# A Framework for Assessing the Complexity of Auto Generated Questions from Ontologies

**Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma**
University of Liverpool, Liverpool, United Kingdom

S.Alkhuzaey@liverpool.ac.uk
F.Grasso@liverpool.ac.uk
T.R.Payne@liverpool.ac.uk
V.Tamma@liverpool.ac.uk

**Abstract**: Automatic difficulty calibration (ADC) is the application of computational techniques to estimate the difficulty levels of assessment questions before administering them. Compared to traditional difficulty calibration approaches, ADC eliminates the need for pretesting, and minimises the time and efforts typically involved in manual calibration. In recent years, Ontology-based Automatic Question Generation (OAQG) has emerged as a powerful tool to generate assessment questions effortlessly and in massive numbers with minimal human intervention. Despite these benefits, not being able to control the characteristics of generated questions hinders their suitability to be used in pedagogical settings. However, much of the research up to now has tended to focus on the quantity of the questions rather than their quality. Indeed, most current automatic question generators produce simple questions which consist of a few facts, and simply test the recall of knowledge. Furthermore, the majority of existing frameworks are mostly technical and are not supported with a strong theoretical underpinning. In this paper, we propose a novel framework to assess the complexity of ontology-based, automatically generated questions. We discuss various aspects that are involved in determining the complexity of assessment questions, and attempt to quantify important characteristics of question complexity through the use of novel ontological metrics. To further support the plausibility of our computational framework, we shed light on its consistency with theories from education and cognitive psychology. This will provide us with a solid theoretical foundation which ensures that questions are generated according to principled methods that are grounded in theories of learning and cognition. The proposed approach is agnostic to different domains and independent of the question format, therefore, is highly general and applicable to a variety of contexts.

**Keywords:** Question Generation, Difficulty Estimation, Ontologies, Education, Question Complexity, Assessment.

## 1. Introduction

Automatic Question Generation (AQG) is a popular and challenging research area where technology-enabled solutions are used to automate assessment. It provides a scalable solution that can generate massive numbers of questions in a short time and with minimal human intervention. Various approaches have been proposed to generate questions from unstructured sources (e.g. text) with the help of Natural Language Processing (NLP) (Faizan and Lohman, 2018; Gao et al., 2018; Huang et al., 2017 ), or from structured sources such as ontologies and knowledge bases (Alsubait et al., 2016; Seyler et al., 2017; Vinu and Kumar, 2017). A small number of studies have explored the problem of controlling the characteristics of the generated questions to ensure their appropriateness for students' assessment. Some of them have also addressed the essential issue of controlling the level of difficulty of the generated questions. However, a desirable characteristic is the ability to differentiate between weak and competent students, through the design of questions with differing difficulty. This becomes even more important when moving toward the automatic generation of high quality and well-balanced exams. However, and despite the extensive literature on AQG, few studies have dealt with the task of difficulty estimation by developing computational frameworks to automate the process (Kurdi et al., 2020).

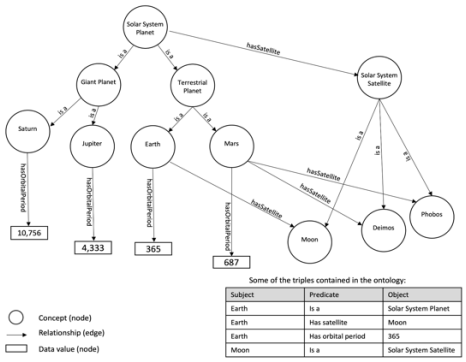



**Figure 1: An example of an ontology describing some aspects of the solar system**

Ontology-based Automatic Question Generation, OAQG, (Papasalouros et al., 2008) exploits domain specific knowledge bases (*ontologies*) as the main source of knowledge upon which question generation can be based. An ontology is a formal, explicit specification of a shared conceptualisation (Studer et al., 1998) that provides a representation vocabulary that can be used to describe the subject domain of interest in the form of a hierarchy or taxonomy of concepts that are connected to each other using different types of relationships. The taxonomic structure is typically modelled as a hierarchical classification, with more general concepts appearing higher within the taxonomy, and more specific concepts (connected by an is-a relationship) appearing lower down. In Figure 1, a small ontology fragment describes some of the concepts in the solar system; the concept "*Solar System Planet*" has two sub-concepts representing terrestrial and giant planets, whereas the concept "*Solar System Satellite*" has the sub-concepts (*Moon*, *Demios* and *Phobos*). These hierarchical relationships are represented by an "*is-a*" predicate. In addition, we have two other relationships: *hasOrbitalPeriod* and *hasSatellite*, which describe properties for the planets in our vocabulary. Thus, the triples (*Earth, hasOrbitalPeriod, 365*) and (*Earth, hasSatellite, Moon*) provide a knowledge-based description of the concept *Earth*.

Ontologies have a number of advantages over other forms of knowledge representation when generating questions with a controlled difficulty level. Ontologies have an inherent structure which can be complex (typically represented by a formally defined schema) and machine readable, making them more amenable to navigation and querying than unstructured forms of knowledge, such as natural language text. Not only does this eliminate the need to pre-process the input data, but also that existing ontologies can be readily reused, expanded and combined (Papasalouros et al., 2008). Furthermore, additional knowledge can be inferred from the ontology using rules or logics such as Description Logics (DLs) (Baader et al., 2017) which allow the specification of new knowledge that can be derived from the subject domain. From a cognitive perspective, the hierarchical architecture typically found within ontologies, coupled with their inferential techniques, resemble the way knowledge is represented in students' memory as well as reflecting the way it is accessed. Indeed, research in the cognitive sciences has previously demonstrated that learners tend to represent subject knowledge in the form of a hierarchical structure where nodes represent units of information (concepts) that are connected to each other (Siew, 2020; Gobbo and Chi, 1986; Kalyuga, 2007). Furthermore, these studies have suggested a correlation between complex hierarchical knowledge structures and the actual time learners use to retrieve concepts from semantic memory to answer a question (Collins and Quillian, 1969). For example, Chang, (1986) argues that learners tend to respond faster when asked to verify if robins are birds compared to when asked if robins are animals. This is because the mental process of verifying that robins are animals would be consistent with having to traverse through an additional level on the hierarchy (i.e. robins are birds). This is exactly the case when considering ontology-generated questions, as we aim to capture the characteristics of knowledge that are tested by a question that was generated based on a hierarchical structure. In addition to all these advantages, semantic knowledge models have been found to be an effective tool that can produce *general* difficulty estimation frameworks. Indeed, ontology-based approaches were praised for being generalisable across different domains and questions formats (Alkhuzaey et al., 2023). All these reasons support the argument that there is real value in using ontologies to generate difficulty-controllable educational questions.

In this paper, we propose a novel framework for assessing the complexity of automatically generated questions from domain ontologies. We underpin our framework by grounding it on pedagogical and psychological theories of learning and cognition. Before introducing our framework, we focus on discussing the fundamental issue of what makes a good question.

## 2. Assessment Quality: Background

Generally speaking, educational measurement refers to the process of gathering reliable information on the knowledge and skills of learners and assigning it a numerical value (Berry and Adamson, 2011). These properties are not necessarily directly observable, but can be hypothetical attributes that are complex, abstract, and/or not consistent across different learners. The fundamental purpose of assessment questions is to be used as a diagnostic tool to provide ongoing feedback to a learner on their understanding of the subject domain, and/or to measure the extent to which learners have met the learning objectives. A good quality assessment question should allow teachers and test developers to distinguish between high and low mastery students (i.e. weak vs. competent students). This is based on the premise that the characteristics of knowledge acquired by different level learners will differ with regard to its quantity and quality (Gobbo and Chi, 1986; Chi et al., 1981). Indeed, the characteristics of learning and performance differ significantly as knowledge increases (i.e. when there is the development of expertise) (Kalyuga, 2007). Thus, knowledge is considered to be an important property that, if

measured, can differentiate these two groups. However, knowledge is a property that is notoriously challenging to measure precisely. Any quantification of this feature must be an approximation that can vary based on the learner, test developer or the measurement method used. The definition of knowledge varies in the literature according to the research field considered. However, and despite the varying contexts, it is important to observe that these definitions often offer a simplification of what we mean by knowledge. This simplification is necessary, especially in pedagogical contexts, in order to facilitate the quantification of the term. One such simplification defines knowledge as a network structure of *nodes* and *edges* where nodes represent units of information (i.e. concepts) relevant to a specific domain, which are connected by edges that represent relations between concepts (Siew, 2020). We find that this definition of knowledge can be mapped to the way ontologies are structured.

An important criterion that is usually considered when designing high quality questions is *difficulty calibration*, which refers to the process of assigning a numerical or categorical score to represent the difficulty level of a question (Pelánek et al., 2021). This metric is crucial for test developers, as it allows them to design well-balanced tests with varying levels of difficulty in order to differentiate between different levels of achievement. Difficulty is a multi-faceted property that is affected by external and internal factors. External factors are influenced by variables such as students' background knowledge or the difficulty of the subject domain under consideration. Conversely, internal factors are derived from the intrinsic structure of the question itself and are commonly referred to as the *complexity* of the question. Hence, and despite being sometimes used as synonyms, difficulty and complexity are two distinct concepts (Pelánek et al., 2021). In this sense, complexity is one aspect that can indicate question difficulty. Indeed, among the ideal sources of difficulty are: 1) the underlying knowledge; and 2) the cognitive ability required to generate a correct answer (NRC, 2001). A question that entails more knowledge and requires cognitive processing is hence considered difficult. The latter source of difficulty is concerned with how learners understand and decompose a question before answering it correctly (also referred to as *the question-answering process*) (Ahmed and Pollitt, 1999; Hebel et al., 2017). If presented with a complex question, expert learners will be more able to decompose the question, identify relations and perform reasoning to arrive at the correct answer, whereas, given the limited abilities of low-knowledge learners, they will find the question more challenging. Based on this understanding of considering complexity as an *intrinsic* property of a question that is given by its internal structure, we use complexity as a proxy to calculate difficulty, meaning that complex questions (with complex internal structures) are also considered difficult. This assumption is supported by previous work which found that, when evaluating human-authored or automatically generated questions, estimating difficulty based on complexity is effective (Alkhuzaey et al., 2023; Choi and Moon, 2020; Yaneva et al., 2019; Loukina et al., 2016).

## 3. Theoretical Foundation

In this section, we provide an overview of the pedagogical and psychological literature on which we ground our approach.

### 3.1 Question Complexity

Different points of view exist in the pedagogical literature regarding what makes a question complex. Complexity has been interpreted as a quantification of the cognitive demand that the question poses on the learner. In their investigation into what makes a question cognitively demanding, Ahmed and Pollitt (1999) proposed that question complexity was an important dimension that increases cognitive demands. According to their view, complexity was defined as the number of operations that must be carried out before generating a correct answer, or the number of ideas that must be brought together (as well as the nature of the relationships between these ideas) in order to determine a solution. Thus, low complexity questions involve simple ideas and operations without the need to link them together; whereas higher complexity questions would require the learner to identify and combine different operations and ideas, and then link them together. When considered on a cognitive scale, a complex question would require the learner to recall and evaluate their subject knowledge, whereas simple questions only require the recognition of knowledge (Ahmed and Pollitt, 1999). Complexity in this definition has been successfully used to rate exam questions in terms of their cognitive demand, in subjects such as Geography, History and Chemistry (Hughes et al., 1998).

This understanding of complexity is also shared by researchers from fields which share commonalities with the AQG community. For example, the question-answering community distinguishes between *simple* and *complex* questions in terms of the number of facts they include, and the need to infer inner relationships (Lan et al., 2021;

Etezadi and Shamsfard, 2023; Luo et al., 2018). Questions that have multiple facts and constraints, express compound relations, and/or include numerical operations, are considered complex questions that require the learner to infer intrinsic semantic components (i.e. understand the inner relationships). To illustrate, consider the question "*Which terrestrial planet has the second highest orbital period?*". For a learner to answer this question correctly, they need to infer several semantic components: 1) that the answer is a member of the *terrestrial planet class*; 2) the answer has a numerical attribute presenting the orbital period value; and 3) the learner must choose the answer that is ranked as the second highest orbital period value amongst all *terrestrial* planets. Meanwhile, a question such as "*What is the orbital period of Mars?*" only requires the simple recall of one fact (i.e. The orbital period of Mars is 687 days).

### 3.2 The Gradual Progression towards Competence

According to theories from cognitive psychology, expert and novice learners differ considerably regarding the quantity and quality of knowledge they possess about a certain subject (Gobbo and Chi, 1986; Kalyuga, 2007; Chi et al., 1981; Chi and Koeske, 1983). Expert learners tend to be superior to novices in two knowledge related aspects: 1) they have a more integrated and cohesive knowledge structure of the subject; and 2) because their knowledge is well-organised, it allows them to perform higher cognitive tasks such as reasoning and deduction. In this context, knowledge cohesion is defined as the pattern of interrelations among the subject's concepts through direct or indirect links. The latter aspect was found to be affected by the former one, meaning that the more complex the knowledge structure the students have, the more capable they will be in performing complex cognitive tasks. Studies have shown that more proficient learners are more capable of understanding complex structures of physics problems compared to novice learners who focus on surface structures (Chi et al., 1981). This unique distinction is consistent with the pedagogical point of view that considers a good quality question to be one that must account for these expert-novice differences. Furthermore, the National Research Council (NRC) (2001) report highlighted the need to develop assessment instruments that are informed by cognitive models of progression towards competence. The rationale for this recommendation is mostly based on research on cognitive expert-novice differences.

These two aspects of proficiency, degree of structure and the ability to perform reasoning, can be mapped to how well-connected concepts are in the ontology, and on the choice of logical formalism used to represent the ontology, which in turn determines the type of reasoning, and therefore the new knowledge that can be derived. To illustrate, refer to the example in Figure 2 which shows two questions generated from the example ontology in Figure 1. Figure 2 assumes that the example ontology in Figure 1 is expressed using the Resource Description Framework – RDF (W3C, 2014), which is a knowledge representation formalism for representing knowledge bases in a machine-readable format. An RDF document is composed of statements about concepts and their relationships. These statements are represented by triples in the form *(subject, predicate, object)* where a concept (the *subject*) has some property (the *predicate*) expressing the type of relation that exists with another concept (the *object*). A statement represents a portion of a graph where the subject and the object are vertices, and the predicate is the edge representing the relationship between them. RDF knowledge bases are queried through the SPARQL query language (W3C, 2013), that exploits the RDF graph data model and formulates queries ranging from simple graph pattern matching (in the form of triple patterns) to complex queries that can span across different knowledge bases.

We can see clear differences in the structure of these two questions when we consider the degree of complexity of the questions' graph patterns. Question A is based on a simple triple pattern, that searches for the objects of triples that have :Earth as a subject and :hasSatellite as a predicate. Question B looks for entities that are of type :Planet and for the value for their property :hasOrbitalPeriod. SPARQL also allows the specification of solution modifiers (e.g ORDER BY, that orders results) and filters (conditions used to restrict solutions to those for which the condition evaluates to TRUE)., which are used in Question B to return only the largest value of :hasOrbitalPeriod for each entity returned.

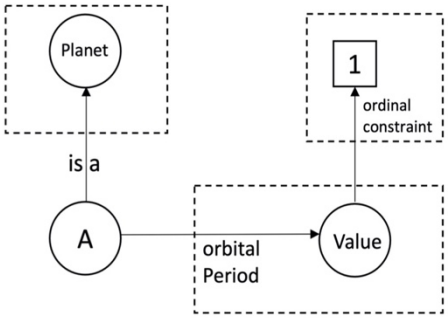
| | A | B |
|---|---|---|
| **Question** | What is the satellite of Earth? | Which planet has the longest orbital period? |
| **Graph Pattern** |  |  |
| **SPARQL Query** | `SELEC ?object`<br>`WHERE {`<br>`  :Earth :hasSatellite ?object.`<br>`}` | `SELEC ?subject`<br>`WHERE {`<br>`  ?subject rdf:type :Planet.`<br>`  ?subject :hasOrbitalPeriod :object.`<br>`}`<br>`ORDER BY DESC(?object)LIMIT 1` |

**Figure 2: Two examples of ontology-based auto-generated questions from ontologies (*A = answer*).**

## 4. The Proposed Complexity Framework

The main goal of our framework is to be able to measure question complexity with ontological metrics to distinguish between students of different mastery levels. This assumes that questions are based on a knowledge base whose elements are expressed using the vocabulary and structure modelled by an ontology. Given that the gradual progression from novice to expert involves significant changes in *what* and, more importantly, *how* knowledge is represented in students' minds, we argue that these aspects are consistent with how ontologies are structured and accessed. This similarity in structure has motivated the development of our framework. In this paper, we adopt the understanding of complexity, where the cognitive demands are represented by the number of operations that must be carried out during the question-answering process, i.e. we propose to quantify these operations using ontological metrics. Therefore, we argue that in order to control the complexity of generated questions, we must be able to calculate an estimate for the knowledge structure required to answer the question (i.e. the complexity of the graph pattern) and the level of reasoning required during the question-answering process.

We introduce a metric that correlates the complexity of an ontology-generated question to the complexity of its underlying graph pattern. We define two categories of questions that can be generated from ontologies: *simple* and *complex* questions. Simple questions are questions which only require the recall of a single fact, whereas complex questions are those which require the integration of multiple facts and constraints, and that are subject to some level of reasoning (e.g. numerical operations). These two types of questions differ with respect to the volume of knowledge they test and the cognitive demand they require. The distinction between these two questions is consistent with how expert learners differ from novices with regard to their subject knowledge. Expert learners who can work with complex knowledge structures and in turn perform higher cognitive tasks are more capable of correctly answering complex questions. In contrast, novices' low ability to work with complex structures can compromise their ability to answer complex questions. This definition is useful and practical as it can be operationalised due to the structured nature of ontologies, which allows us to precisely and explicitly measure metrics such as the complexity of graph patterns.

We compute question complexity by considering the graph pattern which models the question. We define that a complex graph pattern is one that: 1) contains multiple facts; and 2) requires some level of reasoning. The first criterion can be measured by counting the number of triples that the question covers. For the second criterion, the presence of any semantic constraints that are expressed with restrictions such as filters or modifiers which restrict the number of possible answers can indicate a question which needs further reasoning. In contrast, simple questions have fewer triple patterns and only require the recall of the fact. Based on this assumption, we present the following definition to assess the complexity of ontology-based questions.

**Definition**. We define the complexity of a question *Q* as the level of complexity of its corresponding Graph Pattern *GP*. The complexity of a Graph Pattern *C_GP* is defined as the total number of *TP*, *SM* and *F* it contains, where a triple pattern $TP_i$ is triple in which each element (the subject, predicate and object) can be a variable; a solution modifier, $SM_i$ modifies the result set for a query, and more than one modifier can be specified in a query; and a filter $F_i$ is a condition that restricts the results returned.

The definition illustrates that the increase in the number of triple patterns and the constraints posed on the answer set results in increasing the complexity of the generated question. To illustrate, consider our running example of a complex question in Figure 2. Question B is a complex question that entails the following triples ":Answer is_a :planet" and ":Answer :hasOrbitalPeriod :value". When considering the SPARQL query for this question, we also notice that an additional ordinal constraint was carried out by using the ORDER BY and LIMIT modifiers to reason about the ontology and only return the planet with the highest orbital period. For a student to answer question B correctly, they must have knowledge of 1) all the planets in the solar system, and 2) the orbital period of each planet. In addition, the student needs to perform a ranking task to choose the highest planet in the ranked solar system planet set based on the length of its orbital period. This is to say that a student with a higher knowledge level and good reasoning abilities will likely be able to answer both questions (simple and complex). Consequently, this complexity framework will enable us to discriminate between students of different mastery levels.

## 5. Related Work

Rich semantic models such as ontologies and knowledge bases have been previously used to aid in the difficulty calibration of ontology-based questions. In this section, we discuss some of the previously proposed difficulty estimation frameworks.

Although an important deal of research effort has been made in the field of ontology-based automatic question generation, early studies did not directly tackle the problem of controlling the difficulty of the generated questions. In early studies such as Fischer (2001), Holohan et al. (2005) and Papasalouros et al. (2008), the aspect of difficulty estimation was only implicitly tackled by suggesting that semantic similarity (i.e. the distance between concepts in the hierarchical structure of the ontology) increases the learner's effort to provide an answer. One of the earliest attempts to control the difficulty of Multiple-Choice Questions (MCQs) using ontological features was the work of Alsubait et al., (2013). They proposed a similarity-based theory to measure the *Concept similarity* between the key and distractors. They argued that semantically similar sets of options increase the cognitive demand on learners when trying to choose the correct answer. The underlying assumption was that identifying the correct answer would be more difficult if the distractors had a high semantic similarity to the answer. However, this approach is only applicable to MCQs, and the similarity-based measures only focus on distractors while neglecting the stem. This is problematic since it was found that the stem forms an important part of the question that can be a crucial deciding factor for its level of difficulty (Alkhuzaey et al., 2023). Seyler et al. (2017) proposed a difficulty estimation framework as a part of a question generator that employed Knowledge Graphs (KGs) such as DBpedia as an input source. Difficulty sources are extracted based on statistics computed from the knowledge graph, namely *entity popularity* and entity coherence (i.e. *the tendency of two entities to appear together)* as their difficulty metrics. However, given that their difficulty features are extracted from public KGs, this makes their approach less employable in specific educational contexts which utilise other types of knowledge sources. Vinu and Kumar (2015; 2016; 2017; 2020) have published several studies on ontology based AQG with various ontological metrics to control the difficulty of questions. Amongst the metrics they proposed were entities' *popularity, selectivity (*i.e. the number of individuals that satisfy it*), coherence* (i.e. semantic relatedness of entities) and *specificity* (i.e. the depth of the entity within the ontology hierarchy). Faizan and Lohman also exploited the ontological class hierarchy to predict difficulty based on the depth of the correct answer in the class hierarchy. The underlying assumption is that deeper concepts require more knowledge as compared to surface or general concepts. The last two approaches indicate that the hierarchical structure of ontologies can be effectively employed to control the characteristics of the generated questions.

Some limitations were observed in existing ontology-based difficulty estimation models. All the approaches (except for Alsubait et al., 2013) are technical and lack a theoretical underpinning which guides the generation process. We have also observed that most current approaches focus on simple graph pattern matching when accessing the content of the ontology while dismissing advanced expressions that allow us to reason over the knowledge described by the ontology. These approaches start with a fixed linguistic construct that is instantiated with triples (or one part of a triple such as an object) from the input ontology. This method does not account for advanced constructs which capture the semantic relationship between the concepts. This limitation contributes to the simplicity of structure which was observed when comparing automatically generated questions to human-authored questions. Leo et al. (2019) have indeed noted that most of the current approaches are restricted in terms of their linguistic form. Another limitation of many of the current approaches is that the majority of questions generated are simple and only test the recall of knowledge (Khodeir et al, 2014). This has also been highlighted and addressed by Leo et al. (2019) who proposed an approach to control the difficulty of complex case-based MCQs in the medical domain. Although their model generates questions that test higher cognitive abilities (i.e. analysing knowledge), it is both domain- and source-specific as they only address medical case-based MCQs that are generated from one specific dataset, namely The Elsevier Merged Medical Taxonomy (EMMeT).

In this study, we focused on addressing the limitations of current approaches by proposing a generalisable framework which assesses the level of complexity based on ontological patterns to account for the underlying semantic structures. Our approach is employable in both generic and domain-specific contexts and is agnostic to the question format. Furthermore, it is grounded in principled theories of learning and cognition to increase its applicability in educational contexts.

## 6. Conclusion and Future Work

In this paper, we have proposed a framework to assess the complexity of ontology-based auto-generated questions. We have shown that our framework is practical, and consistent with well-established theories of learning and cognition. This work adds to the rapidly expanding field of AQG by providing researchers in the field with an evaluation framework that can help them to automatically calibrate the generated questions according to their complexity level. In future work, we aim to empirically validate our proposed approach by utilising it to practically assess automatically generated questions from domain ontologies.

## References


Ahmed, A. and Pollitt, A., (1999). Curriculum demands and question difficulty. In *IAEA conference*, Bled, Slovenia.

AlKhuzaey, S., Grasso, F., Payne, T., Tamma, V. (2023) Text-based Question Difficulty Prediction: A Systematic Review of Automatic Approaches. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00362-1

Alsubait, T., Parsia, B., Sattler, U. (2016). Ontology-based multiple choice question generation. *KI-KÅNunstliche Intelligenz* , 30 (2), 183–188.

Alsubait, T., Parsia, B. and Sattler, U., (2013). A similarity-based theory of controlling MCQ difficulty. *In 2013 second international conference on e-learning and e-technologies in education (ICEEE)*, 283-288. IEEE.

Berry, R. and Adamson, B. eds., (2011). *Assessment reform in education: policy and practice,* 14. *Springer Science & Business Media*.

Chang, T.M., (1986). Semantic memory: Facts and models. *Psychological Bulletin*, *99*(2), p.199.

Chi, M.T. and Koeske, R.D., 1983. Network representation of a child's dinosaur knowledge. *Developmental psychology*, 19(1), p.29.

Chi, M.T., Feltovich, P.J. and Glaser, R., (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2), 121-152.

Choi, I.-C., & Moon, Y. (2020). Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, 17 (1), 18–42.

Collins, A.M. and Quillian, M.R., (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2), 240-247.

Etezadi, R. and Shamsfard, M., (2023). The state of the art in open domain complex question answering: a survey. *Applied Intelligence*, 53(4), 4124-4144.

Faizan, A., & Lohmann, S. (2018). Automatic generation of multiple-choice questions from slide content using linked data. *Proceedings of the 8th international conference on web intelligence, mining and semantics*. 1–8.

Fischer, S. (2001), Course and exercise sequencing using metadata in adaptive hypermedia learning systems. *Journal of Educational Resources in Computing*, 1(1), p. 5

Gao, Y., Bing, L., Chen, W., Lyu, M.R., King, I. (2018). Difficulty controllable generation of reading comprehension questions. *arXiv preprint arXiv:1807.03586* .

Gobbo, C. and Chi, M., (1986). How knowledge is structured and used by expert and novice children. *Cognitive development*, 1(3), 221-237.

Holohan, E., Melia, M., McMullen, D. and Pahl, C. (2005), Adaptive E-learning content generation based on semantic web technology. *Proceedings of Workshop on Applications of Semantic Web Technologies for e-Learning,* Amsterdam, The Netherlands, 29-36.

Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., . . . Hu, G. (2017). Question difficulty prediction for reading problems in standard tests. *Aaai,* 1352–1359.

Hughes, S., Pollitt, A. and Ahmed, A., (1998). The development of a tool for gauging the demands of GCSE and A Level exam questions. BERA, Queen's University Belfast.

Khodeir, N., Wanas, N., Darwish, N. and Hegazy, N., (2014). Bayesian based adaptive question generation technique. *Journal of Electrical Systems and Information Technology*, *1*(1), 10-16.

Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S., (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30, 121-204.

Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W.X. and Wen, J.R., (2021). A survey on complex knowledge base question answering: Methods, challenges and solutions*. arXiv preprint arXiv:2105.11644*.

Le Hebel, F., Montpied, P., Tiberghien, A. and Fontanieu, V., (2017). Sources of difficulty in assessment: example of PISA science items. *International Journal of Science Education*, 39(4), 468-487.

Leo, J., Kurdi, G., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Dowling, W. (2019). Ontology-based generation of medical, multiterm mcqs. *International Journal of Artificial Intelligence in Education*, 29 (2), 145–188.

Loukina, A., Yoon, S.-Y., Sakano, J., Wei, Y., Sheehan, K. (2016). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. Proceedings of coling 2016, *the 26th international conference on computational linguistics: Technical papers*, 3245–3253.

Luo, K., Lin, F., Luo, X. and Zhu, K., (2018). Knowledge base question answering via encoding of complex query graphs. *In Proceedings of the 2018 conference on empirical methods in natural language processing*, 2185-2194.

National Research Council (NRC), (2001). Knowing what students know: The science and design of educational assessment. *National Academies Press*.

Papasalouros, A., Kanaris, K. and Kotis, K., (2008). Automatic Generation of Multiple-Choice Questions from Domain Ontologies. *e-Learning*, 1, 427-434.

Pelánek, R., Effenberger, T. and Čechák, J., (2022). Complexity and difficulty of items in learning systems. *International Journal of Artificial Intelligence in Education*, 32(1), 196-232.

Studer, R., Benjamins, V.R. and Fensel, D., (1998). Knowledge engineering: Principles and methods, *Data & Knowledge Engineering*, Volume 25, Issues 1–2, 1998, 161-197.

Seyler, D., Yahya, M., Berberich, K. (2017). Knowledge questions from knowledge graphs. *Proceedings of the acm sigir international conference on theory of information retrieval*, 11–18.

Siew, C.S., (2020). Applications of network science to education research: Quantifying knowledge and the development of expertise through network analysis. *Education Sciences*, 10(4), p.101.

Vinu, E., & Kumar, P. (2020). Difficulty-level modeling of ontology-based factual questions. *Semantic Web*, 11 (6), 1023–1036.

Vinu, E.V., & Kumar, P., (2017). Automated generation of assessment tests from domain ontologies. *Semantic Web*, 8 (6), 1023–1047.

Vinu, E., Alsubait, T., Kumar, P. (2016). Modeling of item-difficulty for ontology-based MCQs. arXiv preprint arXiv:1607.00869

Vinu, E.V., & Kumar, P. (2015). A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Journal of Web Semantics*, 34 , 40–54.

World Wide Web Consortium, (2014). RDF 1.1 concepts and abstract syntax. Available at: https://www.w3.org/TR/rdf12-concepts/#data-model (Accessed: 11 Jun 2023).

World Wide Web Consortium, (2013). SPARQL 1.1 Overview. Available at: https://www.w3.org/TR/sparql11-overview/#:~:text=SPARQL%201.1%20is%20a%20set,A%20query%20language%20for%20RDF. (Accessed: 11 Jun 2023).

Yaneva, V., Baldwin, P., Mee, J., et al. (2019). Predicting the difficulty of multiple-choice questions in a high-stakes medical exam. *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, 11–20.