An NLP framework to automatically evaluate the adequacy and relevance of assessment items

Luca Benedetto

November 2025

This project aims at extending previous research on Question Difficulty Estimation from Text and, more broadly, on the evaluation of learning and assessment content [4, 6]. More specifically, it has the objective of developing a framework that leverages Natural Language Processing (NLP) techniques to assess newly created exams and questions, and to evaluate the relevance of such questions to a given learning path (e.g., the content of the lectures that students attend before taking the exam). This framework has potential applications in both traditional and non-traditional settings, and in both high-stakes and low-stakes contexts, by providing professors, educators, and exam curators with an additional evaluation of the assessment items before using them to score students. Depending on the specific application scenario, the framework could be used to automatically filter out potentially problematic items (e.g. in low-stakes settings) or to flag them for further human inspection.

Question Difficulty Estimation from Text (QDET), and question evaluation from text in general, are research areas that have gained increased interest in recent years, primarily following the advent of the Transformer architecture in 2018 and more recently of Large Language Models – [10, 9, 7, 11, 17], inter alia. Their ultimate goal is to overcome (or at least reduce) the need for pretesting exam items [8]: indeed, before being used to assess students, new items need to be evaluated for validity and quality, which is traditionally done by deploying them in exams and studying the response patterns of students. Although this leads to accurate question validation, it is very time-consuming, expensive, and cannot be done in all educational contexts (e.g., in traditional on-site settings with small student cohorts). QDET aims at overcoming this by developing approaches to automatically predict question difficulty from text using NLP, either by training supervised models to predict difficulty from text [4, 2], or by using (large) language models to simulate student responses [14, 12, 3]. QDET is the most popular task in the content evaluation community, but previous literature also explored techniques to evaluate distractors [1, 5, 16], learning paths [15, 13], and other aspects of educational content.

Crucially, the majority of previous research has focused on exam items only, without taking into consideration their relationship with the associated courses or lectures. Conversely, in this project, we will work towards filling the gap, by

evaluating questions and assessment items *in relation to course content*. This project will use an ensemble of traditional machine learning and Information Retrieval (IR) techniques, as well as more recent semantic embeddings and neural networks (including LLMs, prioritising open-weight models), to reach its primary objective: the development of a prototypical framework (implemented in Python) to evaluate exam items and their relevance to the related learning content. This framework will be structured to receive as input the content (lectures, exam questions, and possibly additional material) and provide an interpretable (and explainable) report about their validity, relevance, and difficulty. Once validated, such system has the potential to enable scalable, data-driven evaluation of exam quality and alignment with learning objectives: it will provide human experts with an additional quality control of assessment items, thus further supporting the development of fair and safe educational assessment.

References

[1] Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. Distractor Generation in Multiple-Choice Tasks: A Survey of Methods, Datasets, and Evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14437–14458, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[2] Samah AlKhuzaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. A Systematic Review of Data-Driven Approaches to Item Difficulty Prediction. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, volume 12748, pages 29–41. Springer International Publishing, Cham, 2021.

[3] Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. Using llms to simulate students' responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368, 2024.

[4] Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37, 2023.

[5] Luca Benedetto, Shiva Taslimipoor, and Paula Buttery. A survey on automated distractor evaluation in multiple-choice tasks. In *Workshop on Building Educational Applications with NLP*, 2025.

[6] Luca Benedetto, Shiva Taslimipoor, Andrew Caines, Diana Galvan-Sosa, George Dueñas, Anastassia Loukina, and Torsten Zesch. Workshop on automatic evaluation of learning and assessment content. In *International*

Conference on Artificial Intelligence in Education, pages 473–477. Springer, 2024.

[7] Wanyong Feng, Peter Tran, Stephen Sireci, and Andrew S. Lan. Reasoning and Sampling-Augmented MCQ Difficulty Prediction via LLMs. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, *Artificial Intelligence in Education*, volume 15880, pages 31–45. Springer Nature Switzerland, Cham, 2025.

[8] Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy, 2019. Association for Computational Linguistics.

[9] Ming Li, Hong Jiao, Tianyi Zhou, Nan Zhang, Sydney Peters, and Robert W. Lissitz. Item Difficulty Modeling Using Fine-tuned Small and Large Language Models. *Educational and Psychological Measurement*, page 00131644251344973, July 2025.

[10] Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. Jump-Starting Item Parameters for Adaptive Language Tests. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Steven Moore, Huy A. Nguyen, Tianying Chen, and John Stamper. Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods, July 2023.

[12] Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. Large Language Models are Students at Various Levels: Zero-shot Question Difficulty Estimation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] Nur W Rahayu, Ridi Ferdiana, and Sri S Kusumawardani. A systematic review of learning path recommender systems. *Education and Information Technologies*, 28(6):7437–7460, 2023.

[14] Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. Question Difficulty Prediction Based on Virtual Test-Takers and Item Response Theory.

[15] Fan Yang and Zhenghong Dong. Learning path construction in e-learning. *Lecture Notes in Educational Technology. Springer, Heidelberg*, 3, 2017.

[16] Chak Yan Yeung, John Lee, and Benjamin Tsou. Difficulty-aware Distractor Generation for Gap-Fill Items. page 6, 2019.

[17] Leonidas Zotos, Ivo Pascal de Jong, Matias Valdenegro-Toro, Andreea Ioana Sburlea, Malvina Nissim, and Hedderik van Rijn. NLP Methods May Actually Be Better Than Professors at Estimating Question Difficulty, 2025.