Benchmarking Item Difficulty Classification in German Vocational Education and Training

Alonso Palomino **Benjamin Paaßen**
German Research Center for Artificial Intelligence (DFKI), Germany
Bielefeld University, Germany
first.last@dfki.de, first.last@techfak.uni-bielefeld.de

Abstract

Predicting the difficulty of exam questions or items is essential to effectively assembling and calibrating exams. While item response theory (IRT) models can estimate item difficulty, they require student responses that are costly and rarely available at scale. Natural language processing methods offer a text-only alternative; however, due to the scarcity of real-world labeled data, prior work often relies on synthetic or domain-specific corpora, limiting generalizability and overlooking the nuanced challenges of real-world text-based item difficulty estimation. Addressing this gap, we benchmark 122 classifiers on 935 German Vocational Education and Training (VET) items labeled via previous IRT analysis to assess feasibility under real-world conditions. In our setup, a stacked ensemble that combines linguistic features, pre-trained embeddings, and external semantic resources outperforms both transformer-based models and few-shot large language models, achieving moderate performance. We report findings and discuss limitations in the context of German VET.

1 Introduction

In psychometrics, pretesting is the standard practice for item difficulty estimation: recruiting a representative student sample, recording their item responses, and fitting an IRT model to obtain precise difficulty estimates (Lord and Novick, 1968; Baker, 2001). Gathering representative high-stakes student response data is complex and seldom feasible. High financial and administrative costs for sourcing a domain-specific, diverse examinee pool, challenges embedding items in live exams, and limited pretesting capacity all contribute to its scarcity.

Educators gauge item difficulty via expert review of phrasing combined with their assessment experience. For example, one may consider the lexical, syntactic, and semantic attributes of the item's text (Dale and Chall, 1949; Martinc et al., 2021). Following this difficulty estimation approach, natural language processing models have recently been applied to estimate item difficulty, e.g., via text regression or classification techniques (AlKhuzaey et al., 2023; Benedetto et al., 2023).

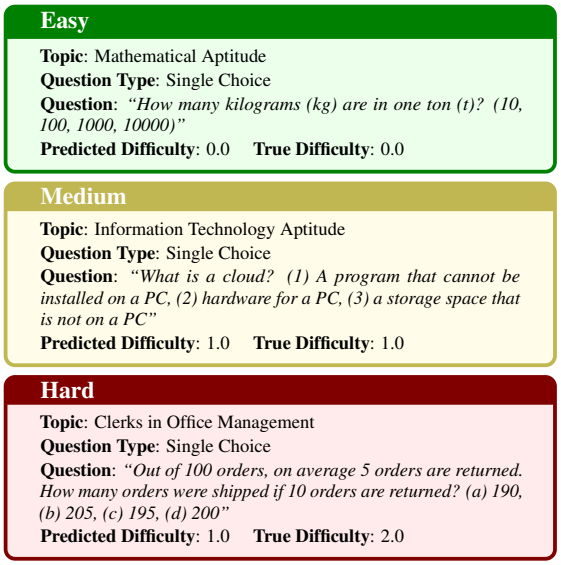

Figure 1: On bfz's VET data, the ensemble classified easy and medium correctly but overestimated hard.

In education, estimating item difficulty is key for comprehensive exam assembly and calibration (Palomino et al., 2024, 2025). Prior research relies on domain-specific or synthetic student response datasets in English; therefore, difficulty estimation in German Vocational Education and Training (VET) remains unexplored, and evidence on existing methods' performance on real data is lacking. To address this gap, we ask: **RQ:** *How accurately do popular text classifiers estimate the difficulty of German VET test items?* Partnering with bfz[1], Germany's largest VET services provider, we evaluated 122 text-classification methods on an industry-

[1] https://www.bfz.de/

Proceedings of Recent Advances in Natural Language Processing,pages 870–875
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-099

grade VET dataset of 935 IRT-labeled items and found that items requiring ambiguity resolution, technical or domain-specific knowledge or visual inference were most often misclassified, underlining the limits of current approaches[2]. Among the tested approaches, a stacked ensemble achieved the strongest production-level performance (balanced accuracy 0.60). Figure 1 compares predicted and true difficulty, showing correct easy and medium classifications but overestimation of hard items. Sections 2-5 outline the literature, dataset, methods, results, and conclusions.

2 Related Work

Psychometrics traditionally estimates item difficulty via Classical Test Theory (CTT) and Item Response Theory (IRT), where CTT defines difficulty as the proportion of correct responses (0-1) and IRT fits latent-trait models, both requiring pretesting data (Lord and Novick, 1968; Baker, 2001; Embretson and Reise, 2013). By contrast, NLP methods predict difficulty directly from text: Loukina et al. (2016) used a random forest on text-complexity features for listening items, Huang et al. (2017) applied a CNN to sentence-level reading items, Ha et al. (2019) combined ELMo, Word2Vec and retrievability features for medical exams, and Xue et al. (2020) demonstrated that transfer learning from completion-time estimation could improve estimation. Yaneva et al. (2021) clustered linguistic features to split 18,961 medical exam items into low/high response-process complexity. Benedetto et al. (2021) compared term-frequency, BERT/DistilBERT, and IRT models on proprietary and ASSISTments data, finding transformers with additional text features most effective. Byrd and Srivastava (2022) simulated pretesting with QA ensembles to infer IRT scores on HotpotQA. Park et al. (2024) proposed a zero-shot, resource-intensive framework generating synthetic responses via LLM clusters and aggregating their outputs. Surveys by Benedetto et al. (2023) and AlKhuzaey et al. (2023), covering 122 and 88 studies, highlighted dataset scarcity due to confidentiality and concluded that semantic/syntactic features remain the strongest predictors. Yaneva et al. (2024) organized a medical exam shared task with 17 teams; framed as regression, models showed minimal improvements over baselines. Zotos et al. (2025) found LLM uncertainty correlates with difficulty on 451 Biopsychology items, though generalizability is unclear.

[2]Research artifacts available at: https://git.hub/kipwb

3 Text-based Item Difficulty Estimation

3.1 Dataset

We used data from bfz's assessment platform, where item difficulties (-3.0 to +3.0) were estimated with a 1-parameter IRT model on 935 items from a vocational cohort (Paaßen et al., 2022b,a). For exam assembly and interpretability, scores were discretized into easy, medium, and hard classes to reduce cognitive load (Schwarz, 2007; Weijters et al., 2010) and mitigate noise and bias (Caruana and Niculescu-Mizil, 2006; Ribeiro et al., 2016; Furnham and Boo, 2011). Though granularity is lost, discrete classes are directly usable by test designers and support categorical constraints. Thresholds were set empirically via quantile inspection and stability checks to preserve ordinal structure while limiting noise. Following Yaneva et al. (2024), we aimed for balanced splits, but real-world class imbalance posed challenges. Despite testing oversampling, class weighting, and alternative binning, we retained the configuration that best preserved item characteristics: 19% easy, 75% medium, 5% hard; 71% single-choice, 20% multiple-choice, 8% matrix; and topic distributions of 30% IT, 16% German, and others <1%. To safeguard privacy and intellectual property, raw data remain unavailable, but anonymized embeddings are shared.

3.2 Task and Experimental Setup

We model difficulty as a three-class task (0 easy, 1 medium, 2 hard) using text and meta-attributes (topic, type, skills). Stratified 10-fold cross-validation preserves distributions. We report balanced accuracy, macro F1, and weighted precision/recall; models are retrained on the full set and evaluated on a stratified hold-out. Labels are 19% easy, 75% medium, 5% hard; formats 71% single-choice, 20% multiple-choice, 8% matrix. Table 2 lists the five most frequent topics; Figure 2b shows IRT scores; Table 1 details the split.

3.3 Models

Classic: We trained baseline classifiers (uniform random and majority-class) and supervised algorithms k-nearest neighbors, decision trees, random forests, support vector machines, logistic regression variants, and neural networks using `scikit-learn` and Keras (Pedregosa et al., 2011;

(a)

Difficulty Level	Count
Easy	139
Medium	566
Hard	43
Total	748
Training	

(b)

Difficulty Level	Count
Easy	40
Medium	139
Hard	8
Total	187
Testing	

Table 1: The class difficulty distribution for training and testing data folds.

(a)

Topic	Count
Information technology aptitude	272
German language competence	150
Warehousing & logistics	78
Professional Advice & self-assessment	69
Clerks in office management	59
Total Items	935

(b)

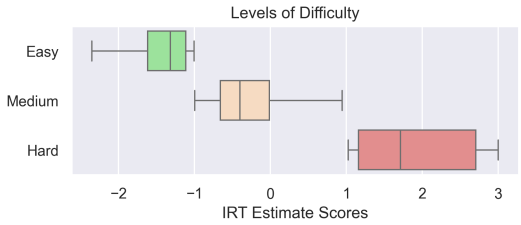

Table 2: (a) shows the top-5 topic distribution, while (b) displays the summary of IRT scores.

Chollet et al., 2015).

Ensemble: We used stacked generalization (Wolpert, 1992; Pedregosa et al., 2011), combining balanced linear (PA, logistic regression, SVM, perceptron) and tree-based models (randomized trees, AdaBoost). Their outputs fed a dense meta-network (128-64-32-10-3) to predict difficulty (Chollet et al., 2015).

DistilBERT and FSL: We fine-tuned and evaluated existing BERT models for text complexity estimation, a semantically similar task (Face, 2024; Anschütz and Groh, 2022; Ruben Klepp, 2022). Additionally, we evaluated few-shot learning (FSL) via SetFit (Tunstall et al., 2022).

LLMs: We built a German few-shot pipeline with GPT-4, GPT-4o, and o1-preview (Achiam et al., 2023; Hurst et al., 2024; Jaech et al., 2024). Prompts defined labels (0-2) with up to 10 stratified exemplars; for tests, the 10 most similar items (cosine similarity) were retrieved. Parameters were fixed for determinism, and three runs per item were aggregated by majority voting using the `Instructor` library (Liu and Contributors, 2024).

Linguistic Features Each item's stem, answers, distractors, topics, and skills were encoded with embeddings and features:

1. **Stylometry and Format:** We extracted sentence count and punctuation frequencies; noun-to-word ratio; character distribution and average characters per word; the Coleman-Liau readability index (Coleman and Liau, 1975); and encoded question types (single choice, multiple choice, matrix) as one-hot vectors.
2. **LIWC and Text Complexity:** We extracted 39 German LIWC dimensions with high correlation to difficulty labels, including analytical thinking, advanced vocabulary, quantifiers, insight, certainty, negative emotion, discrepancy and tone measures (Boyd et al., 2022). We reduced features to two components via PCA and used the Anschütz and Groh (2022) text complexity model as an auxiliary difficulty signal.
3. **Embeddings and Visual Content:** We extracted embeddings from ten pretrained models, including the top five from the MTEB text classification leaderboard (other languages) (Muennighoff et al., 2022) and five trending German similarity models from HuggingFace[3]. Encoding only the stem and answer improved classification. Among these, embeddings (8), (4), and (10) yielded the best results. Visual content was captured via one-hot encoding of categorical metadata.
4. **Topics and Skills:** After comparing several pretrained models, we found that the model of Rogge (2024) best captured topics and skill themes for difficulty prediction; we therefore use it to embed those attributes.
5. **Semantic Relations and Negation Cues:** We applied cosine similarity matching with the ODENet lexical base (Siegel and Bond, 2021) to identify hypernyms and hyponyms and used a manually curated negation cue list to count explicit negation terms.
6. **Semantic Similarities and Named Entities:** We computed cosine similarities between each item's stem and its answers and distractors and used the NER model of Schiesser (2024) to

[3] https://huggingface.co/spaces/mteb/leaderboard

Model Type	Classifier	Embedding Version	Features	Acc (Train)	Acc (Test)	F1 (Train)	F1 (Test)
Dummy	Stratified	N/A	N/A	0.32	0.34	0.61	0.59
	Most Freq.			0.33	0.33	0.65	0.63
Classic	SVM (modified huber, l1)	severinsimmler/xlm-roberta-longformer-base-16384	stem + answer + stylometry + topics + complexity + hypernyms + skills	0.52	0.58	0.65	0.62
	Logistic Regression (log, l1)	intfloat/multilingual-e5-base	stem + answer + stylometry + topics + complexity + hypernyms + skills + negation cues	0.51	0.56	0.59	0.55
	Logistic Regression (log, l1)	severinsimmler/xlm-roberta-longformer-base-16384	stem + answer + stylometry + topics + complexity + hypernyms + skills + negation cues + PCA LIWC	0.52	0.54	0.63	0.61
	Logistic Regression (log, l1)	danielheinz/e5-base-sts-en-de	stem + answer + stylometry + topics + complexity + hypernyms	0.49	0.54	0.59	0.59
	Logistic Regression (log, l1)	danielheinz/e5-base-sts-en-de	stem + answer + stylometry + topics + format/type + LIWC + complexity	0.50	0.53	0.59	0.57
Ensemble	Stacking	intfloat/multilingual-e5-base	stem + answer + stylometry + format/type + topics + complexity + hypernyms + skills + PCA LIWC	0.54	0.60	0.65	0.73
Transformer	DistilBERT	distilbert-base-german-cased	N/A	0.58	0.56	0.68	0.73
	DistilBERT	MiriUll/distilbert-german-text-complexity	N/A	0.56	0.54	0.66	0.72
	DistilBERT	krupper/text-complexity-classification	N/A	0.42	0.48	0.67	0.70
FSL	SetFit	MiriUll/distilbert-german-text-complexity	N/A	0.48	0.48	0.53	0.48
LLM	GPT-4	N/A	N/A	0.37	0.38	0.47	0.49
	GPT-4o			0.41	0.34	0.41	0.43
	o1-preview			0.33	0.33	0.65	0.63

Table 3: The top-performing balanced accuracy and weighted F1 metrics of the evaluated item difficulty classifiers.

count location, organization and person entity frequencies.

4 Results

Addressing RQ, Table 3 reports balanced accuracy and weighted F1 for the best models among 122 configurations. Combining stem and answer embeddings with stylometric, semantic, topic, and skill features yielded the strongest results. A stacked ensemble of balanced linear and tree-based learners achieved 0.60 balanced accuracy and 0.73 weighted F1. DistilBERT variants followed (0.58, 0.56), while logistic regression and SVMs gave moderate scores. LLMs (GPT-4o, o1-preview) performed worst, with several complex models overfitting due to data scarcity and imbalance, underscoring the trade-off between balanced accuracy and weighted F1 in practice. Examination of ensemble predictions highlights key limitations. As Table 3 shows, easy items reached an F1 of 0.55 (moderate), medium items 0.81 across 139 instances (strong), and hard items only 0.24 on eight instances (weak). By topic, "content creation," "basic technology use," and "German cultural competence" achieved perfect balanced accuracy (1.0), while "women in the workforce," "warehousing and logistics," and "IT aptitude" averaged 0.22. Misclassifications often involved items requiring contextual or ambiguous reasoning, complex instructions, technical/domain expertise, or visual/geopolitical inference (e.g., cargo securing, Schengen Agreement impacts).

5 Conclusions

Partnering with bfz, a major VET provider, we benchmarked 122 models for item difficulty classification on real German VET data, unlike prior work on medical or synthetic English datasets. The best model was a stacked ensemble combining stem and answer embeddings, stylometric features, and semantic metadata, achieving 0.60 balanced accuracy and 0.73 weighted F1, outperforming fine-tuned transformers and LLM prompting. While first framed as a standard text classification task, our results highlight its complexity and the importance of non-linguistic features. Future work will apply active learning for data acquisition and fine-tune larger encoders.

Acknowledgements

This work was supported by the KI-Akademie OWL project, funded by the Federal Ministry of Research, Technology and Space (BMFTR) and administered by the Project Management Agency of the German Aerospace Centre (DLR) under grant no. 01IS24057A.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2023. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, pages 1–53.

Miriam Anschütz and Georg Groh. 2022. TUM social computing at GermEval 2022: Towards the significance of text statistics and neural embeddings in text complexity prediction. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 21–26, Potsdam, Germany. Association for Computational Linguistics.

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–157, Online. Association for Computational Linguistics.

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

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130, Dublin, Ireland. Association for Computational Linguistics.

Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

François Chollet et al. 2015. Keras. https://keras.io.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Hugging Face. 2024. distilbert/distilbert-base-german-cased. https://huggingface.co/distilbert/distilbert-base-german-cased. Accessed: 2024-06-30.

Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. 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. Association for Computational Linguistics.

Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1352–1359. AAAI Press.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jason Liu and Contributors. 2024. Instructor: A library for structured outputs from large language models.

F. M. Lord and M. R. Novick. 1968. *Statistical theories of mental test scores*. Addison-Wesley.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253, Osaka, Japan. The COLING 2016 Organizing Committee.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Benjamin Paaßen, Malwina Dywel, Melanie Fleckenstein, and Niels Pinkwart. 2022a. Interpretable knowledge gain prediction for vocational preparatory e-learnings. In *International Conference on Artificial Intelligence in Education*, pages 132–137. Springer.

Benjamin Paaßen, Christina Göpfert, and Niels Pinkwart. 2022b. Faster confidence intervals for item response theory via an approximate likelihood. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*.

Alonso Palomino, Andreas Fischer, David Buschhüter, Roland Roller, Niels Pinkwart, and Benjamin Paassen. 2025. Mitigating bias in item retrieval for enhancing exam assembly in vocational education services. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 183–193, Albuquerque, New Mexico. Association for Computational Linguistics.

Alonso Palomino, Andreas Fischer, Jakub Kuzilek, Jarek Nitsch, Niels Pinkwart, and Benjamin Paassen. 2024. EdTec-QBuilder: A semantic retrieval tool for assembling vocational training exams in German language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 26–35, Mexico City, Mexico. Association for Computational Linguistics.

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177, Miami, Florida, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Niels Rogge. 2024. Lilt + xlm-roberta-base. `https://huggingface.co/nielsr/lilt-xlm-roberta-base`.

Ruben Klepp. 2022. text-complexity-classification (revision 47a5713).

Markus Schiesser. 2024. mschiesser/ner-bert-german. `https://huggingface.co/mschiesser/ner-bert-german`. Accessed: 2024-06-28.

Norbert Schwarz. 2007. Cognitive aspects of survey methodology. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 21(2):277–287.

Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a GermanWordNet from other resources. In *Proceedings of the 11th Global Wordnet Conference*, pages 192–198, University of South Africa (UNISA). Global Wordnet Association.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Bert Weijters, Elke Cabooter, and Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, Seattle, WA, USA → Online. Association for Computational Linguistics.

Victoria Yaneva, Daniel Jurich, Le An Ha, and Peter Baldwin. 2021. Using linguistic features to predict the response process complexity associated with answering clinical MCQs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–232, Online. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2025. Can model uncertainty function as a proxy for multiple-choice question item difficulty? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316, Abu Dhabi, UAE. Association for Computational Linguistics.