



OPEN ACCESS

EDITED BY
Mark D. Reckase,
Michigan State University, United States

REVIEWED BY
Alina Anca Von Davier,
Duolingo, United States

*CORRESPONDENCE
Lucy Skidmore
✉ lucy.skidmore@britishcouncil.org

RECEIVED 05 November 2025
REVISED 24 February 2026
ACCEPTED 27 February 2026
PUBLISHED 12 March 2026

CITATION
Dunn KJ, Skidmore L and
Rogers T (2026) When measurement
meets machine learning: interpretability
and scalability in modelling item
difficulty for language assessment.
Front. Educ. 11:1740237.
doi: 10.3389/feduc.2026.1740237

COPYRIGHT
© 2026 Dunn, Skidmore and Rogers.
This is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).
The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which does
not comply with these terms.

When measurement meets machine learning: interpretability and scalability in modelling item difficulty for language assessment

Karen J. Dunn, Lucy Skidmore* and Thomas Rogers

British Council, English Language Research, London, United Kingdom

Estimation of item difficulty is essential in language test development, but recent attention has shifted toward the need also to explain and predict it. This has practical implications for item development, adaptive testing, and construct validation. Measurement specialists have traditionally explored factors contributing to item difficulty through explanatory item response theory (EIRT). In language assessment, explaining difficulty remains challenging due to the complex, context-sensitive nature of linguistic constructs. Advances in artificial intelligence (AI), most notably in machine learning (ML) and natural language processing (NLP) have expanded possibilities, offering scalable and flexible solutions, but may compromise interpretability, i.e., the capacity to link results to the underlying construct of ability. In sensitive areas, such as immigration and citizenship, generating validation evidence is critical, giving rise to a pressing need to understand the implications of using ML models in this context. This conceptual paper explores the meeting ground between measurement and machine learning, examining how these traditions converge and diverge in modelling item difficulty. Trade-offs between model interpretability and scalable application are highlighted, and implications discussed in the light of the increasingly interdisciplinary nature of this field, including possibilities offered by hybrid IRT-ML solutions.

KEYWORDS

educational measurement, explanatory item response theory, item difficulty prediction, language assessment, machine learning, model interpretability, natural language processing

1 Introduction

Estimation of test item difficulty is a foundation of educational measurement, but recent attention in language testing research has shifted toward the need to also predict and explain it (Ferrara et al., 2022; Ha et al., 2024; Settles et al., 2020; Skidmore et al., 2025). High stakes testing systems increasingly require the generation of item difficulty estimates at scale for purposes such as item development, adaptive testing, investigations into bias and construct validation, among other areas. In second language (L2) assessment, estimation of difficulty is fraught with challenges due to the highly complex nature of the constructs under examination, as well as the high level of context sensitivity of the assessment (Sayin and Bulut, 2024). Item difficulty in this field has traditionally been statistically modelled within a structured measurement framework which emphasises construct validity, interpretability, and fairness (Aryadoust et al., 2021; Min and Aryadoust, 2021). This is typically carried out using item response theory (IRT) or Rasch, which offer robust approaches, but require large amounts of authentic test-taker

responses. As well as being time-consuming and expensive, in the pre-testing context this also requires exposure of items before operational delivery, bringing implications for test security. Therefore, while traditional IRT-based approaches bestow theoretical soundness and interpretability, they limit the capacity of developers to publish new items at scale.

The shift toward digital assessment in recent years has afforded new opportunities for “computational psychometrics,” combining theoretical psychometric models with approaches from machine learning (ML) and artificial intelligence (AI) (von Davier et al., 2021). Indeed, the potential to automate estimation, using ML and pattern recognition to enable generalisable prediction of item difficulty, is increasingly a consideration among measurement specialists (Zheng et al., 2024). This lessens the requirement to employ (pre-)test data, thus mitigating exposure risks and reducing the operational burden and, as such, offers a strong advantage over more traditional Rasch/IRT approaches in terms of scalability. However, the complexity of model parameters and input data representations in ML models limit the interpretability of the findings, thus weakening our ability to link insights back to the originally defined construct of interest, with implications for building, evaluating and defending validity arguments in high-stakes contexts. With respect to language assessment, political and global sensitivities are a consideration, with many operating as gate keepers (e.g., for migration and university entry, as well as in many professional domains), meaning it is critical that validation evidence is generated supporting their implementation (cf. McNamara and Ryan, 2011). It is therefore a pressing challenge for language assessment researchers to fully understand the implications of employing ML models for item difficulty estimation. While some initial work with a language assessment focus is bringing measurement and machine learning approaches closer to one another using hybrid models (Sharpnack et al., 2024a; Yancey et al., 2024) it is not fully clear to the uninitiated where trade-offs lie with respect to the interpretability brought by established IRT methods and the potential for data-efficient application at scale offered by ML. It is therefore a timely moment to provide a review of some approaches to this increasingly interdisciplinary topic.

This mini review focuses on examples and ideas currently being applied in language assessment. It provides an opportunity for readers across disciplines to navigate the space between measurement and ML, taking as its framework a simple heuristic to indicate where various methodologies sit with respect to interpretability and scalability. The overarching goal is to situate key models used in item difficulty research and demonstrate how recent advances in both fields can be applied to support real-world measurement practices.

2 Interpretability and scalability

Two key criteria for evaluating available methods of difficulty estimation are interpretability and scalability. *Interpretability* relates to the simplicity and parsimony of the model applied. It can be understood as the degree to which transparency is built into the measurement approach. Interpretability is important, because it connects how scores are linked theoretically and practically to the assessment construct. In other words, it provides inferential scaffolding by linking the theoretical latent trait that is the focus of measurement with the actual tasks test-takers complete. An interpretable approach exhibits an

understandable relationship between model parameters, item characteristics and observed outcomes. In turn, this supports development of construct validity arguments, operational trust, and fairness, as it allows for quantitative examination of potential sources of bias, as well as broader qualitative scrutiny.

Scalability is the degree to which production of difficulty parameters can be automated, namely the amount of item response data and subsequent data processing steps required to extract information about the patterning of expected responses. Automation drives efficiency, cost effectiveness and sustainability because once deployed it is less resource intensive by design. Automation also promotes consistency and can reduce the risk of human error as well as allowing testing programmes to respond with agility to new content needs, and potentially to deploy innovative delivery methods.

Figure 1 presents a heuristic by which a range of approaches to modelling task difficulty can be understood with respect to their relative interpretability and scalability. The term “explanatory” here refers to models that integrate information about the systematic relationship between an item and its associated difficulty. The plot captures the main trade-off between measurement and ML methods put forward in this paper: traditional measurement approaches rely on explicit theoretical assumptions about test item difficulty and prioritise interpretable parameter estimation and model fit, which in turn require large amounts of test data when introducing new items. In contrast, ML methods allow greater flexibility in model structure, leveraging complex predictive models for scalable predictive performance with little or no additional pre-testing of new items once training is complete. It can be seen, however, that the two approaches are increasingly converging on an explanatory middle-ground, as this paper explores.

3 Measurement approaches

3.1 Rasch and item response theory (IRT)

The Rasch measurement model (Rasch, 1960), along with close extensions such as the rating scale model (RSM; Andrich, 1978) and

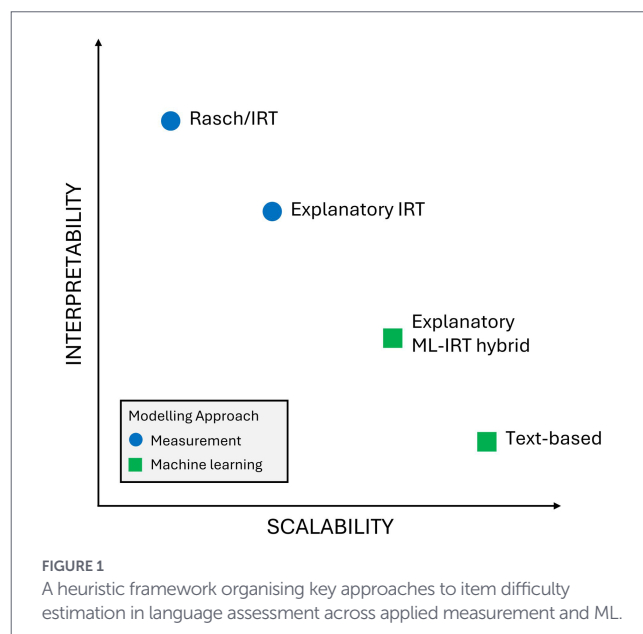


FIGURE 1
A heuristic framework organising key approaches to item difficulty estimation in language assessment across applied measurement and ML.

partial credit model (PCM; Masters, 1982), estimate item and person parameters within a probabilistic model to explain observed test-taker responses. A construct representing test-taker ability is defined as a single underlying dimension, which is formally linked to item difficulty within a strict theoretical and statistical framework. The practical process of parameterisation means that strong assumptions regarding the nature of this latent trait must be made, and in a well-fitting Rasch model, person and item scores contain all the modelled information about person measures and item calibration (Wright and Stone, 1979). Despite Rasch model extensions such as many-facet Rasch models (MFRM; Linacre, 1989) that account for multiple aspects of the assessment process, e.g., raters, test forms, task types, the focus is on estimating, rather than explaining, difficulty.

There are two quite distinct philosophical approaches toward dealing with data that do not meet model assumptions which in themselves mark the difference between Rasch and IRT. For proponents of a *model-orientated* philosophy (i.e., those who focus on preserving the model, rather than adapting it to the data) the core issue is that the theoretical conceptualisation of the latent variable is distorted by expanding or generalising beyond the Rasch model. On the other hand, a *fit-orientated* approach is more consistent with statistical modelling, in that its aim is to more flexibly account for variation in the data. This is where a range of models which fall under the broader umbrella of Item Response Theory (IRT) come into play. The conceptualisation of the relationship between ability and difficulty is complexified by various statistical and philosophical extensions. Such models may accommodate a greater range of parameters (2pl, 3pl [Birnbaum, 1968]; Graded Response Model [Samejima, 1969]), additional latency (Multidimensional IRT; Reckase, 2009), or more complex test structures (Rasch Testlet Model; Wang and Wilson, 2005). Further possibilities are evinced by reparameterisation of IRT models within a factor analytic structure (Asparouhov and Muthén, 2020; Raykov and Marcoulides, 2018) or as network models (Marsman et al., 2018). In flexing core Rasch assumptions, use of such models moves toward framing distinctions between items and how they function in a more nuanced fashion. This does not provide explanatory insight *per se*, but accounting for variations in response patterning is an important conceptual leap that leads the way toward incorporating an explanatory element in IRT.

3.2 Explanatory IRT (EIRT)

The introduction of an explanatory element in IRT was seen in the Linear Logistic Test Model (LLTM) (Fischer, 1973). This constrains item difficulty estimates as linear combinations of item design features (e.g., position effect, or response format) under the assumption of zero error in the prediction. Early work in L1 reading comprehension saw use of the LLTM to explain item difficulty (Latimer, 1982). Interest in this approach emerged for L2 language assessment in the context of Computer Adaptive Testing (CAT) for evaluating item generating systems (Sonnleitner, 2008). Such models have also been used to investigate test administration effects (Kubinger, 2008, 2009), cognitive diagnostic modelling (Baghaei and Kubinger, 2015) and for test validation (Isbell and Son, 2021).

The incorporation of an error term in EIRT models in the LLTM+e provided a significant step in the way that item difficulty explanation was conceptualised in IRT (De Boeck et al., 2011; Janssen et al., 2004). Extending Rasch to incorporate a random effect (latency)

for items as well as for persons, in the form of the Random Person Random Item (RPRI) model (De Boeck, 2008), by building the models within a Generalised Linear Mixed Model (GLMM) framework saw a straightforward move to hypothesising the role of item features as fixed effect covariates (Dunn, 2024). In this context, systematic explanation of variation in item difficulty is accompanied by an error term, or residual, in a statistical model that simultaneously estimates and explains item difficulty.

EIRT models are less interpretable than standard Rasch or IRT equivalents, as incorporating item features into the modeling framework adds complexity that requires domain expertise to interpret. Leveraging these features improves scalability, as these models generate generalisable insights into difficulty that can be applied in the development of new items. However, such models require the hypothesising of pre-conceived and specified item features, the identification and definition of which can be resource intensive and requires extensive pre-testing. This places them lower in scalability than ML approaches, which typically use automated feature representations.

4 Machine learning approaches

4.1 Text-based models

In recent years, ML approaches have emerged as a new direction for item difficulty modelling (Ferrara et al., 2022). Much of this growing interest is a result of huge advancements in natural language processing (NLP), the branch of ML concerned with modelling text. Text-based approaches to item difficulty modelling frame the problem as predicting a difficulty label directly from the item text (see Benedetto et al., 2023; Alkhuzaey et al., 2024; Peters et al., 2025 for systematic overviews of the field). Rather than using person and item response pairs, models are trained on item text and their associated difficulty labels, typically derived from expert judgments, crowdsourced annotations, or pre-testing data. Item features are extracted from text using NLP techniques which can range from interpretable linguistic features (Pandiarova et al., 2019), similar to those used for EIRT, to highly complex and context-sensitive embeddings (Skidmore et al., 2025). Models using the latter in conjunction with transformer-based model architectures (Vaswani et al., 2017) generally achieve the highest predictive accuracy in recent work (Yaneva et al., 2024).

Text-based models afford the highest level of scalability of the four approaches considered here, as once trained, a model can estimate the difficulty of a new item directly from its text without pre-testing. The generalisability of text embeddings used in more recent work further contributes to this scalability, as the same automatic feature extraction approach can be used across different test formats. The interpretability of text-based models can vary widely depending on input features and model architecture. This reflects a common trade-off seen in NLP models where the most accurate models tend to have a very high degree of feature complexity, and therefore limited interpretability (Yaneva et al., 2023). Explainable AI methods (see Holzinger et al., 2022), such as feature attribution like SHAP (Lundberg and Lee, 2017) have been employed to investigate item difficulty predictions 'post-hoc' (cf. Skidmore et al., 2025), however, in practice, these methods can only offer a small window of insight into the complex workings of the model. Given the general shift in the field toward 'black box'

approaches to modelling in NLP, further evidenced by the emerging application of large language models (LLMs) to the item difficulty prediction (Li et al., 2025; Mojinyinola et al., 2025), current text-based approaches have been assigned the least interpretable of the four models investigated.

4.2 Explanatory ML-IRT hybrid models

Explanatory ML-IRT hybrid models represent the latest development in item difficulty estimation. Such models can be seen as a re-framing of EIRT in an ML context, with the aim of combining the generalisability and scalability of text-based ML models with the interpretability and theoretical grounding of traditional measurement approaches. In such approaches, models are trained to predict the probability of a correct response for a given test-taker and item pair using item features as inputs. The explanatory elements of the model comprise NLP-derived textual features, drawing on the same types of representations used in text-based models but incorporating them explicitly as explanatory variables within an IRT framework.

A key challenge in hybridisation is ensuring that the models' underlying functions are meaningfully related to interpretable parameters (Fokkema et al., 2022). In language assessment research, two strategies for addressing this challenge have been explored. The first directly defines the ML model's function in alignment with the assumptions of explanatory IRT (McCarthy et al., 2021; Yancey et al., 2024). The second trains fully data-driven and unconstrained models, the predictions of which are subsequently mapped onto an IRT-based proxy model to derive the parameters (Sharpnack et al., 2024a, 2024b).

By aligning with the IRT framework, hybrid ML-IRT models achieve greater interpretability than purely text-based approaches, as their predictions can be linked to underlying latent parameters. Nonetheless, because these models include input features derived from text, the relationship between these representations and item difficulty remains as challenging to interpret as text-based approaches. As a result, hybrid ML-IRT models are more interpretable than text-based approaches but less so than measurement-based models. In terms of scalability, hybrid models require significantly less pre-testing data for new items compared to measurement-based approaches (Yancey et al., 2024). While relatively few studies have explored such approaches within the language assessment domain, this is an active area of research with developments underway in high-stakes assessment contexts more generally. Notably, work by Ulitzsch et al. (2025), which proposes the integration of ML-based item difficulty predictions as priors in a Bayesian IRT model.

5 Discussion

This short review situates some key measurement and machine learning approaches for item difficulty modelling in relation to two dimensions critical to language assessment concerns: interpretability and scalability. This framing provides a clear means of distinguishing where each approach sits along a spectrum of possibilities for understanding and predicting item functioning with respect to these (often competing) needs. Of course, each approach includes many adaptations and extensions not covered here, and its position along the assigned dimension may vary depending on factors such as parameterisation, feature selection, model complexity, theoretical alignment

and pre-testing requirements. Notably, however, within both measurement and ML traditions, new methods are emerging that bring these priorities closer together. In particular, the hybrid ML-IRT models combine the predictive accuracy and scalability of ML and text-based methods with aspects of the interpretability and construct alignment offered by EIRT frameworks. The scope of considerations in the current paper addresses only the explanation of item difficulty, however explanation of other parameters such as discrimination and (pseudo-) guessing are also relevant dimensions (Benedetto et al., 2020; Byrd and Srivastava, 2022; Ulitzsch et al., 2025).

Research on hybrid ML-IRT is very promising from a practical language test development perspective given the rapid shift to digital delivery and demand. While such approaches may have pre-test requirements for building initial models, or calibrating new items, once trained they have a much lighter ongoing data requirement. They can produce item difficulty estimates accurately and at pace (see Yancey et al., 2024), reducing the cost and logistical burden of continuous piloting and pre-testing, and potentially improving the test-taker experience if the use of seeded items becomes redundant. However, work in this area is still nascent, and despite offering a stronger psychometric foundation than solely text-based models, current hybrid ML-IRT approaches do not fully account for the theoretical framework from an IRT perspective. As Sharpnack et al. (2024a) note, there is room to extend model coverage and dimensionality. In practice, the complexity of such models often necessitates an indirect approach to interpretation, relying on post-hoc inspection rather than on a theoretically integrated estimation of item difficulty (Yancey et al., 2024). This constraint is confounded by the use of text embeddings as input features, which cannot easily be linked to construct-specific features of difficulty. As a result, although these models operate within an explanatory framework, their explanations remain somewhat intangible, which may limit their usefulness for communicating aspects of the validity argument to stakeholders in high-stakes assessment. However, as such advancements bring new ideas and techniques to the fore, established definitions and expectations will need to be re-examined and re-framed. Parallel considerations also apply in other areas, for instance in the auto-generation of items (e.g., Circi et al., 2023), the findings from which might have relevance for feeding insights into current concerns.

A key area for future work in item difficulty estimation for language assessment therefore lies in finding new methods to incorporate explanatory item features that are not only predictive but also grounded in measurement theory. In the absence of a broad evidence base supporting purely ML approaches, there is a higher burden of proof to demonstrate construct validity when low interpretability approaches are employed. Related work on ML applications to educational process data echoes this sentiment, highlighting the importance of balancing predictive methods for ML with theoretical considerations and domain expertise (Huang et al., 2025). Furthermore, with the swift emergence of new methods comes the additional necessity to reconsider existing validation frameworks. As AI technology becomes increasingly integrated into educational assessment, measurement researchers are flagging the need to consider the risks as well as the possibilities (Bulut et al., 2024; Zheng et al., 2024). As expressed by Zheng et al. (2024), frameworks designed for traditional measurement approaches may not be sufficient for ML-enhanced alternatives. Co-ordinated development in both areas is essential to ensure successful integration and acceptance in language assessment.

While there are undoubtedly many more avenues to explore, this mini review has sought to contextualise recent methodological progressions in both ML and measurement with respect to the mutual goal of explaining and predicting item difficulty in language assessment contexts. The application of ML is a valuable step toward bringing efficiencies and scalability for test developers. It is hoped that the reader can perceive that it is essential to evaluate how such offerings align with the specific characteristics and accountabilities of language assessment, in which interpretability and construct validity remain fundamental to defensible assessment design.

Author contributions

KD: Conceptualization, Writing – original draft, Writing – review & editing. LS: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. TR: Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

References

- Alkhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. (2024). Text-based question difficulty prediction: a systematic review of automatic approaches. *Int. J. Artif. Intell. Educ.* 34, 862–914. doi: 10.1007/s40593-023-00362-1
- Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika* 43, 561–573. doi: 10.1007/BF02293814
- Aryadoust, V., Ng, L. Y., and Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: recommendations and guidelines for research. *Lang. Test.* 38, 6–40. doi: 10.1177/0265532220927487
- Asparouhov, T., and Muthén, B. (2020). IRT in Mplus. Version 4. Technical report. Available online at: <https://www.statmodel.com>
- Baghaei, P., and Kubinger, K. D. (2015). Linear logistic test modeling with R. *Pract. Assess. Res. Eval.* 20, 1–11. doi: 10.7275/8f33-hz58
- Benedetto, L., Cappelli, A., Turrin, R., and Cremonesi, P. (2020). R2DE: A NLP approach to estimating IRT parameters of newly generated questions. LAK '20: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, (412–442). doi: 10.1145/3375462.3375517
- Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., et al. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Comput. Surv.* 55, 1–37. doi: 10.1145/3556538
- Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds. F. Lord and M. Novick (Reading, MA: Addison-Wesley).
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., et al. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *Chin. Engl. J. Educ. Meas. Eval.* 5, 1–32. doi: 10.59863/MIQL7785
- Byrd, M. A., and Srivastava, S. (2022). Predicting Difficulty and Discrimination of Natural Language Questions. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers, 2, 119–130. doi: 10.18653/v1/2022.acl-short.15
- Circi, R., Hicks, J., and Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Front. Educ.* 8, 1–5. doi: 10.3389/feduc.2023.858273
- De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73, 533–559. doi: 10.1007/s11336-008-9092-x
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *J. Stat. Softw.* 39, 1–28. doi: 10.18637/jss.v039.i12
- Dunn, K. J. (2024). Random-item Rasch models and explanatory extensions: a worked example using L2 vocabulary test item responses. *Res. Methods Appl. Linguist.* 3, 1–16. doi: 10.1016/j.rmal.2024.100143
- Ferrara, S., Steedle, J. T., and Frantz, R. S. (2022). Response demands of reading comprehension test items: a review of item difficulty modeling studies. *Appl. Meas. Educ.* 35, 237–253. doi: 10.1080/08957347.2022.2103135
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychol.* 37, 359–374. doi: 10.1016/0001-6918(73)90003-6
- Fokkema, M., Iliescu, D., Greiff, S., and Ziegler, M. (2022). Machine learning and prediction in psychological assessment: some promises and pitfalls. *Eur. J. Psychol. Assess.* 38, 165–175. doi: 10.1027/1015-5759/a000714
- Ha, H. T., Nguyen, D. T. B., and Stoeckel, T. (2024). What is the best predictor of word difficulty? A case of data mining using random forest. *Lang. Test.* 41, 828–844. doi: 10.1177/02655322241263628
- Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2022). "Explainable AI methods - a brief overview," in *xxAI - Beyond Explainable AI*, eds. A. Holzinger, R. Goebel, R. Fong, T. Moon, K. R. Müller and W. Samek (Cham: Springer), 13–38.
- Huang, J., Xin, Y. P., and Chang, H. H. (2025). The Application of Machine Learning to Educational Process Data Analysis: A Systematic Review. *Education Sciences.* 15, 1–21. doi: 10.3390/educsci15070888
- Isbell, D. R., and Son, Y.-A. (2021). Measurement properties of a standardized elicited imitation test: an integrative data analysis. *Stud. Second. Lang. Acquis.* 44, 859–885. doi: 10.1017/s0272263121000383
- Janssen, R., Schepers, J., and Peres, D. (2004). "Models with item and item group predictors," in *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, eds. P. De Boeck and M. Wilson (New York, NY: Verlag: Springer), 189–212.
- Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: from constructing tests using item generating rules to measuring item administration effects. *Psychol. Sci. Q.* 50, 311–327. https://www.psychologie-aktuell.com/fileadmin/download/PsychologyScience/3-2008/01_Kubinger.pdf
- Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educ. Psychol. Meas.* 69, 232–244. doi: 10.1177/0013164408322021

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Latimer, S. L. (1982). Using the linear logistic test model to investigate a discourse-based model of reading comprehension. *Educ. Res. Perspect.* 9, 73–94. <https://www.rasch.org/erp6.htm>
- Li, H., Aldib, R., Marchong, C., and Fan, K. (2025). “Comparing AI tools and human raters in predicting reading item difficulty,” in *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, eds. J. Wilson, C. Ormerod and M. B. Parrish (Pittsburgh, PA, USA: NCME).
- Linacre, M. (1989). *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.
- Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30, 4768–4777. doi: 10.48550/arXiv.1705.07874
- Marsman, M., Borsboom, D., Kruijs, J., Epskamp, S., van Bork, R., Waldorp, L. J., et al. (2018). An introduction to network psychometrics: relating Ising network models to item response theory models. *Multivar. Behav. Res.* 53, 15–35. doi: 10.1080/00273171.2017.1379379
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/BF02296272
- McCarthy, A. D., Yancey, K. P., LaFlair, G., Egbert, J., Liao, M., and Settles, B. (2021). “Jump-starting item parameters for adaptive language tests,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, eds. M.-F. Moens, X. Huang, L. Specia and S. W.-t. Yih (Punta Cana, Dominican Republic: ACL).
- McNamara, T., and Ryan, K. (2011). Fairness versus justice in language testing: the place of English literacy in the Australian citizenship test. *Lang. Assess. Q.* 8, 161–178. doi: 10.1080/15434303.2011.565438
- Min, S., and Aryadoust, V. (2021). A systematic review of item response theory in language assessment: implications for the dimensionality of language ability. *Stud. Educ. Eval.* 68, 1–10. doi: 10.1016/j.stueduc.2020.100963
- Mojoyinola, M., Kehinde, O. J., and Tang, J. (2025). “Enhancing item difficulty prediction in large-scale assessment with large language model,” in *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, eds. J. Wilson, C. Ormerod and M. Beiting Parrish (Pittsburgh, PA, USA: NCME).
- Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., and Brefeld, U. (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *Int. J. Artif. Intell. Educ.* 29, 342–367. doi: 10.1007/s40593-019-00180-4
- Peters, S., Zhang, N., Jiao, H., Li, M., and Zhou, T. (2025). “Review of text-based approaches to item difficulty modeling in large-scale assessments,” in *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, eds. J. Wilson, C. Ormerod and M. Beiting Parrish (Pittsburgh, PA, USA: NCME).
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Institute of Educational Research.
- Raykov, T., and Marcoulides, G. A. (2018). *A Course in Item Response Theory and Modeling with Stata*. College Station, TX: Stata Press.
- Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika* 34:1–97. doi: 10.1007/BF03372160
- Sayin, A., and Bulut, O. (2024). The difference between estimated and perceived item difficulty: an empirical study. *Int. J. Assess. Tool. Educ.* 11, 368–387. doi: 10.21449/ijate.1376160
- Settles, B., LaFlair, G., and Hagiwara, M. (2020). Machine learning-driven language assessment. *Trans. Assoc. Comput. Linguist.* 8, 247–263. doi: 10.1162/tacl_a_00310
- Sharpnack, J., Hao, K., Mulcaire, P., Bicknell, K., LaFlair, G., Yancey, K., et al. (2024a). “BanditCAT and AutoIRT: machine learning approaches to computerized adaptive testing and item calibration,” in *Proceedings of Large Foundation Models for Educational Assessment*, eds. S. Li, Z. Cui, J. Lu, D. Harris and S. Jing (Vancouver, BC, Canada: PMLR).
- Sharpnack, J., Mulcaire, P., Bicknell, K., LaFlair, G., and Yancey, K. (2024b). AutoIRT: calibrating item response theory models with automated machine learning. *CoRR, abs/2409.08823*. doi: 10.48550/arXiv.2409.08823
- Skidmore, L., Felice, M., and Dunn, K. J. (2025). “Transformer architectures for vocabulary test item difficulty prediction,” in *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, eds. E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach and R. Laarmann-Quanteet al. (Vienna, Austria: ACL).
- Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychol. Sci. Q.* 50, 345–362.
- Ulitzsch, E., Belov, D., Lüdtke, O., and Robitzsch, A. (2025). Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test. *J. Educ. Meas.* 63, 1–52. doi: 10.1111/jedm.12426
- Vaswani, A., Jones, L., Shazeer, N., Gomez, A. N., Parmar, N., Uszkoreit, J., et al. (2017). “Attention is all you need,” in *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, eds. U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus and S. V. N. Vishwanathan et al. (Long Beach, CA, USA: NeurIPS).
- von Davier, A. A., Mislevy, R. J., and Hao, J. (2021). *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python*. Cham, Switzerland: Springer.
- Wang, W. C., and Wilson, M. (2005). The Rasch testlet model. *Appl. Psychol. Meas.* 29, 126–149. doi: 10.1177/0146621604271053
- Wright, B., and Stone, M. (1979). *Best test design*. Chicago: MESA Press.
- Yancey, K. P., Runge, A., LaFlair, G., and Mulcaire, P. (2024). “BERT-IRT: accelerating item piloting with BERT embeddings and explainable IRT models,” in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, eds. E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach and R. Laarmann-Quanteet al. (Mexico City, Mexico: ACL).
- Yaneva, V., Baldwin, P., Ha, L. E., and Runyon, C. (2023). “Extracting linguistic signal from item text and its application to modeling item characteristics,” in *Advancing Natural Language Processing in Educational Assessment*, eds. V. Yaneva and M. von Davier (New York, NY: Routledge), 167–182.
- Yaneva, V., North, K., Baldwin, P., Ha, L. A., Rezayi, S., Zhou, Y., et al. (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*, eds. E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach and R. Laarmann-Quanteet al. (Mexico City, Mexico: ACL).
- Zheng, Y., Nydick, S., Huang, S., and Zhang, S. (2024). MxML (exploring the relationship between measurement and machine learning): current state of the field. *Educ. Meas. Issues Pract.* 43, 19–38. doi: 10.1111/emip.12593