words or sentences are constructed, processed and interpreted by the brain. For example, the Age of Acquistion (AOA) variable (which refers to the age at which a certain word is learned) was examined in two studies to evaluate its affect on difficulty. Other psycholinguistic features included word concreteness and word imageability.

**Syntax-based Features:** Structure-level features refer to linguistic components that govern the textual structure of an item. This level of language typically incorporates syntactic, lexical and grammatical components. The main motivation behind analysing this type of feature is to determine the underlying characteristics which indicate the level of textual complexity and readability. Moreover, this source of difficulty is estimated by considering word- or sentence-level measures, achieved by counting words, sentences and syllables and examining the relationship between these textual components. Table 1 summarises the most common syntactic features. For example, [29] noted that the lexical frequency of the words was the best predictor of difficulty. Another study found that the part-of-speech (POS) count could accurately predict item difficulty [40]. Word count is the most common feature investigated; in many studies, it is referred to as *word frequency* or *word familiarity*, as both terms include counting the number of words to examine the frequency of the word or its familiarity. Word count can target special words types such as verbs, nouns, negation and named entities. Furthermore, some studies further examined the frequency of complex types of words which require advanced cognitive skills; for example, academic, complex and uncommon words. This is also the case for sentence-level analysis which utilises measures to count the number of sentences or special types of sentences (e.g. type of clause) to assess the complexity level of an item.

**Table 1.** Common Syntax-based Difficulty Features

| **Syntactic Difficulty Feature** | **Studies** |
|---|---|
| Word count | [2] [3] [5] [6] [9] [10] [12] [18] [33] [37] [40] |
| Frequency of complex words | [10] [18] [29] [37] |
| Word length | [3] [4] [5] [10] [12] [18] [19] [29] [33] |
| POS count | [3] [19] [40] |
| Grammatical forms | [2] [3] [18] [33] [37] |
| Negation count | [10] [18] |
| Verb variation | [3] [10] |
| Sentence length | [3] [4] [18] [19] [21] [33] [38] [43] |
| Sentence count | [5] [10] [18] [33] [37] |
| Type of clause | [10] [18] [33] |

Another proxy of textual complexity is the word/sentence length [16, 19]. It is believed that long words/sentences are more difficult to understand than shorter ones. Therefore, utilising measures to count the number of characters in a word or words in a sentence is very common in the literature. Separating content words from function words is the main purpose for using part of speech (POS) tagging measures. This distinction is necessary to identify content words which represent lexical meaning and function words that represent syntactic relations. Further analysis would incorporate POS counting to count the number of appearances of each POS tag (e.g. verbs, nouns and pronouns) in order to explore features like verb variation which increases text complexity.

**Semantic-based Features.** The second type of features focus on the relationship between difficulty and semantic properties of an item or its components (see Table 2). Features that address this level of language were absent in many earlier studies. However, more recently there has been a recognition of the importance of deeper levels of analysis for examining sources of difficulty at the semantic level. *Semantic similarity* is the most investigated feature; both for considering the similarity between words or between item components. The latter includes the semantic relationship between item stem and distractors, distractors and correct responses, or between distractors. The intuition behind using such features is that highly semantically-related components increase the cognitive load on examinees when choosing the correct answer, resulting in an increase in difficulty level. For example, in gap filling items the semantic relatedness between the gap and its context is a significant factor which affects difficulty [4].

**Table 2.** Common Semantic-based Difficulty Features

| **Semantic Difficulty Feature** | **Studies** |
|---|---|
| Semantic similarity between words | [28] |
| Semantic similarity between options | [1] [20] [27] |
| Semantic similarity between item stem and options | [20] [38] [43] |
| Semantic similarity between context (i.e. learning material or passage) and item elements (stem, options and answer) | [3] [38] [49] |

It is worth mentioning that recent publications have utilised ontology-based measures to measure semantic similarity between items' components [26, 27, 44–46]. Ontologies have been increasingly utilised because they provide means to describe semantic relations of domain knowledge in a formal, structured and machine-processable format. Therefore, several ontology-based metrics have been developed in the literature by considering the relationship between concepts, predicates and individuals in the ontology. For example, word popularity on a semantic level can be determined by counting the number of object properties which are linked to an individual from other individuals [46].

## 5 Conclusion

In this paper, we have provided a systematic literature review on data-driven item difficulty prediction, and presented a classification which distinguishes between cognitive and systematic approaches to item difficulty prediction. The review establishes the data-driven approaches as a recent trend, that has emerged to overcome limitation of previous methods. The majority of the reviewed papers were domain- and item-specific. Furthermore research also suggests that linguistic features play a major role in determining items' difficulty level.

The reviewed papers failed to identify specific data-driven approaches that are able to provide generic frameworks that can be applicable across multiple domains and item types. This would have served as a first step towards providing automatic, reliable and objective evaluation methods to automatically validate items with regard to difficulty. This is the objective of our future research.

## References

1. Alsubait, T., Parsia, B., Sattler, U.: Generating multiple choice questions from ontologies: Lessons learnt. In: Keet, C.M., Tamma, V.A.M. (eds.) Proc. 11th Int. Workshop on OWL: Experiences and Directions (OWLED 2014). CEUR Workshop Proceedings, vol. 1265, pp. 73–84 (2014)
2. Aryadoust, V.: Predicting item difficulty in a language test with an adaptive neuro fuzzy inference system. In: 2013 IEEE Workshop on Hybrid Intelligent Models and Applications (HIMA). pp. 43–50. IEEE (2013)
3. Beinborn, L., Zesch, T., Gurevych, I.: Predicting the difficulty of language proficiency tests. Trans. Association for Computational Linguistics **2**, 517–530 (2014)
4. Beinborn, L., Zesch, T., Gurevych, I.: 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. pp. 1–11 (2015)
5. Benedetto, L., Cappelli, A., Turrin, R., Cremonesi, P.: Introducing a framework to assess newly created questions with natural language processing. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) Artificial Intelligence in Education. pp. 43–54. Springer (2020)
6. Benedetto, L., Cappelli, A., Turrin, R., Cremonesi, P.: R2de: a NLP approach to estimating IRT parameters of newly generated questions. In: Proc. of the 10th Int. Conf. on Learning Analytics & Knowledge. pp. 412–421 (2020b)
7. Bilotti, M.W., Ogilvie, P., Callan, J., Nyberg, E.: Structured retrieval for question answering. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 351–358 (2007)
8. Boldt, R.F.: GRE analytical reasoning item statistics prediction study. ETS Research Report Series **1998**(2), i–23 (1998)
9. Boldt, R.F., Freedle, R.: Using a neural net to predict item difficulty. ETS Research Report Series **1996**(2), i–19 (1996)
10. Choi, I.C., Moon, Y.: Predicting the difficulty of efl tests based on corpus linguistic features and expert judgment. Language Assessment Quarterly **17**(1), 18–42 (2020)
11. Crisp, V., Grayson, R.: Modelling question difficulty in an a level physics examination. Research Papers in Education **28**(3), 346–372 (2013)
12. Fei, T., Heng, W.J., Toh, K.C., Qi, T.: Question classification for e-learning by artificial neural network. In: Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint. vol. 3, pp. 1757–1761. IEEE (2003)
13. Franzen, M.: Item difficulty. Encyclopedia of Clinical Neuropsychology pp. 100–100 (2011)
14. Gao, Y., Bing, L., Chen, W., Lyu, M.R., King, I.: Difficulty controllable generation of reading comprehension questions. arXiv preprint arXiv:1807.03586 (2018)
15. Grivokostopoulou, F., Hatzilygeroudis, I., Perikos, I.: Teaching assistance and automatic difficulty estimation in converting first order logic to clause form. Artificial Intelligence Review **42**(3), 347–367 (2014)
16. Grivokostopoulou, F., Perikos, I., Hatzilygeroudis, I.: Estimating the difficulty of exercises on search algorithms using a neuro-fuzzy approach. In: 2015 IEEE 27th Int. Conf. on Tools with artificial intelligence (ICTAI). pp. 866–872. IEEE (2015)
17. Grivokostopoulou, F., Perikos, I., Hatzilygeroudis, I.: Difficulty estimation of exercises on tree-based search algorithms using neuro-fuzzy and neuro-symbolic approaches. In: Advances in combining intelligent methods, pp. 75–91. Springer (2017)

18. Ha, V., Baldwin, P., Mee, J., et al.: Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In: Proc. of the 14th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 11–20 (2019)
19. Hoshino, A., Nakagawa, H.: Predicting the difficulty of multiple-choice close questions for computer-adaptive testing. Natural Language Processing and its Applications p. 279 (2010)
20. Hsu, F.Y., Lee, H.M., Chang, T.H., Sung, Y.T.: Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. Information Processing & Management **54**(6), 969–984 (2018)
21. Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., Hu, G.: Question difficulty prediction for reading problems in standard tests. In: AAAI. pp. 1352–1359 (2017)
22. Hutzler, D., David, E., Avigal, M., Azoulay, R.: Learning methods for rating the difficulty of reading comprehension questions. In: 2014 IEEE International Conference on Software Science, Technology and Engineering. pp. 54–62. IEEE (2014)
23. Khodeir, N.A., Elazhary, H., Wanas, N.: Generating story problems via controlled parameters in a web-based intelligent tutoring system. The International Journal of Information and Learning Technology **35**(3), 199–216 (2018)
24. Khoshdel, F., Baghaei, P., Bemani, M.: Investigating factors of difficulty in c-tests: A construct identification approach. International Journal of Language Testing **6**(2), 113–122 (2016)
25. Kitchenham, B.A., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Tech. Rep. EBSE 2007-001, Keele University and Durham University Joint Report (07 2007)
26. Kurdi, G., Leo, J., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Donato, G., Dowling, W.: A Comparative Study of Methods for a Priori Prediction of MCQ Difficulty. Semantic Web – Interoperability, Usability, Applicability (2020)
27. Lin, C., Liu, D., Pang, W., Apeh, E.: Automatically predicting quiz difficulty level using similarity measures. In: Proceedings of the 8th International Conference on Knowledge Capture. pp. 1–8 (2015)
28. Lin, L.H., Chang, T.H., Hsu, F.Y.: Automated prediction of item difficulty in reading comprehension using long short-term memory. In: 2019 International Conference on Asian Language Processing (IALP). pp. 132–135. IEEE (2019)
29. Loukina, A., Yoon, S.Y., Sakano, J., Wei, Y., Sheehan, K.: 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. pp. 3245–3253 (2016)
30. Mitra, N., Nagaraja, H., Ponnudurai, G., Judson, J.: The levels of difficulty and discrimination indices in type a multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. International e-Journal of Science, Medicine & Education (IeJSME) **3**(1), 2–7 (2009)
31. Narayanan, S., Kommuri, V.S., Subramanian, N.S., Bijlani, K., Nair, N.C.: Unsupervised learning of question difficulty levels using assessment responses. In: Int. Conf. on Computational Science and Its Applications. pp. 543–552. Springer (2017)
32. Ozuru, Y., Rowe, M., O'Reilly, T., McNamara, D.S.: Where's the difficulty in standardized reading tests: The passage or the question? Behavior Research Methods **40**(4), 1001–1015 (2008)
33. Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R.D., Brefeld, U.: Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. International Journal of Artificial Intelligence in Education **29**(3), 342–367 (2019)

34. Parry, J.R.: Ensuring fairness in difficulty and content among parallel assessments generated from a test-item database. Online Submission (2020). https://doi.org/http://dx.doi.org/10.13140/RG.2.2.32537.03689
35. Perikos, I., Grivokostopoulou, F., Hatzilygeroudis, I., Kovas, K.: Difficulty estimator for converting natural language into first order logic. In: Intelligent Decision Technologies, pp. 135–144. Springer (2011)
36. Perikos, I., Grivokostopoulou, F., Kovas, K., Hatzilygeroudis, I.: Automatic estimation of exercises' difficulty levels in a tutoring system for teaching the conversion of natural language into first-order logic. Expert Systems **33**(6), 569–580 (2016)
37. Perkins, K., Gupta, L., Tammana, R.: Predicting item difficulty in a reading comprehension test with an artificial neural network. Language testing **12**(1), 34–53 (1995)
38. Qiu, Z., Wu, X., Fan, W.: Question difficulty prediction for multiple choice problems in medical exams. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 139–148 (2019)
39. Rust, J., Golombok, S.: Modern psychometrics: The science of psychological assessment. Routledge (2014)
40. Sano, M.: Automated capturing of psycho-linguistic features in reading assessment text. In: annual meeting of the National Council on Measurement in Education, Chicago, IL (2015)
41. Seyler, D., Yahya, M., Berberich, K.: Knowledge questions from knowledge graphs. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. pp. 11–18 (2017)
42. Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., Upmeier zu Belzen, A.: Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. Assessment & Evaluation in Higher Education **41**(5), 721–732 (2016)
43. Susanti, Y., Tokunaga, T., Nishikawa, H., Obari, H.: Controlling item difficulty for automatic vocabulary question generation. Research and practice in technology enhanced learning **12**(1), 1–16 (2017)
44. Vinu, E.V., Alsubait, T., Sreenivasa Kumar, P.: Modeling of item-difficulty for ontology-based mcqs. CoRR **abs/1607.00869** (2016)
45. Vinu, E.V., Sreenivasa Kumar, P.: A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. Journal of Web Semantics **34**, 40–54 (2015)
46. Vinu, E.V., Sreenivasa Kumar, P.: Automated generation of assessment tests from domain ontologies. Semantic Web **8**(6), 1023–1047 (2017)
47. Vinu, E.V., Sreenivasa Kumar, P.: Difficulty-level modeling of ontology-based factual questions. arXiv preprint arXiv:1709.00670 (2017)
48. Xue, K., Yaneva, V., Runyon, C., Baldwin, P.: 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. pp. 193–197 (2020)
49. Yeung, C.Y., Lee, J.S., Tsou, B.K.: Difficulty-aware distractor generation for gap-fill items. In: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association. pp. 159–164 (2019)
50. Zhou, Y., Zhang, H., Huang, X., Yang, S., Babar, M.A., Tang, H.: Quality assessment of systematic reviews in software engineering: A tertiary study. In: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. pp. 1–14 (2015)