Creating a software application to help university educators to reflect on the cognitive complexity of their exam questions, using Bloom’s Taxonomy and automated classification

Andrew Valentine and Eduardo Araujo Oliveira
The University of Melbourne

Previous research has shown that many university educators struggle to accurately evaluate the cognitive complexity of exam questions (and overall exams) which they write, based on Bloom's Taxonomy. This can lead to concerns about the design of exams. Software tools could possibly assist educators via automated classification methods. This paper reports a work-in-progress project that is creating a software application (tool) to assist university educators with writing exams. We evaluate 3 methods of automated classification including keywords-based approach, OpenAI evaluation, and an existing algorithm. The tool is designed to be able to help educators to reflect on their exam, by providing educators with meaningful feedback on question complexity and the overall exam, assisting in exam design. The software tool developed in this study is expected to benefit educators by providing objective feedback and serving as a professional development resource.

Keywords: blooms taxonomy, exams, teacher judgement, cognitive workload

Introduction

In university settings it is common for educators to write exams that students complete at the end of a subject. When constructing the exams, it is often expected that the questions will vary in difficulty. The purpose is that students will be required to complete a series of questions which are at different levels of complexity, meaning it will be possible to differentiate between students who demonstrate a lower level of understanding of the subject material, and students who demonstrate a higher level of understanding of the subject material.

However, it can sometimes be difficult for educators to accurately evaluate the complexity of the exam questions that they write, and the overall exam as a whole. A review by van de Watering & van der Rijt (2006) investigated the ability of teachers to accurately estimate the difficulty levels of assessment items, concluding that “in higher education, results show that teachers are able to estimate the difficulty levels correctly for only a small proportion of the assessment items. They overestimate the difficulty level of most of the assessment items”. In contrast, a recent review evaluated 40 years of research on the accuracy of teacher judgements and concluded that teachers tend to overestimate student performance (Urhahne, &Wijnia, 2021). In both cases poses a problem because it means that the design of an exam may be compromised if the educator cannot accurately identify the difficulty of each question, and the overall exam.

For example, educators may perceive that it is straightforward for most students in their subject to complete an exam that they have written within the allocated time, but it may actually be quite challenging for many students. This is especially true for new or inexperienced educators, who have less experience to draw upon. In many universities it is common practice that an exam may be written by one educator and then the exam is checked or evaluated by a departmental colleague as a means of quality control. While this is an effective means of quality control, it poses risks if both educators are unable to accurately evaluate the difficulty of the exam. This raises a question as to whether it may be possible to support this means of quality control by also using technology which can more objectively evaluate the difficulty of exam questions and overall exam, using Bloom’s Taxonomy as a guide.

This paper reports a work-in-progress project that is creating a software application (tool) to assist university educators with writing exams. The tool is designed to be able to help educators to reflect on the cognitive complexity (or difficulty) of their exam questions, and the overall exam as a whole, with the intention that educators can be more aware of ways to enhance the quality of their exam questions, and overall exam. The tool is not intended to tell educators what difficulty the exam should be, but to provide an external means of gaining additional feedback on their exam.

Background

Bloom's taxonomy

This study utilises the cognitive dimension of the revised version of Bloom's Taxonomy (Krathwohl, 2002) to assist with the evaluation of the complexity of exam questions. The cognitive dimension of the revised version of Bloom's Taxonomy asserts that there are 6 dimensions (or levels) of complexity; Remember, Understand, Apply, Analyse, Evaluate, Create. Within the taxonomy, there are verbs that are associated with each level of complexity. One challenge is that different versions of the mapping of verbs to levels of complexity exist, meaning that the level of complexity may vary depending on which reference is used. In this study, we focused on using the table of verbs presented in the study by Shaikh, Daudpotta, & Imran (2021). Another challenge is that some verbs are often repeated across more than one level (Das, Das, & Basu, 2021), meaning it introduces ambiguity about the level of complexity of a question when this verb is used in the wording.

Research has shown that university educators were unable to accurately evaluate exam questions based on Bloom's Taxononomy (Karpen & Welch, 2016), and that the evaluations given by university educators may be highly inconsistent (Karpen & Welch, 2016). This is highly concerning as it demonstrates that educators may not be able to accurately determine the complexity of their exam questions, and may need more professional development or external assistance to ensure that their exams are written to the complexity that the educator intends. University educators have also been shown to perceive that multiple choice questions can test higher order thinking skills, but primarily only test the levels of "apply" and "analyse" (Liu et al., 2023).

Automated classification of exam questions

Many previous studies have investigated the use of technology to assist with automation of classification of exam questions or assessment items. However, these studies highlight that just how educators can have difficulty accurately evaluating the level of question complexity (Karpen & Welch, 2016), many of the technology assisted methods are also often inaccurate. For example, Bengio et al. (2007) investigated automatically classifying course learning outcomes and question statements to different level of Bloom's taxonomy, using only a keywords-based approach. It was found that the average accuracy across the six cognitive domain levels of Bloom's Taxonomy was only 47%. This highlights that only using a keywords-based approach may lead to classifications which are quite inaccurate.

Shaikh, Daudpotta, & Imran (2021) investigated automatically classifying course learning outcomes and assessment question items to different level of Bloom's taxonomy, using both a keywords-based approach and an LSTM (long short-term memory networks) based deep learning model. It was found that the keywords-based approach had a low level of accuracy (55%) for both the course learning outcomes and assessment question items, similar to the findings of Bengio et al. (2007). However, the LSTM based deep learning model had an accuracy of 87% for course learning outcomes and 74% for assessment question items (Shaikh, Daudpotta, & Imran, 2021). Mohammed & Omar (2020) investigated classification of examination questions based on the cognitive domain of Bloom's taxonomy using modified TF-IDF and word2vec. TF-IDF and word2vec are two popular techniques used in natural language processing (NLP) to represent and analyse text. Using three different classifiers and two datasets of 141 and 600 questions, it was found to have an accuracy of between 71% and 90%, depending on the classification and dataset used. Zhang et al. (2021) investigated using automatically classifying multiple choice computing education questions using Bloom's Taxonomy, using Google's BERT as the base model. Initially the accuracy was 59%, but this increased to 82% once questions in three levels of bloom's taxonomy were removed from the training dataset due to low numbers of questions. Overall, these studies show that using a keywords-based approach alone may not be sufficient to accurately identify the complexity of exam questions, and that more comprehensive NLP techniques may be necessary.

Description of work undertaken, results, and future work

There are two main components of the software application tool; (i) the front-end of the system which educators will interact with using a user interface, and (ii) the back-end part of the system which will analyse the information entered into the user interface by the user. The algorithm used to determine the cognitive complexity of each exam question, and the overall exam, is part of the back-end part of the system. Figure 1 provides an overview of the example intended usage of the system. The educator will enter the exam questions into the user inface of the system, which will also have various options to allow for different types of questions such as multiple choice questions, short questions, or extended questions. The ability to characterise questions by type is necessary because the algorithm may vary, depending on the type of question being asked.

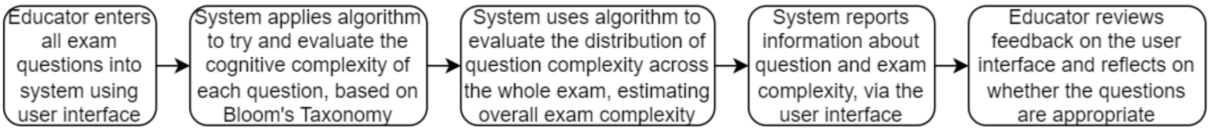
Educator enters all exam questions into system using user interface
System applies algorithm to try and evaluate the cognitive complexity of each question, based on Bloom's Taxonomy
System uses algorithm to evaluate the distribution of question complexity across the whole exam, estimating overall exam complexity
System reports information about question and exam complexity, via the user interface
Educator reviews feedback on the user interface and reflects on whether the questions are appropriate

Figure 1: Explanation of how the system operates and provides feedback to educators

Selection of algorithm

A primary component of the back-end of the system is the algorithm which determines the complexity of each of the exam questions, and the overall exam. It was necessary to either select an existing algorithm (or approach) or to create a new one, which could be used to evaluate the complexity of the exam questions. However, as noted previously in the background section of this study, there are challenges because there are not currently any automated classification approaches which have achieved an accuracy of 100%. Nevertheless, being pragmatic it was necessary to select an algorithm or approach which would allow for a high degree of accuracy. Using an algorithm with high accuracy is imperative, because the lower the accuracy, the less helpful the feedback that can be provided to educators.

We decided to implement a version of the TF-IDF (Term-frequency Inverse Document Frequency) automated classification algorithm implemented in the study by Mohammed & Omar (2020). TF-IDF is a well-known statistical feature and has been used for classifying exam style questions according to Bloom's Taxonomy (Aninditya, Hasibuan, & Sutoyo, 2019; Mohammed & Omar, 2020). For an in-depth explanation of TF-IDF please refer to the study by Aninditya, Hasibuan, & Sutoyo, E. (2019). The classification model that we created uses the classic TF-IDF feature, with a logistic regression classifier. This was selected due to high performance reported in the study by Mohammed & Omar (2020). One of the challenges that we faced was that some of the finer details of the specific algorithm implemented by Mohammed & Omar (2020) were not all explained clearly (such as exclusion of "stop words"), so although the algorithm is very similar, there are slight differences which may account for differences in performance.

Testing of algorithm, and selection of datasets for training algorithm

Table 2: Existing datasets of exam questions tagged with the appropriate Bloom's Taxonomy level, number of questions tagged to each level

Number of Questions Tagged to Level in Bloom's Taxonomy	**Gani et.al. (2023)**	**Mohammed & Omar (2020)**	**Sangodiah et al. (2017)**	**Yaha et al. (2012)**
Knowledge	149	22	50	100
Comprehension	669	20	135	100
Application	100	14	72	100
Analysis	99	19	56	100
Evaluation	107	21	57	100
Synthesis	76	21	45	100
Total Number of Questions	1200	127	415	600

Table 3: Existing datasets of exam questions tagged with the appropriate Bloom's Taxonomy level, by topic area, and f1-score across 15 runs

Dataset sourced from	**Question types**	**Broad Topic Areas**	**Average f1-score Across 15 Runs**
Gani et.al. (2023)	1200 questions	Wide variety	0.80
Mohammed & Omar (2020)	127 open-ended questions	Wide variety	0.65
Sanders et al. (2013)	654 multiple-choice questions	Computing education CS1 and CS2 topics	N/A – future work
Sangodiah et al. (2017)	415 open-ended questions	Wide variety	0.65
Yaha et al. (2012)	600 open-ended questions	Wide variety	0.74

We subsequently investigated the accuracy of the algorithm by checking the performance against existing larger datasets of exam questions that had already been expertly tagged with the appropriate Bloom's Taxonomy level

(Table 2). The datasets cover a range of different topic areas (Table 3). We independently tested the algorithm separately 15 times with each dataset, then averaged the results across the 15 runs, to evaluate the accuracy of the algorithm. We recorded the weighted average for "accuracy", "precision", "recall" and "f1-score" across each run. For each run, we selected that 90% was used for training the model, and 10% was used to test the dataset; this is a widely used standard (e.g. Gani et.al., 2023).

Table 3 shows the average f1-score across the various datasets ranged from 0.65 to 0.80. The f1-score for the Mohammed & Omar (2020) and Sangodiah et al. (2017) datasets was possibly lower due the lower number of questions, and because the number of questions was not evenly distributed across the different levels of Bloom's Taxonomy. The number of questions in the dataset by Yaha et al. (2012) is evenly distributed, which means that the model is trained more effectively to accurately identify questions at all levels of Bloom's Taxonomy. The f1-score of 0.74 for the dataset from Yaha et al. (2012) was lower than the f1-scores of between 0.85 and 0.89 reported by Mohammed & Omar (2020) (who used the same dataset, and whose algorithm ours is based upon). However, Mohammed & Omar (2020) used 100% of questions in training their model, and 10% were used in testing. This means that the same questions were used to both train and test their model, which likely increased accuracy. We also tested training the model on the Yaha et al. (2012) dataset, and the testing this on the independent Mohammed & Omar (2020) dataset, which had an f1-score of 0.68.

The implications are that the Yaha et al. (2012) dataset is the preferred dataset for training the classification model, and that it is likely to be reasonably accurate (at least relatively accurate in comparison to the findings of recent research on this topic area) at correctly identifying questions from other datasets. This means that the dataset is ideal to train the classification model used in the software application, and that it is likely to be reasonably accurate when used to evaluate exam questions written by educators from various disciplines.

Providing educators with feedback on their exam questions, and overall exam

Following this, we will evaluate effective means for providing the educator (the user) with meaningful and actionable feedback based on the results of the question classifications according to Bloom's Taxonomy. Rather than (only) providing the user with raw information about the Bloom's Taxonomy classification for every question, we will provide feedback about the complexity of the overall exam. This will be done by considering the complexity of each of the exam questions, the type of questions, and the Bloom's Taxonomy classification, and comparing this against the number of marks which are allocated to the question in the exam. For example, if there are a large number of questions which have high complexity but are only worth a small percentage of the overall marks, this can be highlighted to the educator for them to reflect on, and consider whether the wording of the questions or the question marks may need to be adjusted. Likewise, if the complexity of the overall exam is quite high, the system will provide feedback to the educator that it may be challenging for students to complete the exam during the allocated time, and the educator may need to reflect on whether the number of questions should be adjusted. The style and wording of feedback messages will be created in conjunction with an educational expert, to ensure that the messages will encourage reflective thinking by the educators.

User testing by educators

Once the final system is completely implemented, we will engage teaching staff with exams from The University of Melbourne during the second half of 2023 to conduct extensive user testing, that will be used to refine the system. We intend to conduct further research to evaluate how the software tool can assist university educators with writing high quality exam questions.

Significance & Expected Outcomes

Many university educators are unable to accurately evaluate the complexity of exam questions based on Bloom's Taxonomy (Karpen & Welch, 2016; Watering & van der Rijt 2006), and often overestimate student performance (Urhahne, &Wijnia, 2021). This leads to creating exams that are often either too complex or not complex enough. Educators need ways of receiving additional support to make sure that they are the appropriate level of complexity as a pedagogical tool for to appropriately evaluating students' learning during the subject. We expect that educators will derive great benefits from being able to use the software tool to aide with the creation of examinations, because it will provide an external method for receiving feedback regarding the complexity of their exams, including specific areas within the exam which may require additional attention or reflection. The software tool will serve as an objective measure to ensure the exams are appropriately aligned with the desired learning outcomes. Educators will receive valuable feedback regarding the cognitive demands of the questions, enabling them to make informed adjustments and enhance the overall effectiveness of the assessment process. The software tool will also serve as a professional development resource for educators, helping them improve their understanding of Bloom's Taxonomy and its application in designing exams.

References

Aninditya, A., Hasibuan, M. A., & Sutoyo, E. (2019). Text mining approach using TF-IDF and naive Bayes for classification of exam questions based on cognitive level of bloom's taxonomy. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)* (pp. 112-117). IEEE. https://doi.org/10.1109/IoTaIS47347.2019.8980428

Bengio, Y., & Senécal, J. S. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. IEEE Transactions on Neural Networks, 19(4), 713-722. https://doi.org/10.1109/TNN.2007.912312

Das, S., Das Mandal, S. K., & Basu, A. (2022). Classification of action verbs of Bloom's taxonomy cognitive domain: An empirical study. Journal of Education, 202(4), 554-566. https://doi.org/10.1177/00220574211002199

Gani, M. O., Ayyasamy, R. K., Sangodiah, A., & Fui, Y. T. (2023). Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. Education and Information Technologies, 1-22. https://doi.org/10.1007/s10639-023-11842-1 (early online access)

Karpen, S. C., & Welch, A. C. (2016). Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. Currents in Pharmacy Teaching and Learning, 8(6), 885-888. https://doi.org/10.1016/j.cptl.2016.08.003

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. Theory into practice, 41(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Liu, Q., Wald, N., Daskon, C., & Harland, T. (2023). Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers. Innovations in Education and Teaching International, 1-13. https://doi.org/10.1080/14703297.2023.2222715 (early online access)

Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. PloS one, 15(3), e0230442. https://doi.org/10.1371/journal.pone.0230442

Sanders, K., Ahmadzadeh, M., Clear, T., Edwards, S. H., Goldweber, M., Johnson, C., ... & Spacco, J. (2013, June). The Canterbury QuestionBank: Building a repository of multiple-choice CS1 and CS2 questions. In Proceedings of the ITiCSE working group reports conference on Innovation and technology in computer science education-working group reports (pp. 33-52). https://doi.org/10.1145/2543882.2543885

Shaikh, S., Daudpotta, S. M., & Imran, A. S. (2021). Bloom's learning outcomes' automatic classification using lstm and pretrained word embeddings. IEEE Access, 9, 117887-117909. https://doi.org/10.1109/ACCESS.2021.3106443

Sangodiah, A., Ahmad, R., & Wan Ahmad, W. F. (2017). Taxonomy Based Features in Question Classification Using Support Vector Machine. Journal of Theoretical & Applied Information Technology, 95(12). http://www.jatit.org/volumes/Vol95No12/22Vol95No12.pdf

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. Educational Research Review, 32, 100374. https://doi.org/10.1016/j.edurev.2020.100374

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. Educational Research Review, 1(2), 133-147. https://doi.org/10.1016/j.edurev.2006.05.001

Yahya, A. A., Toukal, Z., & Osman, A. (2012). Bloom's taxonomy–based classification for item bank questions using support vector machines. In Modern advances in intelligent systems and tools (pp. 135-140). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30732-4_17

Zhang, J., Wong, C., Giacaman, N., & Luxton-Reilly, A. (2021, February). Automated classification of computing education questions using Bloom's taxonomy. In Proceedings of the 23rd Australasian Computing Education Conference (pp. 58-65). https://doi.org/10.1145/3441636.3442305

Valentine, A., Oliveira, E. A. (2023). Paper Name. In Editors names (Eds.), *People, partnerships and pedagogies*. Proceedings ASCILITE 2023. Christchurch (pp. xxx–xxx). DOI: xxx

Note: All published papers are refereed, having undergone a double-blind peer-review process.