Demaidi *et al. Smart Learning Environments* (2017) 4:7
DOI 10.1186/s40561-017-0046-6

Smart Learning Environments

RESEARCH Open Access

Evaluating the quality of the ontology-based auto-generated questions

Mona Nabil Demaidi[1*], Mohamed Medhat Gaber[1*] and Nick Filer[2*]

*Correspondence: monanabildemaidi@gmail.com; mohamed.gaber@bcu.ac.uk; nick@cs.man.ac.uk
[1]School of Computing and Digital Technology, Birmingham City University, Birmingham, UK
[2]School of Computer Science, University of Manchester, Manchester, UK

Abstract

An ontology is a knowledge representation structure which has been used in Virtual Learning Environments (VLEs) to describe educational courses by capturing the concepts and the relationships between them. Several ontology-based question generators used ontologies to auto-generate questions, which aimed to assess students' at different levels in Bloom's taxonomy. However, the evaluation of the questions was confined to measuring the qualitative satisfaction of domain experts and students. None of the question generators tested the questions on students and analysed the quality of the auto-generated questions by examining the question's difficulty, and the question's ability to discriminate between high ability and low ability students. The lack of quantitative analysis resulted in having no evidence on the quality of questions, and how the quality is affected by the ontology-based generation strategies, and the level of question in Bloom's taxonomy (determined by the question's stem templates). This paper presents an experiment carried out to address the drawbacks mentioned above by achieving two objectives. First, it assesses the auto-generated questions' difficulty, discrimination, and reliability using two statistical methods: Classical Test Theory (CTT) and Item Response Theory (IRT). Second, it studies the effect of the ontology-based generation strategies and the level of the questions in Bloom's taxonomy on the quality of the questions. This will provide guidance for developers and researchers working in the field of ontology-based question generators, and help building a prediction model using machine learning techniques.

Keywords: Ontology, Ontology-based question generator, Ontology-based generation strategy, Bloom's taxonomy, Classical Test Theory, Question difficulty index, Question discrimination index, Reliability, Discrimination power, Item Response Theory

Introduction

Ontology is a formal and explicit specification of a shared conceptualisation (Uschold and Gruninger 1996; Studer et al. 1998; Borst 1997). It is a knowledge representation structure, which models a specific domain of interest by providing a formal machine readable representation of entities in the domain. Entities include classes, individuals, and properties. Classes represent sets of individuals, individuals represent actual objects in the domain, and properties represent relationships in the domain between individuals.

Ontologies have been used in Virtual Learning Environments (VLEs) to capture the concepts in an educational course (Gruber 1993). Sakathi (Murugan et al. 2013) developed an ontology, which captures concepts in the computer networks domain such as the network topology, the communication's medium, and the Open Systems Inter-

connection (OSI) model. Lee et al. (2005), Kouneli et al. (Kouneli et al. 2012), and Ganapathi et al. (2017) developed ontologies, which capture the educational concepts in the Java language introductory courses (Arnold et al. 1996). The ontologies aimed to teach students the fundamental concepts of programming in Java.

On the other hand, some ontologies were not developed to capture particular domains. Instead, they aimed to have the world's largest and complete knowledge base that covers different domains. Among these ontologies is the OpenCyc ontology (OpenCyc). OpenCyc covers several domains such as; mathematics, physics, medicine, computer networks and many others, and it consists of hundreds of thousands of concepts and properties.

Ontologies have been used by several ontology-based question generators to auto-generate true and false, multiple choice, and short answer assessment questions. The question generators used several ontology-based generation strategies which exploit the ontology classes, individuals, and properties. The ontology-based generation strategies could be categorised into the following three main strategies (Papasalouros et al. 2017;2011, Cubric and Tosic 2017; Grubisic 2012; Grubisic et al. 2013; Al-Yahya 2014):

1. The class-based strategy, which uses the relationship between the ontology classes and individuals.
2. The terminology-based strategy, which uses the relationship between the class and sub-class in ontologies.
3. The property-based strategy, which uses the object, datatype, and annotation properties in the ontologies.

Papasalouros et al. (2017;2011) defined the class, terminology, and property-based generation strategies which traverse the domain ontology and auto-generate the multiple choice question's correct answer (key) and incorrect answers (distracters). The three main strategies contain several sub-strategies which specify the classes, individuals or properties from which the question's key and distractors are generated. For example, Table 1 illustrates a multiple choice question generated using Papasalouros's terminology-based strategy. The question was generated from Sakthi's computer network ontology. The question's key is a subclass of the concept OSI model and the question's distractors are sibling classes of the OSI model class. Table 1 also shows that the question had the "Choose the correct sentence" text, which is called the question's stem, and it is used in all the questions generated using Papasalouros's question generator.

Cubric and Tosic (2017) built a question generator which used the ontology-based generation strategies defined by Papasalouros. However, they extended the property-based strategies to include more sub-strategies, which used the annotation properties in the ontology. Moreover, instead of using Papasalouros's stem template, which is not

Table 1 Question generated using a terminology-based strategy

Stem:	Choose the correct sentence
Options:	(1) Transport layer is part of the OSI model (2) Network operating system is part of the OSI model (3) TCP/IP suite is part of the OSI model
Key:	(1) Transport layer is part of the OSI model

related to an educational theory, Cubric and Tosic defined a set of stem templates, which aimed to assess student cognition at different levels in the Bloom's taxonomy, which is widely used in the educational research (BS 1956; Krathwohl 2002; Anderson and Sosniak 1994). Bloom's taxonomy categorise the assessment questions into the following six major levels, which are arranged in a hierarchical order according to the complexity of the cognitive process involved (BS 1956; Krathwohl 2002; Assessment 2002): 1) Knowledge: at this level the students need only to recall certain concepts in the domain. For example students need to list, define, and describe specific concepts in the domain without understanding how they are related to other concepts. 2) Comprehension: at this level the students need to start thinking about the meaning of the concepts in terms of their relationship with other concepts in the domain. 3) Application: at this level the students need to demonstrate their ability to use the concepts they have learned in real situations. For example the students need to provide and show examples that prove their understanding of the domain concepts. 4) Analysis: at this level the students need to understand the domain terminology structure. For example the students need to have a good overview of the concepts in the domain by analysing how they are classified and related to each other. 5) Synthesis: at this level the students should be able to relate concepts from different domains to create and develop new ideas. 6) Evaluation: at this level the students need to make judgments, assess and compare ideas and evaluate the data.

Each level in Bloom's taxonomy is subsumed by the higher levels, for example a student functioning at the application level had mastered the educational concepts in the knowledge and comprehension levels (BS 1956). Bloom's associated the levels hierarchical order with the question's difficulty (BS 1956), for example knowledge level questions are easier than questions which assess other levels in Bloom's taxonomy, and synthesis and evaluation question are more difficult than the comprehension level question (BS 1956).

Cubric and Tosic (2017) generated questions which assess students at the knowledge, comprehension, application and analysis levels only. Grubisic (2012); Grubisic et al. (2013) followed a similar approach to Cubric and Tosic by defining a set of question stem templates which assess students' cognition at the knowledge, comprehension, application and analysis levels. However, unlike the previous work, Grubisic generated different types of questions (true and false, multiple choice, and short answer). Moreover, she ignored the class-based strategies, and only used the terminology-based and property-based strategies to traverse the ontology and generate assessment questions.

Grubisic (2012); Grubisic et al. (2013) used ontology-based generation strategies similar to Papasalouros. However, fewer restrictions were applied for selecting the distractors in the generated questions. For example, if a question is generated to assess students on the educational concept *EC*, Papasalouros defined that the distractor should be one of class *EC* siblings, while Grubisic allowed selecting any class randomly from the ontology as long as it has no relationship with *EC*.

Al-Yahya (2014;2011) also built a question generator for auto-generating true and false, multiple choice, and short answer questions using class-based and property-based strategies. She defined question stem templates aimed only to assess students' cognition at the knowledge level in Bloom's taxonomy (Al-Yahya 2014;2011). Al-Yahya followed Grubisic's steps in allowing distracters to be randomly selected from the domain ontology.

The ontology-based question generators discussed above evaluated the auto-generated questions. However, the evaluation of the questions was confined to measuring the qualitative satisfaction of domain experts and the students who agreed that the auto-generated questions could be used as assessment questions in learning environments. None of the ontology-based questions' generators tested the questions on students to analyse the quality of auto-generated questions by examining the question's difficulty, and the question's ability to discriminate between high ability and low ability students. In addition, the question generators auto-generated different types of questions using different ontology-based generation strategies. However, none of the ontology-based question generators studied the effect of the ontology-based generation strategies and the level of question in Bloom's taxonomy on the quality of questions generated. Therefore, this paper makes the following contributions to knowledge:

1. Developing an ontology-based question generator which integrates the stem templates and generation strategies introduced by Papasalouros et al. (2017;2011), Cubric and Tosic (2017), Grubisic (2012); Grubisic et al. (2013), and Al-Yahya (2014;2011). The generator could be used to generate questions from any domain ontology. In addition, it helps in evaluating the quality of questions quantitatively. This help researchers auto-generate questions with specific characteristics (e.g., high discrimination);
2. Quantitatively analyse the quality of ontology-based auto-generated question's for the first time;
3. quantitatively analyse the quality of assessment tests formed from the ontology-based auto-generated questions; and
4. study the effect of different ontology-based generation strategies and the level of question in Bloom's taxonomy on the quality of question's generated.

This paper is structured as follows: Section Related work illustrates the analysis used by existing question generators and presents the limitations in these analysis and the importance of our study. "Evaluation methods" explains the evaluation methods used in this paper to evaluate the quality of auto-generated questions. Section Experimental study presents the experimental study. Section Results and discussion illustrates the experiment results. Finally, "Conclusion and future work" concludes the paper and suggests future work.

Related work

Different qualitative and quantitative analyses were carried out to evaluate questions auto-generated from domain ontologies (Alsubait et al. 2014; Al-Yahya 2014; Vinu and Kumar 2017; Seyler et al. 2016; Susanti et al. 2017). Papasalouros et al. (2017;2011) auto-generated multiple choice questions (MCQs) from the Eupalineio Tunnel ontology, which is a domain ontology about the ancient Greek history. The questions were evaluated by two domain experts who found that all the questions were satisfactory for assessment regardless of some errors in the questions' syntax (75% of the MCQs were assessed as syntactically correct) (Papasalouros et al. 2017). Cubric and Tosic (2017) developed an online environment where users could upload their domain ontologies, auto-generate MCQs, and evaluate the questions created by them or other users in the environment. The users evaluate the auto-generated questions by determining the question quality (the question

is easy to understand and the grammar is correct), and the question usability (the question could be used in an assessment test). Cubric and Tosic did not publish any evaluation results.

Grubisic (2012); Grubisic et al. (2013) evaluated the questions auto-generated from the 'computer as system' domain ontology using two groups of students. The first group consisted of 14 students who had good prior knowledge in the 'computer as system' domain. However, the students had no experience working with VLEs. The second group consisted of 16 students who had learned about the 'computer as system' domain three years before the experimental study was carried out and had a good knowledge of different VLEs. 21% of the students in the first group found the questions comprehensible while 29% had a neutral opinion, and 50% found the questions incomprehensible (Grubisic et al. 2013). On the other hand, 38% of the students in the second group found the questions comprehensible, 38% had a neutral opinion, and 24% found the questions incomprehensible (Grubisic et al. 2013). Grubisic concluded that the students in the second group who were more mature (students who took the 'computer as system' course three years before the experiment was carried out) and who had more experience working with different VLEs were more satisfied in terms of understanding the ontology-based generated questions.

Al-Yahya (2014); Al-Yahya (2011) auto-generated true and false, multiple choice, and short answer questions from several domain ontologies such as the travel ontology, which captures information about travel destinations and hotels (Protege ontology library - protege wiki 2017). She evaluated the auto-generated questions by assessing if the questions are syntactically correct and whether the questions were suitable to be used in an assessment test. Al-Yahya's evaluation results revealed that 90% of the questions generated were syntactically correct and could be used as assessment questions (Al-Yahya 2011). Al-Yahya carried out further evaluation to assess if the auto-generated MCQs were syntactically correct and could be used as assessment questions using three domain experts. The experts had experience in formulating MCQs and were asked to assess the MCQs generated from two domain ontologies (an ontology which captures the Arabic vocabulary (Al-Yahya et al. 2010) and a history ontology in Arabic which captures the historical concepts taught to students in the 8^{th} grade (Al-Yahya 2014; 2011)). The experts agreed that 82% of the MCQs generated from the Arabic vocabulary were syntactically correct and could be used as assessment questions, while 60% of the MCQs generated from the history ontology were syntactically correct and could be used as assessment questions (Al-Yahya 2014). Al-Yahya stated that the difference in the evaluation results was due to the content of the domain ontologies, as the MCQs, which were classified as unacceptable in the history ontology, were dealing with common sense or general knowledge. This was not the case in the Arabic vocabulary ontology (Al-Yahya 2014).

In summary, the ontology-based question generators mentioned above have the following limitations: Firstly, the evaluation of the auto-generated questions was confined to measuring the qualitative satisfaction of domain experts and the students who agreed that the auto-generated questions could be used as assessment questions in learning environments. However, none of the ontology-based questions' generators tested the questions on students to analyse the quality of auto-generated questions by examining the question's difficulty, and the question's ability to discriminate between high ability and low ability students. Secondly, none of the ontology-based question generators studied the effect of the ontology-based generation strategies and the level of question in Bloom's taxonomy

on the quality of questions generated. Therefore, Section Evaluation methods presents the evaluation methods used in this paper to evaluate the questions quantitatively.

Evaluation methods

This section presents two statistical methods, which have been used to evaluate the quality of ontology-based generated questions.

Classical Test Theory

Classical Test Theory (CTT) is used to evaluate the quality of questions and assessment tests in learning environments using the statistical measures described in the following sections (Alagumalai and Curtis 2005; Ding and Beichner 2009; Doran 1980; Cohen et al. 2013; Erguven 2014).

Question difficulty index

The question's difficulty index (P) measures the question easiness and it is defined as the proportion of students choosing the correct answer (Ding and Beichner 2009; Doran 1980; Cohen et al. 2013; Schmidt and Embretson 2003):

$$P = \frac{N1}{N} \tag{1}$$

Where N1 is the number of correct answers and N is the total number of students taking the test. P values range from 0 to 1. Table 2 shows that questions with high difficulty indices are easy while questions with low difficulty indices are difficult.

Question discrimination index

The question's discrimination index measures how well the question could discriminate between high ability (students with high scores) and low ability students (students with low scores) (Ding and Beichner 2009; Doran 1980; Cohen et al. 2013). The discrimination index is defined as the difference between the proportion of the top quartile students who answered the question correctly and the proportion of the bottom quartile students who answered the question correctly (Ding and Beichner 2009; Doran 1980):

$$Discrimination\ index = \frac{N_H - N_L}{N/4} \tag{2}$$

Where N_H and N_L are the number of correct answers in the top quartile and bottom quartile, and N is the total number of students taking the test. Table 2 shows that

Table 2 Range of values and descriptions of the question's quality measurements

Question quality measurements	Range of values	Description	Literature
Difficulty index	[0, 0.35)	Very difficult	(Doran 1980)
	[0.35, 0.60)	Moderately difficult	(Doran 1980)
	[0.60, 0.85)	Moderately easy	(Doran 1980)
	[0.85, 1]	Very easy	(Doran 1980)
Discrimination index	[0, 0.3)	Low	(Doran 1980; Ebel 1979)
	[0.3, 0.6)	Medium	(Doran 1980; Ebel 1979)
	[0.6 , 1]	High	(Doran 1980; Ebel 1979)
Reliability (point biserial correlation coefficient)	[0.0, 0.3)	Low	(Dancey and Reidy 2004)
	[0.3, 6.0)	Medium	(Dancey and Reidy 2004)
	[0.6, 1]	High	(Dancey and Reidy 2004)

questions with discrimination indices <0.3 have low discrimination, while questions with discrimination indices ≥ 0.6 have high discrimination.

Question reliability

The question's reliability is measured using the point biserial correlation coefficient, which is the correlation between students scores in the question and students' total scores (Ding and Beichner 2009; Schmidt and Embretson 2003; Brown 1996):

$$R_{\text{pbi}} = \frac{(\bar{x_1} - \bar{x_0})}{\sigma_{\text{x}}} \sqrt{P_{\text{i}} * (1 - P_{\text{i}})} \tag{3}$$

Where R_{pbi} is the point biserial correlation coefficient for question i, $\bar{x_1}$ is the average total score of students who correctly answered question i, $\bar{x_0}$ is the average total score for students who did not answer question i correctly, σ_{x} is the standard deviation of students' total scores, and P_{i} is the difficulty index for question i. R_{pbi} value ranges from [-1, 1] and high R_{pbi} value means that students who selected the correct answer are students with high total scores and students who selected the incorrect answer are students with low total scores. Higher R_{pbi} values are better (Ding and Beichner 2009). The reliability is also used to measure the question's discrimination. Table 2 shows that questions with R_{pbi} <0.3 have low reliability (discrimination) while questions with $R_{\text{pbi}} \geq 0.6$ have strong reliability (discrimination).

Test discrimination power

The test discrimination power is measured using Ferguson's delta (δ) (Ferguson 1949), which investigates how broadly the test scores are distributed over the possible range of scores (Zhang and Lidbury 2013). Ferguson's delta (δ) is measured using the following formula:

$$\delta = \left(\frac{N^2 - \sum_{i=1}^{K} f_{\text{i}}}{N^2 - N^2/(K+1)} \right) \tag{4}$$

Where N is the total number of student who attempted the test, f_{i} is the number of students whose total score is i, k is the number of questions in a test. δ ranges from 0 to 1, where 0 indicates that the test has minimal discrimination and this occurs when all students have the same score. On the other hand, when δ is 1 this means all possible scores occur in the test with the same frequency (Hankins 2007). Ferguson's delta (δ) value greater than 0.9 is considered a good discrimination as it represent the normal distribution of scores (Kline 1986; 2013a; 2013b).

Test reliability

The test reliability is measured using Cronbach's α (Cronbach and Shavelson 2004), which measures the internal consistency of the test by finding the correlation between each question's score in the test and the whole test score. In other words, Cronbach's α examines whether a test is constructed from questions that address the same material and it is measured using the following formula:

$$\text{Cronbach } \alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^{n} P_{\text{i}}(1 - P_{\text{i}})}{{\sigma_{\text{x}}}^2} \right) \tag{5}$$

Where K is the number of questions in a test, P_{i} is the difficulty index of the i^{th} question in the test, ${\sigma_{\text{x}}}^2$ is the variation of the total test scores.

The CTT statistical measures have a range of desired values that questions and tests in learning environments are recommended to achieve (see Table 3).

Even though the CTT is widely used in evaluating the questions and tests in learning environments (Schmidt and Embretson 2003), it is limited in several ways: 1) Question's difficulty, discrimination, and reliability values vary across different samples of students (Haladyna 1994). For example, questions' are easy when the sample of students used in the analysis have high ability, and questions are difficult when the sample of students have low ability (De Ayala 2009). 2) Students and test characteristics can not be separated and they are interpreted in the context of each other (Hambleton 1991). Question's difficulty, discrimination, and reliability values depend on the sample of students and the ability of students depends on the assessment test. For example, if a test is easy this indicates that students have high ability and vice-versa. 3) CTT is test oriented rather than question oriented, as it can not predict how a particular student may do in a particular assessment question (Hambleton 1991).

These limitations have been addressed by the IRT, which is explained in the following section.

Item Response Theory

Item Response Theory (IRT) is a family of probabilistic models that relates students' ability (θ) to the probability of answering a test question within a particular category (Lord 1980). Similar to CTT, IRT models are used to assess the question's difficulty and discrimination. However, IRT addresses the CTT drawbacks by achieving the following (Baker 2001; Reckase 2009): 1) The question's difficulty and discrimination values measured using IRT are sample independent, i.e., question's difficulty discrimination values does not change across different samples of students such as high ability and low ability students. 2) Students and test characteristics in IRT can be separated; the question's difficulty and discrimination are independent of the sample of students used in the analysis. Moreover, students' ability is independent of the assessment questions

Models

IRT includes the following set of probabilistic models, which differentiate in the number of parameters used to describe the characteristics of the assessment questions:

1) One parameter logistic model (1PL): This is the simplest model in IRT as it has one parameter for describing the characteristics of a student (ability), and one parameter for describing the characteristics of an assessment question (difficulty). This model assumes that all questions in the test are equally discriminating. 1PL model is presented in the following equation:

Table 3 CTT statistical measures desired values

CTT statistical measures	Desired values	Literature
Difficulty index average	[0.30-0.90]	(Doran 1980; Ding et al. 2006)
Discrimination index average	≥ 0.30	(Doran 1980; Zhang and Lidbury 2013)
Point biserial coefficient average	≥ 0.20	(Ding and Beichner 2009; Zhang and Lidbury 2013; Kline 1986; Ding et al. 2006)
Cronbach's α	≥ 0.70	(Doran 1980; Kline 2013a)
Fergusons δ	≥ 0.90	(Kline 1986; 2013a; 2013b)

$$P(X_{ij} = 1|\theta_j, b_i) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} \quad (6)$$

Where X_{ij} represents the response of a student j to question i, X_{ij} = 1 means that question i is answered correctly and X_{ij} = 0 means that question i is answered incorrectly. θ_j represents the ability of student j, and b_i is the difficulty parameter of question i.

2) Two parameter logistic model (2PL): This model is a slightly more complex model, as it considers both the question's difficulty and discrimination. The model is presented in the following equation:

$$P(X_{ij} = 1|\theta_j, a_i, b_i) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (7)$$

Where a_i is the question's discrimination parameter. The higher the value of a_i, the more sharply the question discriminates between high ability and low ability students.

3) Three parameter logistic model (3PL): This model is more complex than the previous models. It considers the possibility that the student correct answers could be obtained by guessing. The model is presented in the following equation:

$$P(X_{ij} = 1|\theta_j, a_i, b_i, G_i) = G_i + (1 - G_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (8)$$

Where G_i is the guessing parameter which accounts for the possibility that all students even the ones with very low ability have a non-zero probability of answering a question correctly by guessing.

Assumptions

In order to use the IRT models to analyse an assessment's test data, the following two assumptions underlying the model must be satisfied (De Ayala 2009; Reckase 2009; Hambleton and Swaminathan 1985; Comer and Kendall 2013; Toland 2014):

1) Unidimensionality: This assumption means that the assessment test measures only one ability parameter (θ), while multi-dimensionality means the test measures more than one ability parameter. Uni-dimensionality could be examined using the Principle Component Analysis (PCA) test (Chou and Wang 2010). PCA outputs the number of components underlying the assessment test. If one component is found the uni-dimensional IRT (UIRT) models can be used to analyse the assessment test data, otherwise the multidimensional IRT (MIRT) can be applied to the assessment test data.

2) Local independence: This assumption states that the only influence on an individual question response is that of the ability parameter being measured (De Ayala 2009). This indicates that there is no influence on the individual question response from other questions or other ability variables. The term local is used to indicate that responses are assumed to be independent at the level of individual students having the same ability (θ). Local independence is examined using the Local Dependence chi-square (LD x^2) test which is applied for each pair of questions in the assessment test (Chen and Thissen 1997). The LD x^2 is computed by comparing the observed and expected frequencies of students'

responses for each pair of questions. In addition, it is applied under the *null hypothesis* that there is local independence between each pair of questions.

Model selection methods

Selecting the IRT model, which is the closest fit to the assessment test data is essential to obtain question's difficulty and discrimination values which are invariant across different samples of students (Hambleton and Swaminathan 1985; Gler et al. 2014). In this paper the following methods have been used to select the IRT model with the closest fit to the assessment test data:

1) The likelihood ratio: The Liklihood Ratio (LR) statistical test (De Ayala 2009; Comer and Kendall 2013; Toland 2014) could be used to select the best IRT model from the three nested models (1PL, 2PL, and 3PL). Moreover, it could be used to select the best model from UIRT and MIRT models, which have different dimensions and the same number of parameters. LR is a chi-square based statistical test and it is measured as the difference between deviances for the two IRT models being compared. The deviance statistic is defined as:

$$-2 * \log(MaximumLikelihood(model)) \tag{9}$$

The maximum likelihood (ML) is obtained for the IRT models using Bock and Aitkin's Expectation-Maximization algorithm (BAEM) (Bock and Aitkin 1982). The LR statistical test is applied under the *null hypothesis* that there is no difference between the two compared models (model 1 and model 2). If the difference between the models deviances which has a chi-square distribution is statistically significant then model 2 has better fit to the assessment test data compared to model 1, otherwise, model 1 has a better fit to the assessment test data.

2) Information theoretic methods: The LR test tends to select models with more parameters (e.g., the 2PL model) which are more complex models and may be a better fit to the assessment test data compared to the models with fewer parameters (e.g., 1PL model) (De Ayala 2009; Kang and Cohen 2007). Therefore, the Akaike's Information Criterion (AIC) (Akaike 1974) and the Bayesian Information Criterion (BIC) (Schwarz 1978) are model selection methods, which penalise the IRT models according to their complexity. They are used as a trade-off between the complexity of the model and the goodness of fit between the model and the assessment test data. Akaike's Information Criterion is measured using the following equation:

$$AIC(model) = -2 * \log(MaximumLikelihood(model)) + 2 * N_{\text{parm}} \tag{10}$$

Where $-2 * \log(MaximumLikelihood(model))$ is the deviance and N_{parm} is the number of parameters being estimated. The model with the smallest AIC is the closest fit to the assessment test data (De Ayala 2009; Toland 2014).

Bayesian Information Criterion is measured using the following equation:

$$BIC(model) = -2 * \log(MaximumLikelihood(model)) + \log(N) * N_{\text{parm}} \tag{11}$$

Where N_{parm} is the number of parameters being estimated, N is the sample size which is the total number of students who attempt the assessment test. The model with the smallest BIC is the closest fit to the assessment test data (De Ayala 2009; Toland 2014).

Equation. 10 shows that AIC penalise the model based on the number of parameters estimated and it does not take into account the sample size. This results in AIC favouring more complex models when the sample size increase (Kang and Cohen 2007; DeMars 2012). On the other hand, BIC tends to select models that are simpler than those selected by AIC when the sample size is large (Kang and Cohen 2007). Equation 11 shows that BIC takes into account the sample size and the penalty for model complexity increases for large samples (DeMars 2012).

Experimental study

This section presents the research questions which will be answered using the evaluation methods discussed in the previous section. In addition, it presents the experimental set-up and participants.

Experiment questions

This experiment aims to answer two main questions:

1. Do the questions and tests generated from ontologies have satisfactory difficulty, discrimination and reliability values?
2. Do the ontology-based generation strategies and the level of the questions in Bloom's taxonomy affect the questions' difficulty and discrimination?

Experimental set-up

A question generator prototype was developed in Java and used to generate true and false, multiple choice, and short answer questions using the ontology-based generation strategies defined by Papasalouros et al. (2017); Papasalouros et al. (2011), Cubric and Tosic (2017), Grubisic (2012); Grubisic et al. (2013), and Al-Yahya (2014); Al-Yahya (2011). Figure 1 shows an example of a class-based strategy integrated in the question generator. The question generator also integrated 20 question stem templates defined by Grubisic (2012); Grubisic et al. (2013), Cubric and Tosic (2017) to auto-generate questions aim to assess student's cognition at the knowledge, comprehension, application and analysis levels in Bloom's taxonomy. Table 4 shows part of the stem templates integrated in the question generator.

Grubisic (2012); Grubisic et al. (2013) knowledge level stem templates focused on assessing if students could recall concepts in the domain ontology and understand the subclasses or superclasses properties between concepts. The comprehension level stem templates focused on the meaning of the concepts in terms of their relationship with other concepts in the domain. Application level stem templates assumed that students are more familiar with the domain ontology being tested, as students are asked about the relationship between individuals and concepts in the domain ontology. Analysis level

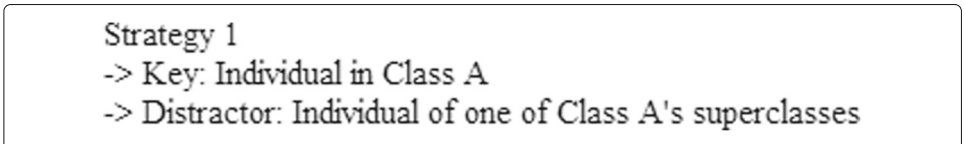

Fig. 1 An example of a Class-based strategy

Table 4 Part of the stem templates analysed in the experiment

Question number	Stem template	Bloom's level	Literature
1	Are Class A and Class B directly connected?	Knowledge	Grubisic et al. (2013)
2	Which one of the following response pairs relates in the same way as: Class A Property Class B	Comprehension	Cubric and Tosic (2017)
3	Which one of the following demonstrates the concept Class A?	Application	Cubric and Tosic (2017)
4	Analyse the following text and decide which one of the following words is a correct replacement for the blank space in the text?	Analysis	Cubric and Tosic (2017)

stem templates focused on assessing the concept's annotation properties and the concept's datatype and object properties with other concepts in the domain ontology. Cubric and Tosic followed a different approach in forming the stem templates. They used words that define each level in Bloom's taxonomy such as demonstrate, define, relate, and analyse (Assessment 2002; Felder and Brent 1997). No generation strategies or stem templates were defined by Papasalouros et al. (2017); Papasalouros et al. (2011), Cubric and Tosic (2017), Grubisic (2012); Grubisic et al. (2013), and Al-Yahya (2014); Al-Yahya (2011) to auto-generate questions which assess students at the synthesis and evaluation levels in Bloom's taxonomy.

The Computer Networks (Murugan et al. 2013) and the OpenCyc (Matuszek et al. 2006) ontologies were used to auto-generate questions which covered the 'transport layer' topic. 44 questions were chosen and syntactically checked by a domain expert who is a lecturer in the School of Computer Science and teaches the Computer Networks course. After that, the questions were imported into Moodle VLE to form three different tests. Tables 5, 6, and 7 illustrate the distribution of the questions generated using the ontology-based generation strategies. Each test contained true and false, multiple choice and short answer questions, and consisted of questions which aim to assess students' cognition at different levels in Bloom's taxonomy. Table 7 shows that the number of short answer questions used in the experiment was small compared to the true and false and multiple choice questions. This is due to that fact that Grubisic (2012); Grubisic et al. (2013) and Al-Yahya (2014); Al-Yahya (2011) defined only two generation strategies and stem templates for generating short answer questions.

The quality of questions generated was evaluated using the CTT and IRT which are explained in details in Sec.2.

Participants

In 2013/2014, third year undergraduate students registered in the Data networking course (TUO a) and the Computer Networks course (TUO b) at the University of Manchester,

Table 5 Distribution of questions based on the generation strategies

Test	Number of questions	Generation strategies		
		Class	Terminology	Property
1	14	1	4	9
2	16	1	4	11
3	14	1	4	9

Table 6 Distribution of questions based on the level of the questions in Bloom's taxonomy

Test	Number of questions	Level of the question bloom's taxonomy			
		Knowledge	Comprehension	Application	Analysis
1	14	4	4	4	2
2	16	4	4	4	4
3	14	4	4	4	2

volunteered to take part in the experiment. In total, 126 students attempted test-one, 88 students attempted test-two, and 89 students attempted test-three. Students accessed the three tests using Moodle VLE. Their responses were recorded and used to analyse the quality of the questions and tests.

Results and discussion

This section illustrates the experiment results obtained using the CTT and IRT. Before applying the IRT models to the assessment test data, the uni-dimensionality and local independence assumptions were investigated. Table 8 illustrates the results obtained by applying the PCA to test-one, which consists of 14 questions and was answered by 126 students. Initially, 14 components were identified; i.e., the number of components equals the number of questions in test-one. Table 8 shows that test-one data results in six components with eigenvalues greater than one. The first component had a 2.225 eigenvalue which is higher than the next five components (1.635, 1.248, 1.213, 1.078, and 1.004). 15.894% of the test variance was explained by the first component and a cumulative variance of 60.02% was explained by the first six components (see Table 8). The results obtained using the PCA suggests that test-one is not unidimensional and it does not measure a single ability parameter. The same analysis were applied to test-two and test-three and the results obtained also suggest that both tests are not unidimensional.

The local dependence assumption was also investigated on test-one, test-two and test-three data using the LD x^2 test. The results revealed that the questions are independent of each other.

After the assumptions were investigated, several IRT models were applied to the three tests and the model selection methods explained in Section Model selection methods were used to select the model with the best fit. The PCA analysis revealed that test-one is not unidimensional, and six components had eigenvalues greater than one. Therefore, the model's data fit analysis was examined using the UIRT, and the MIRT models starting from two dimensions and up to six dimensions. The following abbreviations are used throughout the analysis:

$$\text{UIRT(M)}$$
$$[\,\text{D}] - \text{MIRT(M)}$$

Table 7 Distribution of questions based on types of questions

Test	Number of questions	Types of question		
		True/ False	Multiple choice	Short answer
1	14	4	10	0
2	16	4	11	1
3	14	4	9	1

Table 8 Total variance explained by each component in test-one

Component	Initial eigenvalues		
	Total	% of Variance	Cumulative %
1	2.225	15.894	15.894
2	1.635	11.678	27.572
3	1.248	8.914	36.485
4	1.213	8.661	45.146
5	1.078	7.701	52.847
6	1.004	7.173	60.020
7	0.942	6.731	66.751
8	0.830	5.926	72.677
9	0.785	5.608	78.285
10	0.736	5.258	83.543
11	0.668	4.772	88.315
12	0.620	4.428	92.743
13	0.549	3.923	96.666
14	0.467	3.334	100.000

Where M is the type of IRT model which could be one parameter logistic model (1PL), two parameter logistic model (2PL), or three parameter logistic model (3PL). D is only used with MIRT as it represents the number of dimensions in IRT.

The analysis started with the 2PL model. Table 9 illustrates the likelihood ratio, Akaike's information criterion (AIC), and the Bayesian information criterion (BIC) goodness of fit statistics after applying UIRT (2PL) and MIRT (2PL) models to test-one.

Table 10 shows the chi-square test between several models. The results revealed that AIC, BIC and chi-square tests gave consistent results identifying the 2-MIRT (2PL) model as the best fit for test-one data, as 2-MIRT (2PL) had the smallest AIC and BIC values, and the chi-square test revealed a statistically significant difference between the 2-MIRT (2PL) and the UIRT (2PL) models.

Further investigations were carried out to examine the effect of changing the type of IRT model (e.g., 2PL and 3PL) in 2-MIRT on the goodness of fit statistics. Table 11 shows the goodness of fit statistics for the 2-MIRT (2PL) and the 2-MIRT (3PL) models. The results revealed that 2-MIRT (2PL) fits test-one data better than 2-MIRT (3PL), as it has lower AIC and BIC values, and the chi-square test revealed no statistically significant difference (P-value >0.05) between the 2-MIRT (2PL) and the 2-MIRT (3PL) models (see Table 12). In summary, the 2-MIRT (2PL) model was the closest fit to test-one. The same analysis were applied to test-two and test-three data and the results revealed that the UIRT (2PL) model has the closest fit to test-two and test-three. The 2PL model in the three tests assumes that questions have no guessing parameter.

Table 9 Goodness of fit statistics for UIRT (2PL) and MIRT (2PL) models in test-one

IRT models	-2 log likelihood	Free parameters	AIC	BIC
UIRT (2PL)	2085.82	28	2141.82	2144.63
2-MIRT (2PL)	2057.31	41	2139.31	2143.43
3-MIRT (2PL)	2044.58	56	2156.58	2162.2

Table 10 Chi-square tests comparing the fit of UIRT (2PL) and MIRT (2PL) models

Comparison	x^2	Degree of freedom (DF)	*P*-value
UIRT (2PL) 2-MIRT (2PL)	29.5	13	*P*-value <0.01
2-MIRT (2PL) 3-MIRT (2PL)	11.47	15	*P*-value >0.05

Do the questions and tests generated from ontologies have satisfactory difficulty, discrimination and reliability values?

The questions difficulty indices measured using the CTT when applied to questions administered to third year undergraduate students registered in the Data networking course and the Computer Networks course at the University of Manchester could be summarised as follows: The questions difficulty indices varied from very easy to very difficult in test-one (see Table 13), and very easy to moderately difficult in test-two (see Table 14) and test-three (see Table 15). 16% (7 questions out of 44) of the questions in the three tests were very easy or very difficult which results in low discriminating questions.

The CTT analysis results also revealed when applied to tests administered to third year undergraduate students that the three tests had medium difficulty with 0.525, 0.540, and 0.564 average difficulty index values. The difficulty fall within the CTT desired range of values (see Tables 16, 17, and 18) (Doran 1980; Ding et al. 2006). In addition, the tests' average difficulty index values were very close to 0.5, which is the value that test authors are advised to achieve when constructing questions and where the test have the maximum discrimination (Doran 1980; Mitkov et al. 2017; Mitkov et al. 2006). The maximum discrimination is obtained only when all the students with high ability (students with high scores) answer the questions correctly and all the students with low ability do not answer the questions correctly.

The IRT was also used to assess the question's difficulty due to its invariance assumption. Tables 19, 20, and 21 illustrates the IRT analysis results obtained for test-one, test-two, and test-three accordingly. The results revealed a strong relationship between the difficulty indices obtained using the CTT and IRT *(Pearson R= -0.602, P-value <0.05)*. In addition, the IRT analysis revealed that 22.7% (10 questions out of 44) of the questions were either very easy or very difficult.

The discrimination was also measured for the individual questions and the entire assessment tests. The question discrimination indices obtained using the CTT when applied to the three tests administrated to third year undergraduate students (see Tables 13, 14, and 15) had positive values. This indicates that the auto-generated questions may not need to be reviewed or eliminated from the assessment tests (Doran 1980; Mitkov et al. 2006; Mitkov and Ha 2017). In addition, the three tests had satisfactory average discrimination values above 0.30 (see Tables 16, 17, and 18) which indicates that the questions could efficiently discriminate between high ability and low ability students (Doran 1980; Zhang and Lidbury 2013; Thorndike and Hagen E 2017; Corkins 2009). Similar results were obtained using the IRT, which could be seen in Tables 19, 20, and 21. The results revealed that the questions in the three tests had positive discrimination values

Table 11 Goodness of fit statistics for 2-MIRT (2PL) and 2-MIRT (3PL) models

IRT models	-2 log likelihood	Free parameters	AIC	BIC
2-MIRT (2PL)	2057.31	41	2139.31	2143.43
2-MIRT (3PL)	2042.87	70	2182.87	2389.9

Table 12 Chi-square tests comparing the fit of 2-MIRT (2PL) and 2-MIRT (3PL) models

Comparison	x^2	Degree of freedom (DF)	*P*-value
2-MIRT (2PL) -2-MIRT (3PL)	14.44	29	*P*-value >0.05

and that the auto-generated questions may not need to be reviewed or eliminated from the assessment tests (Baker 2001; Hambleton and Swaminathan 1985).

The CTT was also used to obtain the tests' discrimination power using Ferguson's delta. The results revealed that the three tests had satisfactory discrimination power with Ferguson's delta values above 0.90 which is the discrimination power for normally distributed test scores.

The questions' reliability was measured using the point biserial correlation coefficients (R_{pb}), which is shown in Tables 13, 14, and 15. The results revealed that the question's reliability values in the three tests administrated to third year undergraduate students were positive and the questions' could effectively discriminate between low ability and high ability students as the average point biserial coefficients in each test were satisfactory with values above 0.2.

The test's reliability values was obtained using Cronbach's α, which revealed that test-one and test-two had poor reliability with 0.54, 0.56 reliability values respectively, while test-three had a higher reliability value (0.604), which is considered acceptable. The tests low reliability values obtained using Cronbach's α are due to the fact that the individual questions in each test had satisfactory reliability values (R_{pb}) which are not high enough to improve the tests' overall reliability (Jones 2009). Higher R_{pb} values are desired and lower R_{pb} values indicate that a question is not testing the same educational material or may not be testing the same educational material at the same level (Ding and Beichner 2009). In this experiment the questions are generated from the same domain ontologies (OpenCyc and Computer Networks). As a result the context of the educational material being tested is known. However, the auto-generated questions were designed to assess different educational concepts at different levels of Bloom's taxonomy, which may result in satisfactory reliability values at the questions' level (average R_{pb}) but low reliability values at the test's level (Cronbach's α).

Table 13 Questions' analysis in test-one (number of students' attempts = 126)

Question	Difficulty index	Discrimination index	Rpb
1	Moderately easy (0.770)	Low (0.226)	Low (0.216)
2	Very easy (0.897)	Low (0.258)	Low (0.261)
3	Moderately difficult (0.579)	Medium (0.516)	Medium (0.445)
4	Moderately difficult (0.365)	Medium (0.355)	Low (0.270)
5	Moderately difficult (0.373)	Low (0.290)	Low (0.233)
6	Moderately difficult (0.508)	High (0.710)	Medium (0.479)
7	Moderately difficult (0.532)	High (0.613)	Medium (0.460)
8	Moderately easy (0.635)	High (0.677)	Medium (0.541)
9	Moderately easy (0.651)	High (0.613)	Medium (0.521)
10	Very difficult (0.333)	Low (0.226)	Medium (0.333)
11	Very difficult (0.087)	Low (0.129)	Low (0.267)
12	Moderately difficult (0.540)	Medium (0.516)	Medium (0.347)
13	Moderately difficult (0.595)	High (0.645)	Medium (0.496)
14	Moderately difficult (0.484)	Medium (0.484)	Medium (0.362)

Table 14 Questions' analysis in test-two (number of students' attempts = 88)

Question	Difficulty index	Discrimination index	Rpb
1	Moderately difficult (0.580)	Medium (0.409)	Low (0.297)
2	Moderately easy (0.739)	Medium (0.318)	Low (0.241)
3	Moderately difficult (0.545)	Low (0.273)	Low (0.196)
4	Moderately easy (0.761)	Medium (0.318)	Low (0.274)
5	Moderately difficult (0.432)	Medium (0.409)	Medium (0.382)
6	Moderately difficult (0.420)	Medium (0.500)	Medium (0.488)
7	Moderately easy (0.614)	Low (0.182)	Low (0.193)
8	Very difficult (0.295)	Medium (0.455)	Medium (0.368)
9	Moderately difficult (0.432)	High (0.682)	Medium (0.489)
10	Moderately difficult (0.523)	Medium (0.318)	Low (0.263)
11	Very difficult (0.284)	Medium (0.409)	Medium (0.378)
12	Moderately easy (0.773)	Medium (0.545)	Medium (0.457)
13	Moderately difficult (0.489)	High (0.727)	Medium (0.524)
14	Moderately easy (0.625)	Medium (0.455)	Medium (0.455)
15	Moderately easy (0.636)	Medium (0.500)	Medium (0.371)
16	Moderately difficult (0.500)	Medium (0.545)	Medium (0.518)

Do the ontology-based generation strategies and the level of the questions in Bloom's taxonomy affect the questions' difficulty and discrimination?

This section studies the effect of the ontology-based generation strategies and the level of questions in Bloom's taxonomy on the questions' difficulty and discrimination obtained using the CTT (dependent on the sample of students) and the IRT (independent from the sample of students).

The study was carried out on the CTT difficulty and discrimination indices obtained for the whole 44 questions (total number of assessment question in test-one, test-two, and test-three), and on the IRT difficulty and discrimination indices, which did not experience variance across different samples of students. The invariance of IRT measurements was tested for the whole 44 questions by dividing the students in each test (test-one, test-two, and test-three) into two groups: low ability students (students with test scores less

Table 15 Questions' analysis in test-three (number of students' attempts = 89)

Question	Difficulty index	Discrimination index	Rpb
1	Moderately easy (0.708)	Medium (0.409)	Low(0.294)
2	Moderately easy (0.652)	Medium (0.318)	Low (0.258)
3	Moderately easy (0.764)	Medium (0.409)	Medium (0.360)
4	Moderately easy (0.640)	Medium (0.318)	Medium (0.300)
5	Moderately easy (0.663)	High (0.636)	Medium (0.504)
6	Moderately difficult (0.483)	Medium (0.591)	Medium (0.429)
7	Very difficult (0.337)	Medium (0.455)	Medium (0.433)
8	Moderately difficult (0.427)	Medium (0.591)	Medium (0.524)
9	Moderately easy (0.798)	Medium (0.409)	Medium (0.373)
10	Moderately easy (0.674)	Medium (0.500)	Medium (0.428)
11	Very difficult (0.247)	Low (0.273)	Low (0.239)
12	Moderately difficult (0.360)	High (0.727)	Medium (0.508)
13	Moderately difficult (0.416)	Medium (0.591)	Medium (0.465)
14	Moderately easy (0.730)	Medium (0.545)	Medium (0.476)

Table 16 Test-one analysis (number of students' attempts = 126)

Test statistics	Value	Description	Desired values
Questions' difficulty index average	0.525	Medium difficulty	[0.3-0.9]
Questions' discrimination index average	0.445	Satisfactory	≥0.3
Questions' reliability average (point biserial coefficient)	0.266	Satisfactory	≥0.2
Test reliability index	0.540	Poor	≥0.7
Test discrimination power (Ferguson's delta)	0.540	Satisfactory	≥0.9

than 50%) and high ability students (students with test scores above or equal 50%) following the approach in (Hambleton and Swaminathan 1985). Students in each test could also be divided according to their gender or year of study (De Ayala 2009; Crocker and Algina 1986). However, this was not applicable in the experiment carried out in this paper due to the large difference in students' numbers when the students in each test were divided according to their gender or year of study. The IRT model, which has the best fit to the whole sample of students in each test, was applied to the low ability and high ability sample of students separately to obtain the questions' difficulty and discrimination indices. The standard deviation was measured for the question's difficulty and discrimination across the three groups of students: the whole sample of students, students with low ability, and students with high ability. Questions with large standard deviation values compared to other questions in the assessment test were considered outliers as they experienced high variance across the three groups of students. In total 10 questions out of 44 violated the IRT invariance assumption and were not used in the upcoming evaluations.

Does the ontology-based question generation strategy affect the question difficulty and discrimination?

The results revealed that generating questions using different generation strategies (class, terminology, and property) appear to affect the question difficulty and discrimination obtained using the CTT and IRT. A statistically significant difference in the CTT difficulty indices *($U = 69$, P-value <0.05)* and IRT difficulty indices *($U = 26$, P-value <0.05)* was found between questions generated using the terminology-based strategies and questions generated using the property-based strategies. This suggests that students found questions which assess their knowledge about an educational concept and how it is related to other concepts using the superclass and subclass properties easier than questions which assess their knowledge about the concept's object, datatype, and annotation properties). Questions generated using terminology-based strategies had higher CTT difficulty indices *(Spearman's $R = 0.476$, P-value <0.01)* and lower IRT difficulty indices *(Spearman's $R = -0.583$, P-value <0.01)*. Higher difficulty indices in CTT means the question is easy while in IRT it means the question is more difficult.

Table 17 Test-two analysis (number of students' attempts = 88)

Test statistics	Value	Description	Desired values
Questions' difficulty index average	0.540	Medium difficulty	[0.3-0.9]
Questions' discrimination index average	0.440	Satisfactory	≥0.3
Questions' reliability average (point biserial coefficient)	0.360	Satisfactory	≥0.2
Test reliability index	0.560	Poor	≥0.7
Test discrimination power (Ferguson's delta)	0.955	Satisfactory	≥0.9

Table 18 Test-three analysis (number of students attempts = 89)

Test statistics	Value	Description	Desired values
Questions difficulty index average	0.564	Medium difficulty	[0.3-0.9]
Questions discrimination index average	0.484	Satisfactory	≥0.3
Questions reliability average (point biserial coefficient)	0.399	Satisfactory	≥0.2
Test reliability index	0.604	Acceptable	≥0.7
Test discrimination power (Ferguson's delta)	0.966	Satisfactory	≥0.9

No statistical significant difference was found in the CTT and IRT difficulty indices between questions generated using class-based strategies and terminology-based strategies, and between questions generated using class-based strategies and property-based strategies. This suggests that the students found questions auto-generated using the individual and class relationship in the ontology as difficult as questions generated using the terminology-based strategies and the property-based strategies.

The questions' discrimination indices were also investigated and the results revealed a statistical significant difference in CTT discrimination indices *(U= 74, P-value <0.05)*, CTT R_{pb} *(U = 59, P-value <0.05)*, and IRT discrimination indices *(U = 43, P-value <0.05)* between questions generated using the terminology-based strategies and questions generated using the property-based strategies. Questions generated using the terminology-based strategies have better discrimination values compared to questions generated using the property-based strategies; questions generated using terminology-based strategies had higher CTT discrimination indices *(Spearman's R = 0.454, P-value <0.01)*, higher CTT R_{pb} *(Spearman's R = 0.521, P-value <0.01)*, and higher IRT discrimination indices *(Spearman's R = 0.456, P-value <0.01)* compared to questions generated using the property-based strategies.

The results also revealed that there is a statistical significant difference in CTT discrimination indices *(U = 2, P-value <0.05)* and CTT R_{pb} *(U = 2, P-value <0.05)* between questions auto-generated using class-based strategies and terminology-based strategies. Questions generated using terminology based strategies had higher CTT discrimination indices *(Spearman's R = 0.63, P-value <0.05)*, and higher CTT R_{pb} *(Spearman's R = 0.617, P-value <0.05)* compared to questions generated using class-based strategies.

Table 19 Questions' difficulty and discrimination indices using IRT (test-one)

Q	Difficulty index	Discrimination index
1	Medium (-0.030)	Moderate (0.986)
2	Very easy (-2.603)	Low (0.484)
3	Easy (-1.800)	Perfect (1.712)
4	Medium (-0.374)	Moderate (1.043)
5	Difficult (0.837)	Moderate (0.740)
6	Very difficult (7.354)	Very low (0.071)
7	Medium (-0.131)	Moderate (1.073)
8	Easy (-0.515)	High (1.458)
9	Easy (-0.584)	High (1.421)
10	Difficult (0.854)	Moderate (0.972)
11	Very difficult (3.323)	Moderate (0.779)
12	Medium (-0.133)	High (1.500)
13	Medium (-0.391)	Moderate (1.305)
14	Medium (0.072)	Moderate (1.105)

Table 20 Questions' difficulty and discrimination indices using IRT (test-two)

Question	Difficulty index	Discrimination index
1	Easy (-1.81)	Very low (0.18)
2	Very easy (-2.23)	Low (0.49)
3	Easy (-0.99)	Very low (0.19)
4	Very easy (-10.33)	Very low (0.11)
5	Difficult (0.66)	Low (0.43)
6	Medium (0.42)	Moderate (0.89)
7	Very easy (-8.44)	Very low (0.05)
8	Difficult (1.16)	Moderate (0.86)
9	Medium (0.28)	Moderate (1.22)
10	Medium (-0.29)	Very low (0.32)
11	Difficult (1.87)	Low (0.52)
12	Easy (-1.66)	Moderate (0.84)
13	Medium (0.02)	Perfect (2.75)
14	Easy (-1.5)	Low (0.35)
15	Easy (-0.98)	Low (0.63)
16	Medium (-0.01)	High (1.48)

However, this result depends on the sample group of students as no statistical significant difference was found in the IRT discrimination indices (sample independent) between questions generated using the class-based strategies and questions generated using the terminology-based strategies. In addition, no statistical significant difference was found in the CTT discrimination indices, the CTT R_{pb}, and IRT discrimination indices between the questions generated using class-based and property based strategies. This suggests that the class-based and property-based strategies produce questions, which have similar discrimination indices.

Do Bloom's taxonomy stem templates affect the question difficulty and discrimination?

Grubisic (2012); Grubisic et al. (2013), Cubric and Tosic (2017) defined several question stem templates to auto-generate questions aimed to assess students' cognition at different levels in Bloom's taxonomy. However, they never investigated

Table 21 Questions' difficulty and discrimination indices using IRT (test-three)

Question	Difficulty index	Discrimination index
1	Very easy (-2.51)	Low (0.37)
2	Very easy (-3.32)	Very low (0.19)
3	Very easy (-2.06)	Low (0.62)
4	Easy (-1.51)	Low (0.42)
5	Easy (-0.76)	Moderate (1.17)
6	Medium (0.09)	Moderate (0.70)
7	Difficult (0.88)	Moderate (0.87)
8	Medium (0.31)	Moderate (1.08)
9	Easy (-1.48)	Moderate (1.17)
10	Easy (-0.99)	Moderate (0.86)
11	Very difficult (6.10)	Very low (0.18)
12	Difficult (0.60)	Moderate (1.18)
13	Medium (0.38)	Moderate (1.04)
14	Easy (-1.05)	Moderate (1.32)

whether the question stem templates order the auto-generated questions according to their easiness in Bloom's taxonomy or whether the question stem templates affect the questions' discrimination. Therefore, this section investigates the effect of the level of question in Bloom's taxonomy on the question difficulty and discrimination.

The results revealed that the question stem templates defined by Grubisic (2012); Grubisic et al. (2013), Cubric and Tosic (2017) appear to order questions according their easiness in Bloom's taxonomy. A statistical significant difference in the CTT difficulty indices *(U = 21, P-value <0.05)* and IRT difficulty indices *(U = 13, P-value <0.05)* was found between questions in the knowledge and comprehension levels. Questions generated to assess the students in the knowledge level are easier than questions generated to assess the students in the comprehension level, as they have higher CTT difficulty indices *(Spearman's R = 0.614, P-value <0.01)* and lower IRT difficulty indices *(Spearman's R = -0.616, P-value <0.01)*. The results are expected as the knowledge level stem template shown in Table 4 focused on assessing whether students could recall concepts and are aware of the subclass and superclass relationships between concepts. However, the comprehension level stem templates focused on students' understanding about the similarity of the relationship between concepts (see question 2 in Table 4) and whether students know all the concept's subclasses and superclasses.

The results also revealed that questions in the knowledge level were easier than questions in the application and analysis level. This is due to the fact that the application level stem templates defined by Grubisic (2012); Grubisic et al. (2013), Cubric and Tosic (2017) focused on the relationship between the individual and superclass (see question 3) as students need to provide an example of the concept he/she learned. Similarly, in the analysis level stem templates students are assessed on the annotation and object properties in classes and individuals (see question 4). For students, these stem templates are harder than knowledge level stem templates which focus on recalling concepts in the domain ontology.

However, no statistical significant difference in the CTT difficulty indices and IRT difficulty indices was found between the other levels in Bloom's taxonomy. This suggests that comprehension, application, and analysis level questions appeared to have to the same difficulty to students.

Questions' discrimination was also investigated and the results revealed that the knowledge level questions, which are the easiest questions, tend to have lower discrimination compared to comprehension, application and analysis level questions. On the other hand, no statistical significant difference in the CTT discrimination indices, CTT R_{pb}, and IRT discrimination indices was found between comprehension, application, and analysis level questions which suggest that comprehension, application and analysis question stem templates auto-generate questions which have the same discrimination.

Conclusion and future work

This paper presented the experiment carried out to analyse the quality of the questions, which were generated using Papasalouros et al. (2017); Papasalouros et al. (2011), Cubric and Tosic (2017), Grubisic (2012); Grubisic et al. (2013), and Al-Yahya (2014); Al-Yahya (2011) question generators. It has three main contributions to the field of ontology-based

question generators: 1) Developing an ontology-based question generator which integrates the preexisting stem templates and generation strategies to generate questions that assess students at different levels in Bloom's taxonomy. 2) providing a quantitative analysis for the auto-generated questions using the CTT and IRT statistical methods. 3) Studying the effect of the ontology-based generation strategies and the level of the questions in Bloom's taxonomy on the questions quality measurements.

The results obtained using the CTT revealed that the three assessment tests formed from the auto-generated questions had medium difficulty values, which are very close to the value (0.5) that the test authors are advised to achieve when constructing tests Doran (1980); Mitkov et al. (2017); Mitkov et al. (2006). In addition, the results revealed that the questions and tests had satisfactory positive discrimination values, which indicate that the questions and tests could effectively discriminate between high ability and low ability students, and that the questions may not need to be reviewed or eliminated from the assessment tests Doran (1980); Mitkov et al. (2006); Mitkov and Ha (2017). In addition to the CTT, the Item Response Theory (IRT) was used to assess the quality of the auto-generated questions because of its invariant assumption. The IRT analysis revealed similar results to the CTT, as the questions' discrimination indices had positive values which justify that the auto-generated questions may not need to be reviewed or eliminated from the assessment tests Baker (2001); Hambleton and Swaminathan (1985).

As mentioned earlier, this paper also investigated the effect of the ontology-based generation strategies and the level of the questions in Bloom's taxonomy on the questions quality measurements. The results revealed that the generation strategies and the level of the questions in Bloom's taxonomy affect the question's difficulty and discrimination. This provides guidance for developers and researchers working in the field of ontology-based question generators.

The analysis results obtained were based on 44 questions generated from the 'transport layer' topic and used in three different tests which consequently consists of 14, 16 and 14 questions. The experiment could be enhanced in the future work by: 1) increasing the number of questions in each test. 2) Increasing the number of students participating in the experiment. 3) Generating questions from different topics in the computer networks domain or different domains (e.g., medicine).

The experiment results obtained using the CTT and IRT could be used in future work to build a prediction model using machine learning techniques (e.g., multiple linear regression James et al. (2014)) to predict the question's difficulty (very difficult, moderately difficult, moderately easy, and very easy) and discrimination (low, medium, and high) in the computer networks domain using the following two features: the ontology-based generation strategy and the level of the question in Bloom's taxonomy. This will help researchers and developers save time and effort in terms of testing the auto-generated questions on real students Arnold et al. (1996). In addition, the ontology-based question generator developed for the purpose of analysing the auto-generated questions quantitatively could be enhanced in the future work to auto-generate personalised formative feedback which takes into account the question characteristics (e.g., the level of question in Bloom's taxonomy) Mason and Bruning (2001).

Acknowledgments
Not applicable.

Funding
Not applicable.

Availability of data and materials
Students responses to the three tests which were evaluated in this paper can be found in the link below. https://drive.google.com/open?id=0B25z6hoT8MGnNWZxUkdLbmdoMlE

Authors' contributions
MD is the main author in this manuscript, she collected data, performed analysis on all samples, interpreted data, and wrote manuscript. MMG supervised the development of work. He also helped in reviewing the article and gave the final approval of the version to be submitted. NF helped in the experimental design part and setting up the environment for collecting the participants data. The author also helped in reviewing the article. All authors read and approved the final manuscript.

Ethics approval and consent to participate
The research study carried out in this paper has been approved by COMMITTEE ON THE ETHICS OF RESEARCH ON HUMAN BEINGS at the University of Manchester in 2013.

Consent for publication
Not applicable.

Competing interests
The authors declares that they have no competing interests.

Publisher's Note

Received: 6 July 2017 Accepted: 7 November 2017
Published online: 15 November 2017

References

H Akaike, A new look at the statistical model identification. Automatic Control IEEE Trans. **19**(6), 716–723 (1974)
M Al-Yahya, in *Advanced Learning Technologies (ICALT) 2011 11th IEEE International, Conference on*, Ontoque: a question generation engine for educational assesment based on domain ontologies (IEEE, 2011), pp. 393–395
M Al-Yahya, Ontology-based multiple choice question generation. Sci World J (2014)
T Alsubait, B Parsia, U Sattler, in *OWLED*, Generating multiple choice questions from ontologies: Lessons learnt, (2014), pp. 73–84
M Al-Yahya, Ontology-based multiple choice question generation. Sci World J (2014)
M Al-Yahya, H Al-Khalifa, A Bahanshal, I Al-Odah, N Al-Helwah, An ontological model for representing semantic lexicons: an application on time nouns in the holy quran. Arab J Sci Eng. **35**(2), 21–35 (2010)
S Alagumalai, DD Curtis, *Classical test theory*. (Springer, 2005), pp. 1–14
LW Anderson, LA Sosniak, *Bloom's taxonomy: A forty-year retrospective.ninety-third yearbook of the national society for the study of education*, (1994)
K Arnold, J Gosling, D Holmes, *The Java programming language*, Vol 2. (Addison-wesley Reading, 1996)
FB Baker, *The basics of item response theory*. (ERIC, 2001)
Bloom BS, *C o C. Examiners, University, Taxonomy of educational objectives,* Vol. 1. (David McKay, New York, 1956)
R Bock, M Aitkin, Marginal maximum likelihood estimation of item parameters. Psychometrika. **47**(3), 369–369 (1982)
WN Borst, *Construction of engineering ontologies for knowledge sharing and reuse*. (Universiteit Twente, 1997)
JD Brown, *Testing in language programs*. (Prentice Hall Regents, New Jersey, 1996)
CA Assessment, Assessment of higher order skills (2002). http://www.caacentre.ac.uk/resources/faqs/higher.shtml
Y-T Chou, W-C Wang, Checking dimensionality in item response models with principal component analysis on standardized residuals. Educ. Psychol. Meas. **70**(5), 717–731 (2010)
W-H Chen, D Thissen, Local dependence indexes for item pairs using item response theory. J. Educ. Behav. Stat. **22**(3), 265–289 (1997)
L Cohen, L Manion, K Morrison, *Research methods in education*. (Routledge, 2013)
JS Comer, PC Kendall, *The Oxford Handbook of Research Strategies for Clinical Psychology*. (Oxford University Press, 2013)
J Corkins, *The Psychometric Refinement of the Materials Concept Inventory (MCI)*. (ProQuest, 2009)
LJ Cronbach, RJ Shavelson, My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas. **64**(3), 391–418 (2004)
L Crocker, J Algina, *Introduction to classical and modern test theory*. (ERIC, 1986)
M Cubric, M Tosic, Towards automatic generation of e-assessment using semantic web technologies. Intl J e-Assessment (2017)
CP Dancey, J Reidy, *Statistics Without Maths for Psychology: Using Spss for Windows*. (Prentice-Hall Inc., 2004)
RJ De Ayala, *Theory and practice of item response theory*. (Guilford Publications, 2009)
CE DeMars, Confirming testlet effects. Appl. Psychol. Meas. **36**(2), 104–121 (2012)
L Ding, R Beichner, Approaches to data analysis of multiple-choice questions (2009)
L Ding, R Chabay, B Sherwood, R Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. Phys Rev Special Topics-Physics Educ Res. **2**(1) (2006)
RL Doran, *Basic measurement and evaluation of science instruction*. (National Science Teachers Association, Washington, DC, 1980)
R Ebel, *Essentials of Educational Measurement*. (Prentice-Hall, 1979). http://books.google.co.uk/books?id=eEv0NeqTUXYC

M Erguven, Two approaches to psychometric process: Classical test theory and item response theory. J Educ. **2**(2), 23–30 (2014)
G Ferguson, On the theory of test development. Psychometrika. **14**, 61–68 (1949)
RM Felder, R Brent, Objectively speaking. Chem. Eng. Educ. **31**, 178–179 (1997)
G Ganapathi, R Lourdusamy, V Rajaram, in *World Congress on Engineering*, Towards ontology development for teaching programming language, (2017)
TR Gruber, A translation approach to portable ontology specifications. Knowl Acquisition. **5**(2), 199–220 (1993)
A Grubisic, Adaptive students knowledge acquisition model in elearning systems. Thesis (2012)
A Grubisic, S Stankov, B Žitko, in *ICIIS 2013: International Conference on Information and Intelligent Systems*, Stereotype student model for an adaptive e-learning system, (2013)
N Gler, GK Uyank, GT Teker, Comparison of classical test theory and item response theory in terms of item parameters. Eur. J. Res. Educ. **2**(1), 1–6 (2014)
M Hankins, Questionnaire discrimination:(re)-introducing coefficient d. BMC Med Res Methodol. **7**(1) (2007)
TM Haladyna, *Developing and validating multiple-choice test items/Thomas M.Haladyna*, (Hillsdale, NJ; Hove, UK:Erlbaum, Hillsdale, NJ; Hove, UK, 1994)
RK Hambleton, *Fundamentals of item response theory,* Vol 2. (Sage publications, 1991)
RK Hambleton, H Swaminathan, *Item response theory: Principles and applications*, Vol 7. (Springer, 1985)
G James, D Witten, T Hastie, *An introduction to statistical learning: With applications in r*, (2014)
A Jones, *Using the right tool for the job: An analysis of item selection statistics for criterion-referenced tests*. (ProQuest, 2009)
T Kang, AS Cohen, Irt model selection methods for dichotomous items. Appl.Psychol. Measurement. **31**(4), 331–358 (2007)
A Kouneli, G Solomou, C Pierrakeas, A Kameas, *Modeling the knowledge domain of the java programming language as an ontology*. (Springer, 2012), pp. 152–159
DR Krathwohl, A revision of bloom's taxonomy: An overview. Theory Into Pract. **41**(4), 212–218 (2002)
P Kline, *A handbook of test construction: Introduction to psychometric design*. (Methuen, 1986)
Kline, P, *Handbook of psychological testing*. (Routledge, 2013a)
P Kline, *Personality: The psychometric view*. (Routledge, 2013b)
M-C Lee, DY Ye, TI Wang, *Fifth IEEE International Conference on. Java learning object ontology*. (IEEE, 2005), pp. 538–542
FM Lord, *Applications of item response theory to practical testing problems*. (Routledge, 1980)
C Matuszek, J Cabral, MJ Witbrock, J DeOliveira, in *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, An introduction to the syntax and content of cyc, (2006), pp. 44–49
BJ Mason, R Bruning, *Providing feedback in computer-based instruction: What the research tells us*, (2001). http://dwb.unl.edu/Edit/MB/MasonBruning.html
R Mitkov, LA Ha, A Varga, L Rello, in *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation (Association for Computational Linguistics, 2017), pp. 49–56
R Mitkov, L An Ha, N Karamanis, A computer-aided environment for generating multiple-choice test items. Nat. Lang. Eng. **12**(02), 177–194 (2006)
R Mitkov, LA Ha, in *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, Computer-aided generation of multiple-choice tests (Association for Computational Linguistics, 2017), pp. 17–22
S Murugan, RP Bala, G Aghila, An ontology for exploring knowledge in computer networks. Int. J. Comput. Sci. Appl. (IJCSA). **3**(4), 13–21 (2013)
OpenCyc, Opencyc for the semantic web. http://sw.opencyc.org/
A Papasalouros, K Kanaris, K Kotis, in *e-Learning*, Automatic generation of multiple choice questions from domain ontologies, (2017), pp. 427–434
A Papasalouros, K Kotis, K Kanaris, Automatic generation of tests from domain and multimedia ontologies. Interact Learn Environ. **19**(1), 5–23 (2011)
Protege ontology library - protege wiki (2017). http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library
MD Reckase, *Multidimensional item response theory*. (Springer, 2009)
KM Schmidt, SE Embretson, Item response theory and measuring abilities. Handb Psychol (2003)
G Schwarz, Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
D Seyler, M Yahya, K Berberich, Knowledge questions from knowledge graphs (2016). arXiv preprint arXiv: 1610.09935
MD Toland, Practical guide to conducting an item response theory analysis. J Early Adolesc. **34**, 120–151 (2014)
TUO, Manchester, Data networking. http://www.eee.manchester.ac.uk
TUO, Manchester, Computer networks. http://studentnet.cs.manchester.ac.uk/ugt/COMP28411/syllabus
RL Thorndike, Hagen E, Measurement and evaluation in psychology and education (2017)
R Studer, VR Benjamins, D Fensel, Knowledge engineering: principles and methods. Data Knowl. Eng. **25**(1), 161–197 (1998)
Y Susanti, T Tokunaga, H Nishikawa, H Obari, Evaluation of automatically generated english vocabulary questions. Res Pract Technol Enhanced Learn. **12**(1), 11 (2017)
M Uschold, M Gruninger, Ontologies: Principles, methods and applications. Knowl. Eng. Rev. **11**(02), 93–136 (1996)
EV Vinu, PS Kumar, Automated generation of assessment tests from domain ontologies. Semantic Web. **8**(6), 1023–1047 (2017)
F Zhang, BA Lidbury, Evaluating a genetics concept inventory. Bioinformatics: Concepts Methodol Tools Appl, 29–41 (2013)