

Automated Justification-Depth Scoring for Adaptive Support in AI-Supported Instructional Tasks

Songhee Han, Jueun Shin, and Jiyeon Han

Florida State University
{songhee.han, js25e, jh25}@fsu.edu

Abstract

Open-ended justifications can indicate whether learners need structured support before AI-supported revision. This study evaluates automated detection of low-depth baseline justifications (Level 1) in a 600-response instructional design task dataset. Human-coded responses were evaluated using 20 group-aware repeated train-development-test partitions, ensuring that identical normalized responses did not cross training, development, and held-out test boundaries. TF-IDF logistic regression, Complement Naive Bayes, and a RoBERTa classifier were compared using a recall-prioritized policy with a .50 development-set precision constraint. The RoBERTa classifier provided the strongest held-out performance, with Level 1 recall of .81, Level 1 F2 of .74, PR-AUC of .78, and fewer false negatives than the classical baselines. These results provide preliminary evidence that automated justification-depth scoring can help flag low-depth baseline responses that may warrant formative support before revision; future user-facing work should evaluate whether classifier-triggered support improves revision quality or task performance in general.

1 Introduction

Generative AI tools are increasingly used in educational design tasks as sources of feedback, explanation, and revision support (Kasneci et al., 2023; Maier et al., 2026; Qian, 2025). These systems can respond flexibly, but they do not automatically know whether a learner is ready for open-ended revision or needs more structured support. A short baseline justification of their work can provide an early signal: learners who give minimal or unsupported reasons may benefit from

scaffolds before being routed into a more open AI-supported revision process.

This paper studies one such signal: baseline justification depth in the context of choosing instructional models. The construct captures whether a learner merely names or accepts an instructional design model, gives a general reason, or connects the model to task demands, learner needs, engagement, differentiation, and assessment. The practical goal is not to assign a high-stakes grade. The goal is to identify low-depth justifications that may warrant formative support before revision.

In a formative scaffolding context, errors have unequal consequences. Missing a low-depth response (Level 1) may leave a learner without a checklist, worked example, or human review that could support revision. A false positive may provide extra scaffolding to a learner whose response was not a Level 1. Both errors matter, but the intended support use makes Level 1 recall especially important. With this priority in mind, we pursued the following research questions (RQs):

RQ1. How accurately can automated scoring systems detect low-depth baseline justifications (Level 1) in held-out student responses?

RQ2. When Level 1 detection is prioritized, what tradeoffs in recall, false positives, and false negatives appear across the evaluated systems?

RQ3. What do the selected system's threshold sensitivity and remaining errors imply for using it as a formative support signal?

The contribution is applied and measurement-oriented. First, the study operationalizes baseline justification depth as a priority-class detection problem for adaptive learning. Second, it compares classical and transformer-based scoring systems under the same development-selected threshold policy. Third, it reports thresholded metrics, area

under the precision-recall curve (PR-AUC), and error tradeoffs for using the score as a low-stakes formative support signal.

2 Related Work

2.1 Open-Ended Responses and Automated Scoring

Automated short answer grading (ASAG) has long examined how natural-language student responses can be scored efficiently and consistently (Burrows et al., 2015). Shared tasks such as SemEval-2013 Student Response Analysis framed short-answer scoring as an educational natural language processing problem involving semantic distinctions in student responses, including both multi-level and binary scoring variants (Dzikovska et al., 2013). More recent ASAG work has examined pretrained language models such as BERT and RoBERTa (Camus and Filighera, 2020; Devlin et al., 2019; Liu et al., 2019), but the measurement challenge remains tied to the construct being scored, the quality of the human labels, and the intended use of the automated score.

The present task differs from many ASAG settings because the score is not intended to certify correctness against a single expected answer. The target is justification depth in a formative instructional design task. This makes the task closer to detecting whether a response provides enough task-specific reasoning to support an adaptive next step. Related work on learning-system detectors shows that automated models can be useful when they identify learner states or behaviors that may warrant instructional attention. For example, gaming-the-system research has examined learner behaviors associated with poorer learning or disengagement (Baker et al., 2008; Zhang et al., 2025). The present study uses a similar decision logic: an automated scoring system is meant to identify learners who may need more structured support, not to certify mastery of instructional design.

2.2 Automated Scoring, Human Oversight, and Error Analysis

Recent automated-scoring research treats model performance as one part of a broader assessment-use problem. Tian et al. (2025) found that teachers valued rapid AI-generated feedback for formative purposes but remained cautious about automated scoring and preferred human oversight. Foster et al.

(2025) examined automated scoring in relation to agreement, reliability, and interchangeability with human scoring designs. Afrin et al. (2025) evaluated automated classification against expert judgments for AI-generated assessment content and highlighted the difficulty of replacing expert evaluation with automated judgments alone. Together, these studies suggest that automated scoring systems should be evaluated not only by predictive accuracy, but also by the decision context, the role of human judgment, and the consequences of different errors.

This issue is especially important for text-based educational scoring because model output can appear precise even when the evidence is uncertain. Jung et al. (2025) showed that automated scoring performance can depend on how task information is represented for the model. Hou et al. (2025) examined LLM-human alignment in evaluating teacher questioning practices and emphasized evidence and explanation quality in addition to ratings. Zhang et al. (2025) used error analysis to show how false positives and false negatives matter when detecting gaming in open-ended self-explanations. Because this study treats Level 1 as the priority class, precision-recall evaluation and threshold selection are also central. PR curves can be more informative than receiver operating characteristic (ROC) curves for imbalanced binary classification (Saito and Rehmsmeier, 2015), and thresholding probabilistic classifiers is part of aligning model outputs with intended decisions (Bennett et al., 2017; Lipton et al., 2014). The present study follows this direction by comparing classical and transformer-based automated scoring systems, prioritizing Level 1 detection, and interpreting false positives and false negatives in relation to formative support.

2.3 Validity for Formative Adaptive Use

Because the automated scoring system is intended to inform support decisions, its value depends on the interpretation and use of its scores. Argument-based validity theory emphasizes that validation concerns the plausibility of claims made from scores and the appropriateness of intended uses, not the score or model in isolation (Kane, 2013). In this study, the intended interpretation is narrow: a Level 1 prediction indicates that a baseline justification may be too underdeveloped for open-ended AI-supported revision without additional structure.

This interpretation connects the automated scoring system to formative assessment and feedback. Formative assessment research emphasizes information that can guide next steps in learning (Black and Wiliam, 1998; Sadler, 1989). Effective feedback should clarify goals, address the task, and provide information learners can act on (Hattie and Timperley, 2007; Shute, 2008), while learners' ability to interpret and use feedback is also central to feedback uptake (Carless and Boud, 2018). Computer-based scaffolding research further suggests that structured support can improve learning in problem-centered contexts, with effects depending on design and implementation conditions (Belland et al., 2017). In this study, the automated scoring system is evaluated as an early signal for choosing whether a learner should receive more structured support before AI-supported revision.

3 Methods

3.1 Learning Task and Data

The dataset contains 600 baseline justifications from an AI-supported instructional design study. Participants ($n = 600$) selected an instructional design model or process for designing a 30-minute mathematics game lesson for diverse third-grade learners and wrote a short justification for their choice. The response was treated as a baseline justification before AI-supported revision.

Two independent researchers coded the responses using the rubric in Table 1. When differences occurred, the final code was assigned after discussion and consensus. The present analysis uses the final consensus code levels.

Level	Operational meaning
Level 1	Minimal or unsupported justification; little task-specific explanation.
Level 2	General task-linked justification; relevant but broad or partly elaborated.
Level 3	Elaborated task-specific justification connecting model features to task demands, learners, engagement, differentiation, assessment, or iterative improvement.

Table 1: Coding levels for baseline justification depth.

3.2 Models

We compared three automated scoring systems. TF-IDF logistic regression provided an interpretable classical text baseline. Complement Naive Bayes provided a second sparse-text baseline that is often effective for text classification. The transformer model was RoBERTa, fine-tuned as a three-class classifier and evaluated by its Level 1 probability score. Classical models were implemented with scikit-learn (Pedregosa et al., 2011), and the transformer model used the RoBERTa architecture (Liu et al., 2019), which builds on pretrained transformer language modeling (Devlin et al., 2019).

3.3 Group-Aware Repeated Splits

The duplicate-response audit found two sets of repeated text after normalization: one set contained two responses, and the other contained three. These repeated responses were kept together during split generation. This retained all real responses while ensuring that identical normalized responses did not appear across training, development, and held-out test splits. This group-aware step helps prevent data leakage by ensuring that identical normalized responses do not appear across training and evaluation splits (Kaufman et al., 2012).

For each of 20 random seeds, we created a group-aware, label-stratified 420/90/90 split: identical normalized responses were kept within the same split, and the distribution of Level 1, Level 2, and Level 3 labels was preserved across training, development, and held-out test sets. Table 2 reports the full coded dataset distribution and the label counts used in each training, development, and held-out test split. The same counts were used across all 20 repeated splits, so model comparisons were not affected by differences in Level 1 prevalence across repetitions.

Split	n	Level 1	Level 2	Level 3
Full dataset	600	136	215	249
Training	420	95	151	174
Development	90	20	32	38
Held-out test	90	21	32	37

Table 2: Label counts in the full dataset and each group-aware repeated split.

3.4 Thresholds and Evaluation

For each repeated split, models were fit on the 420 training responses. Thresholds were selected on the 90-response development split and evaluated on

the 90-response held-out test split. For all three systems, the operating threshold was selected by maximizing Level 1 recall subject to development-set Level 1 precision of at least .50. F2 combines precision and recall with greater weight on recall; F3 places even stronger weight on recall and was used only in the threshold-sensitivity analysis.

This .50 value was a precision constraint, not a fixed probability cutoff; each system selected its own Level 1 probability threshold from the development split because probability scores are not directly comparable across model families. The constraint was used for the intended formative-support use: the threshold policy prioritizes catching low-depth justifications while avoiding an unconstrained high-recall policy that would flag too many Not-Level-1 responses. We also examined alternative threshold policies as a sensitivity check.

We report mean accuracy, balanced accuracy, Cohen's kappa, Level 1 precision, Level 1 recall, Level 1 F2, PR-AUC, ROC-AUC, false positives, and false negatives across the 20 held-out test splits. PR-AUC and ROC-AUC were computed from held-out Level 1 probability scores and are reported as threshold-independent ranking metrics. False positives and false negatives are summed across the 20 held-out test splits.

4 Results

4.1 Model Performance for Level 1 Detection

Across the 20 group-aware repeated splits, RoBERTa provided the strongest performance for detecting Level 1. As shown in Tables 3 and 4, RoBERTa had the highest balanced accuracy, kappa, Level 1 precision, Level 1 recall, Level 1 F2, and PR-AUC. It also produced fewer false negatives.

System	Thr.	Acc.	Bal. Acc.	Kappa	PR (ROC)-AUC
TF-IDF LogReg	.44	.78	.77	.47	.67 (.84)
ComplementNB	.21	.78	.77	.47	.64 (.83)
RoBERTa	.13	.80	.81	.54	.78 (.89)

Table 3: Overall held-out performance across 20 group-aware repeated splits. Thr. = Level 1 operating threshold; Acc. = accuracy; Bal. Acc. = balanced accuracy; Kappa = Cohen's kappa; PR (ROC)-AUC reports PR-AUC and ROC-AUC in parentheses.

System	L1 Prec.	L1 Rec.	L1 F2	FP	FN
TF-IDF LogReg	.53	.75	.69	295	104
Complement NB	.53	.75	.69	294	104
RoBERTa	.57	.81	.74	274	78

Table 4: Priority-class tradeoffs for Level 1 detection. L1 = Level 1; Prec. = precision; Rec. = recall; F2 weights recall more heavily than precision; FP = false positives; FN = false negatives. FP and FN are summed across the 20 held-out test splits. Across the 20 repeated splits, standard deviations (SDs) for the main metrics were generally modest: accuracy SDs ranged from .05 to .06, balanced accuracy SDs from .05 to .06, Level 1 recall SDs from .09 to .13, Level 1 F2 SDs from .07 to .10, and PR-AUC SDs from .08 to .09.

RoBERTa's PR-AUC was .78, compared with .67 for TF-IDF logistic regression and .64 for Complement Naive Bayes. RoBERTa's ROC-AUC was also higher (.89) than the two classical baselines (.84 and .83). Together, these ranking metrics indicate that RoBERTa assigned higher Level 1 probability scores to true Level 1 responses more consistently than the classical baselines. Figure 1 visualizes the same pattern by comparing Level 1 recall, Level 1 F2, and PR-AUC across the three systems.

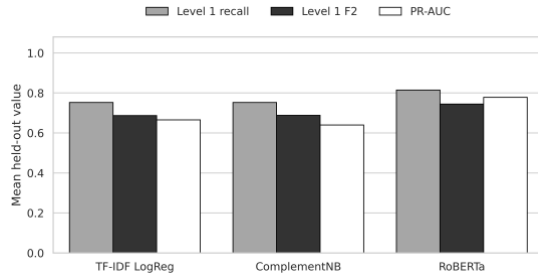


Figure 1: Mean Level 1 recall, Level 1 F2, and PR-AUC across 20 group-aware repeated splits. Values are held-out test means under the precision-constrained recall policy.

4.2 Tradeoffs in Prioritizing Level 1 Detection

The threshold sensitivity analysis shows why the selected threshold policy matters. Without a precision floor, RoBERTa reached .88 Level 1 recall and .75 Level 1 F2, but false positives increased to 426. The more aggressive F3 policy increased the Level 1 recall to .92, but false positives increased to 555. The precision-constrained recall policy retained a strong Level 1 recall (.81) while reducing false positives to 274 (Table 5).

Policy	Thr.	L1 Prec.	L1 Rec.	L1 F2	FP	FN
Recall + prec $\geq .50$	.13	.57	.81	.74	274	78
F2 no floor	.06	.48	.88	.75	426	49
F3 no floor	.03	.42	.92	.74	555	33

Table 5: RoBERTa threshold policy sensitivity across 20 group-aware repeated splits. All threshold policies used the same trained RoBERTa models and selected thresholds on the development split. F2 and F3 weight recall more strongly than precision, with F3 placing greater weight on recall.

4.3 Selected Model Error Profile

Under the selected threshold policy, the RoBERTa classifier identified 342 of 420 Level 1 responses across the 20 held-out test splits and missed 78. It correctly identified 1,106 of 1,380 Not-Level-1 responses and flagged 274 as Level 1. Figure 2 shows these outcomes within each human-coded class, making the Level 1 recall pattern and the remaining false negative cases visible.

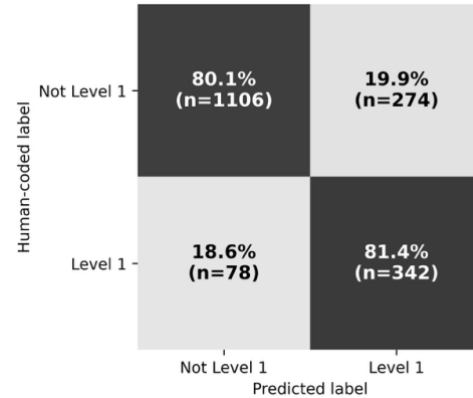


Figure 2: RoBERTa outcomes by human-coded class. Counts are summed across the 20 held-out test splits. Percentages are calculated within each human-coded class under the recall-prioritized policy with a .50 precision constraint.

For the intended formative use, the remaining false negatives are the primary risk because they represent low-depth justifications that would not receive the additional support. False positives also matter because unnecessary scaffolding can constrain a learner's revision process. However, in a low-stakes formative-support context, the selected policy provides a decent balance: it catches most Level 1 responses while keeping false positives lower than the more aggressive F2 and F3 policies shown in Table 5.

5 Discussion and Conclusion

5.1 Main Findings

This study evaluated automated justification-depth scoring as an early formative signal for AI-supported instructional design tasks. The results show that low-depth baseline justifications can be detected in held-out student responses and that the system performance depends on a threshold policy aligned with the intended use. The RoBERTa classifier showed the strongest threshold-independent ranking performance and the strongest thresholded performance under the recall-prioritized policy.

The findings are consistent with ASAG research showing the value of transformer representations for short-answer scoring, while also showing why classical sparse-text baselines remain important comparators. The gap between RoBERTa and the classical models was clearest for PR-AUC and false negatives. This pattern suggests that RoBERTa offered better Level 1 detection and

missed fewer true Level 1 responses after the final threshold was applied.

5.2 Implications for Formative Adaptive Support

The central implication is that adaptive support can begin before revision. The baseline justification functions as a short self-explanation in which learners articulate why an instructional design model fits a specific mathematics game lesson. When that explanation is minimal, unsupported, or weakly connected to task demands, a learning system can provide more structured support before the learner moves into AI-supported revision.

This use fits the formative assessment and feedback literature. Formative assessment is valuable when it informs the next step in learning (Black and Wiliam, 1998; Sadler, 1989), and effective feedback should be actionable, task-focused, and timed to support improvement (Shute, 2008). In this study, the classifier is not the feedback itself. It is a routing signal that can help determine whether a learner receives a checklist, worked example, guiding prompt, or human review.

5.3 Validity, Human Oversight, and Limits of Use

The findings should be interpreted through an argument-based validity perspective, where validation concerns the proposed interpretation and use of scores (Kane, 2013). The relevant claim is not that the classifier measures all aspects of instructional design quality. The narrower claim is that a Level 1 prediction can indicate a need for additional formative support before revision. The evidence reported here supports that narrower claim through held-out classification performance, repeated group-aware splitting, precision-recall evaluation, and sensitivity analysis, consistent with prior work on precision-recall evaluation and threshold selection under precision constraints (Bennett et al., 2017; Lipton et al., 2014; Saito and Rehmsmeier, 2015).

Human oversight remains necessary, consistent with recent automated-scoring work showing that educators value rapid AI-supported feedback but remain cautious about automated scoring decisions (Tian et al., 2025). The classifier should not be used as a high-stakes grader, as a substitute for rubric-based human judgment, or as evidence that an adaptive intervention improves learning. Its

appropriate role is a low-stakes signal that can help allocate additional scaffolding or review. Any user-facing adaptive system should allow learners and instructors to override, inspect, or revise automated support decisions.

5.4 Limitations and Future Work

Several limitations and future directions should be noted. First, the model was developed from a relatively limited amount of human-scored data: 600 consensus-coded justifications. The group-aware repeated-split analysis strengthens the stability evidence, but an external locked dataset would provide a stronger test of generalization. Second, the construct and task are specific to baseline justifications in an instructional design activity. The same scoring approach should be validated in additional tasks and domains before broader claims are made about automated justification-depth scoring. Lastly, the present study evaluated whether the classifier could provide a usable formative signal for identifying low-depth justifications. A future user-facing adaptive study should test whether support triggered by this signal improves revision quality, learner reasoning, or downstream task performance.

5.5 Conclusion

This study contributes an applied, measurement-oriented account of automated justification-depth scoring for AI-supported learning. Across 20 group-aware repeated splits, the RoBERTa classifier provided the strongest Level 1 detection under a precision-constrained recall policy, with higher Level 1 recall, F2, PR-AUC, balanced accuracy, and kappa than two classical baselines. The classifier is best understood as a low-stakes formative support signal: useful for identifying learners who may need additional scaffolding, but not sufficient as a high-stakes grade or as evidence that adaptive support improves learning outcomes.

References

- Tazin Afrin, Le An Ha, Victoria Yaneva, Keelan Evanini, Steven Go, Kristine DeRuchie, and Michael Heilig. 2025. Toward Automated Evaluation of AI-Generated Item Drafts in Clinical Assessment. In Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 172–182, Wyndham

- Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why Students Engage in “Gaming the System” Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research*, 19(2):185–224. ERIC Number: EJ789079.
- Brian R. Belland, Andrew E. Walker, Nam Ju Kim, and Mason Lefler. 2017. Synthesizing Results From Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis. *Review of Educational Research*, 87(2):309–344. <https://doi.org/10.3102/0034654316670999>
- Paul N. Bennett, David M. Chickering, Christopher Meek, and Xiaojin Zhu. 2017. Algorithms for Active Classifier Selection: Maximizing Recall with Precision Constraints. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 711–719, Cambridge United Kingdom. ACM.
- Paul Black and Dylan Wiliam. 1998. Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74. <https://doi.org/10.1080/0969595980050102>
- Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117. <https://doi.org/10.1007/s40593-014-0026-8>
- Leon Camus and Anna Filighera. 2020. Investigating Transformers for Automatic Short Answer Grading. In Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán, editors, *Artificial Intelligence in Education*, volume 12164 of *Lecture Notes in Computer Science*, pages 43–48. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-52240-7_8
- David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8):1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In Suresh Manandhar and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Jonathan Kyle Foster, James Drimalla, and Nursultan Japashov. 2025. Comparative Study of Double Scoring Design for Measuring Mathematical Quality of Instruction. In Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 163–171, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research*, 77(1):81–112. <https://doi.org/10.3102/003465430298487>
- Ruikun Hou, Tim Fütterer, Babette Bühler, Patrick Schreyer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2025. LLM-Human Alignment in Evaluating Teacher Questioning Practices: Beyond Ratings to Explanation. In Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 239–249, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Ji Yoon Jung, Ummugul Bezirhan, and Matthias von Davier. 2025. Input Optimization for Automated Scoring in Reading Assessment. In Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 1–8, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Michael T. Kane. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1):1–73. <https://doi.org/10.1111/jedm.12000>
- Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva,

- Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4):1–21. <https://doi.org/10.1145/2382577.2382579>
- Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 225–239. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.
- Uwe Maier, Moritz Seibold, and Christian Klotz. 2026. LLM-generated formative feedback in education: A qualitative systematic literature review. *Computers and Education Open*, 10:100374. <https://doi.org/10.1016/j.caeo.2026.100374>
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. <https://doi.org/10.48550/arxiv.1201.0490>
- Yufeng Qian. 2025. Pedagogical Applications of Generative AI in Higher Education: A Systematic Review of the Field. *TechTrends*, 69(5):1105–1120. <https://doi.org/10.1007/s11528-025-01100-1>
- D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144. <https://doi.org/10.1007/BF00117714>
- Takaya Saito and Marc Rehmsmeier. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*, 10(3):e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research*, 78(1):153–189. <https://doi.org/10.3102/0034654307313795>
- Zewei Tian, Alex Liu, Lief Esbenshade, Shawon Sarkar, Zachary Zhang, Kevin He, and Min Sun. 2025. Implementation Considerations for Automated AI Grading of Student Work. In Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 9–20, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Jiayi (Joyce) Zhang, Ryan S. Baker, and Bruce M. McLaren. 2025. LLM-Based Approaches for Detecting Gaming the System in Self-Explanation. In Joshua Wilson, Christopher Ormerod, and Magdalen Beiting Parrish, editors, *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 91–98, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).