Predicting Item Difficulty in German Language Tests: A Machine Learning Approach with Linguistic Features

Ivana Kapounová and Jana Dlouhá

Department of Psychology, Faculty of Arts, Charles University
Celetná 20, 116 42 Praha 1
ivana.kapounova@ff.cuni.cz

Abstract

Item difficulty is central to test validity and fair assessment, yet it is difficult to estimate without large-scale pretesting. Machine learning offers a promising alternative by enabling prediction of item difficulty directly from the text prior to administration. This study replicates and extends previous research on predicting item difficulty in language comprehension tests using machine learning algorithms and linguistic features. Specifically, a modelling approach previously developed for English maturita exams was applied to German language test items. Item difficulty parameters were estimated using Item Response Theory, and 696 features covering lexical, frequency-based, surface similarity, and readability properties were extracted from the texts. Regularized linear regression models were trained and evaluated using nested cross-validation. Similar to results obtained in the English dataset, the best-performing model (LASSO) achieved a mean RMSE close to the standard deviation of item difficulty and therefore comparable to predicting the mean. This lack of predictive power was further compounded by marked instability across cross-validation folds. Directly applying coefficients trained on English items resulted in substantially larger prediction errors. The results are attributed primarily to the small number of items relative to the high-dimensional feature space. These findings suggest that feature-based machine learning approaches offer limited added value in such settings and are better treated as a complement to, rather than a replacement for, traditional difficulty estimation methods.

1 Introduction

Accurate estimation of item difficulty is central to educational measurement, supporting test validity, score interpretation, and comparability across test forms. It is also essential in modern applications such as computerized adaptive testing, where items must be selected dynamically to match examinee ability and reduce test length. Scalable estimation further enables the construction of balanced tests, detection of potential biases, and more informed decision-making about examinees.

1.1 Traditional Difficulty Estimation

Traditionally, difficulty is established via pretesting and domain expert judgement. Although pretesting yields statistically robust estimates (Fadillah et al., 2023), it is costly, time-consuming, and delays item deployment. Expert judgement, while useful in early stages, is subjective and potentially inconsistent across raters (Wonde et al., 2024).

Yet item difficulty is not arbitrary. It is partly encoded in an item's structural and linguistic properties. Advances in machine learning and natural language processing now make it possible to infer difficulty directly from the text. This shift opens the possibility of predicting item difficulty prior to administration in a scalable and systematic manner, complementing traditional approaches.

1.2 Evolution of Automated Prediction

Early text-based prediction approaches relied on manually engineered surface-level and linguistic features (Loukina et al., 2016), such as sentence length, vocabulary rarity, and syntactic complexity. These features were then used as inputs to classical machine learning models (e.g., linear regression, random forests).

Later developments introduced static word embeddings (e.g., *Word2Vec*, *GloVe*) to capture semantic similarity (Hsu et al., 2018), followed by deep learning architectures such as convolutional neural networks and long short-term memory networks (He et al., 2021; Huang et al., 2017; Xue et al., 2020), which enabled more effective modelling of contextual relationships in text.

1.3 Recent Approaches

Recent advances have shifted the field decisively toward transformer-based architectures and large language models (LLMs), giving rise to three principal paradigms.
First, the most direct approach is to prompt an LLM to predict item difficulty from text (Gombert et al., 2024; Li, Jiao, et al., 2025). While intuitive, this approach typically achieves only moderate accuracy compared to more structured methods (Razavi & Powers, 2025).
Second, LLMs are used as feature extractors, generating cognitively and linguistically informed variables (e.g., reasoning steps or cognitive load) for downstream models (Razavi & Powers, 2025). This approach has been shown to outperform both direct LLM predictions and traditional text-based features (Mojoyinola et al., 2025).
Third, LLMs are used to simulate student responses across proficiency levels (Maeda, 2025; Scarlatos et al., 2025). Item difficulty is subsequently estimated via IRT models fitted to the synthetic data. A key limitation is that LLMs struggle to reflect human capability constraints, and highly capable models tend to overproduce correct responses even when conditioned on lower proficiency (Acquaye et al., 2026; Li, Chen, et al., 2025; Liu et al., 2025).
Beyond these paradigms, LLMs are utilized for data augmentation to enrich training sets with synthetic items and rationales (Feng et al., 2025; Li, Jiao, et al., 2025), as well as for uncertainty estimation, where the model's internal confidence serves as a proxy for predicting human task difficulty (Zotos et al., 2025). A comprehensive overview of these developments is provided by Peters et al. (2025).

1.4 Limitations and Open Challenges

Despite rapid progress in advanced architectures, several important challenges remain unresolved. Among these, interpretability and cross-linguistic generalizability are particularly relevant to this study.
Automated item difficulty prediction involves a persistent trade-off between predictive accuracy and interpretability. Recent LLM-based approaches often achieve superior performance but rely on high-dimensional, opaque representations, limiting transparency. Although post hoc explainable AI methods have been proposed, they offer only partial insight into model decision-making (Dunn et al., 2026). In high-stakes assessment, where transparency is vital for validity and fairness, traditional feature-based models may be therefore preferable. Defined linguistic features allow these approaches to pinpoint difficulty, offering item writers actionable guidance for precise calibration. They are also more robust in small-sample settings, which are typical in high-stakes testing (Peters et al., 2025).

A second open question concerns the cross-linguistic transferability of predictive models. Current evidence in this area is predominantly indirect. For example, recent research suggests that training models jointly on multiple languages can substantially improve cross-lingual performance as they learn global difficulty patterns (Das et al., 2024; Skidmore et al., 2025). For feature-based machine learning models, however, studies explicitly examining transferability are very rare.

1.5 Research Objectives

Two prior studies established a framework for predicting item difficulty from linguistic features using machine learning (Dlouhá, 2025; Štěpánek et al., 2023) but were limited to English. This study aims to test the framework's cross-linguistic generalizability by replicating it on German maturita reading items and to compare these results with the original findings.

2 Methods

2.1 Data

Data were obtained from *CERMAT* German maturita tests and response data (spring terms 2016–2024, autumn terms were excluded due to different examinee populations). Only reading-comprehension sections (Parts 5 and 7) were analysed, yielding 45 items per part. A German frequency resource *DeReWo-2014-II* (Institut für Deutsche Sprache, 2014) was used for lexical frequency features.

2.2 Procedure

All analyses were conducted in R (R Core Team, 2023).
The replication part followed a three-stage pipeline:

Item difficulty was first estimated via Rasch models (package *eRm*; Mair & Hatzinger, 2007). Responses were analysed per test form without linking.
Items were then segmented into their components (passage, question, key, distractors) from PDFs. A total of 696 base features (including lexical, frequency-based, surface similarity and readability indices) were computed across these components using the *quanteda* (Benoit et al., 2018) and *udpipe* (Wijffels, 2023) packages. Feature–difficulty correlations were modest ($|r| \leq .40$), and substantial multicollinearity combined with a small sample size motivated the use of regularization.
Separate models were subsequently trained for Parts 5 and 7 using ridge, LASSO, and elastic net regression. For that purpose, packages *glmnet* (Friedman et al., 2010) and *caret* (Kuhn, 2008) were used. Nested cross-validation (9

inner and 9 outer folds) served both for unbiased performance estimation and hyperparameter tuning. Before training the model, within each inner loop, the top 100 predictors (by correlation with difficulty) were selected to reduce dimensionality. Performance was assessed via mean RMSE and R^2 from outer folds.

Key differences from prior work include the use of Rasch models (instead of 2PL), a reduced feature set (without *CEFR* or *Word2Vec*), and frequency-based lexical profiling using *DeReWo*.

3 Results

Given the small number of items (n = 45 per part), all results should be interpreted cautiously. In the nested 9×9 cross-validation, each test fold contained only five items, resulting in high variance and limited reliability. Accordingly, the findings are best interpreted as indicative patterns rather than stable effects.

3.1 Model Performance

Across regularization methods (ridge, LASSO, elastic net), differences in predictive performance were marginal, however, LASSO was selected as the primary model due to its feature selection properties and interpretability. Performance metrics are reported in Table 1, which presents mean values and standard deviations across nine outer folds of nested cross-validation. In both test parts, RMSE values were close to the empirical standard deviation of item difficulty (SD_b = 0.67 for Part 5, SD_b = 0.85 for Part 7). This indicates that the models offer only limited improvement over a simple mean-based prediction.

Table 1: Performance Metrics for LASSO Models

	RMSE		R^2	
	Mean	**SD**	**Mean**	**SD**
Part 5	0.69	0.20	.24	.32
Part 7	0.81	0.33	.11	.17

This performance appears slightly weaker than reported by Dlouhá (2025) and broadly comparable to Štěpánek et al. (2023), although differences in feature sets and difficulty characteristics constrain direct comparison.

3.2 Cross-linguistic Transfer

Using the optimal hyperparameters identified in nested cross-validation ($\alpha = 1$, $\lambda = 0.071$), a final LASSO model was trained to examine selected predictors. There was no direct overlap between the predictors retained in this model and those reported by Dlouhá (2025). But given the high multicollinearity among features and the small sample size, this is not unexpected. Under such conditions, LASSO tends to select arbitrary representatives from correlated predictor clusters, limiting interpretability at the level of individual variables.

To examine cross-linguistic transfer at the level of feature space, a model for Part 5 was trained using only the predictors reported by Dlouhá (2025), which resulted in performance comparable to the full feature set (see Table 2 for a comparison of models).

In contrast, transfer at the parameter level was clearly unsuccessful, as applying the original English model with fixed coefficients led to a substantial drop in performance.

Table 2: Comparison of Cross-Linguistic Transfer Models (Part 5)

	RMSE	
	Mean	**SD**
Full Feature Set	0.69	0.20
English Features	0.71	0.20
English Coefficients	1.70	-
Baseline	0.67	-

Note. RMSE values for cross-validated models represent mean and standard deviation across 9 folds. The model with English coefficients was evaluated on the full dataset. Baseline corresponds to the standard deviation of item difficulty.

3.3 Feature Selection Stability

Feature selection exhibits limited stability. Approximately 50 predictors consistently ranked among the 100 most highly correlated features in at least 8 of 9 folds, which indicates a partially consistent core set. However, the total number of unique features that met the selection threshold at least once is large (223 for Part 5; 239 for Part 7), reflecting substantial variability, which was more pronounced for Part 7. Given this instability, interpretation of specific predictors within individual models is not warranted.

No single feature shows consistently strong association with item difficulty across both parts. This absence of robust cross-part predictors is consistent with the low correlation between examinee performance in Parts 5 and 7 (r = .15), suggesting that the two sections might be driven by partially distinct cognitive and linguistic demands. If this is the case, it would mean that item difficulty in this context is not a unitary construct and separate theoretical frameworks and feature engineering strategies may be needed for each.

4 Discussion

This study aimed to test whether a machine learning framework for predicting item difficulty from linguistic features, previously validated on English maturita exams by Štěpánek et al. (2023) and Dlouhá (2025), generalises to German, thereby addressing a gap in cross-linguistic evidence within feature-based difficulty prediction.

The results indicate that linguistic features provide a modest but unreliable predictive signal. Although LASSO performed best among the regularization methods, it showed little improvement over a naïve mean predictor. This pattern mirrors findings from Štěpánek et al. (2023) on English data and suggests that the observed performance ceiling is unlikely to be language-specific. Similar results have been reported in other small-scale educational datasets, where even more complex models fail to outperform simple baselines (Gombert et al., 2024; Kapoor et al., 2025). Overall, these findings point to limitations in the available data and signal strength rather than in the choice of modelling approach.

In this study, the problem is further amplified by the high dimensionality of the feature space relative to the number of observations. Even with regularisation, the model is prone to overfitting and unstable coefficient estimates, a phenomenon commonly referred to as the curse of dimensionality (Fulari & Rusert, 2024; Peters et al., 2025). Another contributing factor may be the homogeneous design of the maturita exam, where the reading items are tailored to a specific proficiency threshold, which constrains variance in both item difficulty and linguistic properties, leaving less signal for models to capture.

Regarding cross-linguistic transfer, the evidence is mixed. At the feature level, a model restricted to predictors identified in prior English-based research (Dlouhá, 2025) achieved performance comparable to the model trained on the full set of available features. However, since both models operate close to the baseline defined by the variance of item difficulty, it likely reflects limited predictive signal rather than successful transfer. In other words, two uninformative models will always converge on approximately the same RMSE regardless of their feature content. The feature-level transfer question therefore remains open rather than confirmed or disconfirmed. At the parameter level, directly applying the English-trained coefficients to German data produced a substantially higher RMSE confirming that parameter estimates are language-specific and retraining is necessary.

Two additional limitations constrain the validity of the results. First, not all features from Dlouhá (2025) were extracted (e.g., CEFR lists, Dale-Chall index, and Word2Vec semantic similarity), which may further reduce comparability with prior work. Second, multicollinearity complicates the interpretation of selected predictors, as several features capture overlapping information (e.g., features reflecting item length, such as raw counts).

To address these limitations, future research should rely on substantially larger datasets when applying this methodology. In contexts where larger samples are not available, the number of predictors should be reduced by prioritising theoretically grounded features, or by applying more advanced feature engineering and dimensionality reduction techniques. Predictive performance may also benefit from incorporating more recent interpretable approaches, such as features derived from LLMs (e.g., indicators of cognitive load or required reasoning complexity). Finally, for cross-linguistic comparisons, it may be useful to adopt approaches based on explicitly formulated and theoretically informed hypotheses, as this could lead to more interpretable and convincing insights into language-specific patterns.

5 Conclusion

Recent advances in machine learning, particularly in large language models, have made automated item difficulty prediction a rapidly developing area. This study aimed to replicate and extend prior research on English maturita items (Dlouhá, 2025; Štěpánek et al., 2023) by evaluating the same framework on German data and examining its cross-linguistic generalizability.

The results show that predictive performance remains limited across all conditions, and instability in feature selection prevents meaningful interpretation of individual predictors. These findings are consistent with earlier work and suggest that the main constraint lies in small item sample size. In such settings, these approaches may offer only limited added value and are better viewed as a complement to traditional difficulty estimation methods.

References

[1] Acquaye, C., Huang, Y. T., Carpuat, M., & Rudinger, R. (2026). *Take out your calculators: Estimating the real difficulty of question items with LLM student simulations*. arXiv. https://doi.org/10.48550/arXiv.2601.09953

[2] Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774–774. https://doi.org/10.21105/joss.00774

[3] Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., & Nakov, P. (2024). EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for

evaluating vision language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7768–7791). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.420

[4] Dlouhá, J. (2025). *Item response theory and machine learning in modeling difficulty of educational assessments* [Doctoral dissertation]. Univerzita Karlova, Filozofická fakulta.

[5] Dunn, K. J., Skidmore, L., & Rogers, T. (2026). When measurement meets machine learning: Interpretability and scalability in modelling item difficulty for language assessment. *Frontiers in Education*, *11, 1740237*. https://doi.org/10.3389/feduc.2026.1740237

[6] Fadillah, S. M., Ha, M., Nuraeni, E., & Indriyanti, N. Y. (2023). Exploring confidence accuracy and item difficulty in changing multiple-choice answers of scientific reasoning test. *Malaysian Journal of Learning and Instruction (MJLI)*, *20*(2), 319–341. https://doi.org/10.32890/mjli2023.20.2.5

[7] Feng, W., Tran, P., Sireci, S., & Lan, A. S. (2025). Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds), *Artificial Intelligence in Education* (pp. 31–45). Springer Nature Switzerland.

[8] Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

[9] Fulari, R., & Rusert, J. (2024). Utilizing machine learning to predict question difficulty and response time for enhanced test construction. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 528–533). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.45/

[10] Gombert, S., Menzel, L., Di Mitri, D., & Drachsler, H. (2024). Predicting item difficulty and item response time with scalar-mixed transformer encoder models and rational network regression heads. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 483–492). Association for Computational Linguistics. https://aclanthology.org/2024.bea-1.40/

[11] He, J., Peng, L., Sun, B., Yu, L., & Zhang, Y. (2021). Automatically predict question difficulty for reading comprehension exercises. *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 1398–1402. https://doi.org/10.1109/ICTAI52525.2021.00222

[12] Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, *54*(6), 969–984. https://doi.org/10.1016/j.ipm.2018.06.007

[13] Huang, Z., Liu, Q., Chen, E., Zhao, H., Gao, M., Wei, S., Su, Y., & Hu, G. (2017). Question difficulty prediction for reading problems in standard tests. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10740

[14] Institut für Deutsche Sprache. (2014). *DeReWo-2014-II-Hauptarchiv-STT.100000: Korpusbasierte Wortformenliste DEREWO, v-xxx, mit Benutzerdokumentation* [Data set]. Institut für Deutsche Sprache. https://www.ids-mannheim.de/kl/derewo/

[15] Kapoor, R., Truong, S. T., Haber, N., Ruiz-Primo, M. A., & Domingue, B. W. (2025). *Prediction of item difficulty for reading comprehension items by creation of annotated item repository*. arXiv. https://doi.org/10.48550/arXiv.2502.20663

[16] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

[17] Li, M., Chen, H., Xiao, Y., Chen, J., Jiao, H., & Zhou, T. (2025). *Can LLMs estimate student struggles? Human-AI difficulty alignment with proficiency simulation for item difficulty prediction*. arXiv. https://doi.org/10.48550/arXiv.2512.18880

[18] Li, M., Jiao, H., Zhou, T., Zhang, N., Peters, S., & Lissitz, R. W. (2025). Item difficulty modeling using

fine-tuned small and large language models. *Educational and Psychological Measurement*, *85*(6), 1065–1090. https://doi.org/10.1177/00131644251344973

[19] Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, *56*(3), 1028–1052. https://doi.org/10.1111/bjet.13570

[20] Loukina, A., Yoon, S.-Y., Sakano, J., Wei, Y., & Sheehan, K. (2016). Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In Y. Matsumoto & R. Prasad (Eds), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3245–3253). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1306/

[21] Maeda, H. (2025). Field-testing multiple-choice questions with AI examinees: English grammar items. *Educational and Psychological Measurement*, *85*(2), 221–244. https://doi.org/10.1177/00131644241281053

[22] Mair, P., & Hatzinger, R. (2007). Extended rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. https://doi.org/10.18637/jss.v020.i09

[23] Mojoyinola, M., Kehinde, O. J., & Tang, J. (2025). Enhancing item difficulty prediction in large-scale assessment with large language models. In J. Wilson, C. Ormerod, & M. Beiting Parrish (Eds), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress* (pp. 218–222). National Council on Measurement in Education (NCME). https://aclanthology.org/2025.aimecon-wip.27/

[24] Peters, S., Zhang, N., Jiao, H., Li, M., & Zhou, T. (2025). Review of text-based approaches to item difficulty modeling in large-scale assessments. In J. Wilson, C. Ormerod, & M. Beiting Parrish (Eds), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers* (pp. 37–47). National Council on Measurement in Education (NCME). https://aclanthology.org/2025.aimecon-sessions.4/

[25] R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

[26] Razavi, P., & Powers, S. J. (2025). *Estimating item difficulty using large language models and tree-based machine learning algorithms*. arXiv. https://doi.org/10.48550/arXiv.2504.08804

[27] Scarlatos, A., Fernandez, N., Ormerod, C., Lottridge, S., & Lan, A. (2025). SMART: Simulated students aligned with item response theory for question difficulty prediction. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 25071–25094). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1274

[28] Skidmore, L., Felice, M., & Dunn, K. (2025). Transformer architectures for vocabulary test item difficulty prediction. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds), *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)* (pp. 160–174). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.12

[29] Štěpánek, L., Dlouhá, J., & Martinková, P. (2023). Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms. *Mathematics*, *11*(19), 4104. https://doi.org/10.3390/math11194104

[30] Wijffels, J. (2023). *udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe' 'NLP' toolkit* (Version 0.8.11) [Computer software]. https://CRAN.R-project.org/package=udpipe

[31] Wonde, S. G., Tadesse, T., Moges, B., & Schauber, S. K. (2024). Experts' prediction of item difficulty of multiple-choice questions in the Ethiopian Undergraduate Medicine Licensure Examination. *BMC Medical Education*, *24*(1), 1016. https://doi.org/10.1186/s12909-024-06012-x

[32] Xue, K., Yaneva, V., Runyon, C., & Baldwin, P. (2020). Predicting the difficulty and response time of multiple choice questions using transfer learning. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, & T. Zesch (Eds),

Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 193–197). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.20

[33] Zotos, L., Rijn, H. van, & Nissim, M. (2025). Are you doubtful? Oh, it might be difficult then! Exploring the use of model uncertainty for question difficulty estimation. In C. Mills, G. Alexandron, D. Taibi, G. L. Bosco, & L. Paquette (Eds), *Proceedings of the 18th International Conference on Educational Data Mining* (pp. 77–89). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.15870153