

An LLM-Enhanced Score Reporting Assistant for Teachers: System Design and Usability Evaluation

Shan Zhang¹, Caitlin Tenison², Diego Zapata-Rivera², Reginald Gooch², Maya Israel¹

¹ University of Florida, ² ETS Research Institute

Correspondence: zhangshan@ufl.edu

Abstract

This study introduces an LLM-powered Smart Report Assistant grounded in audience analysis and assessment design principles. Using retrieval-augmented generation, the system helps teachers interpret assessment data through personalized and conversational reporting. Results from a usability study indicate promise for supporting score interpretation and data-informed instructional decision-making.

1 Introduction

Assessment data can provide valuable information about student learning, but interpreting assessment results often requires substantial time and expertise. As assessment systems become increasingly complex, teachers may benefit from additional support that helps them make sense of score report information and translate findings into instructional actions.

Recent advances in generative artificial intelligence (AI) offer new opportunities to provide personalized and interactive support for assessment interpretation. This study presents an AI-enhanced score reporting system that allows teachers to engage in natural language conversations about assessment results, receive tailored explanations, and obtain instructional recommendations. We report findings from a usability study examining the system's usability, perceived usefulness, and potential to support score interpretation and instructional decision-making.

2 Literature Review

Assessment results are often used to inform important educational decisions, making the appropriate interpretation and use of scores a central component of the validity argument (Kane, 2013; Downing, 2003). Score reports, which communicate assessment evidence to stakeholders and support the interpretation of learner performance, provide

the primary mechanism through which assessment information is delivered to users (Cisterna et al., 2025; Zapata-Rivera, 2018). As such, they play a critical role in supporting valid score use. For teachers, score reports help identify student strengths and weaknesses, guide instructional decisions, and connect assessment results to classroom practice (Goodman and Hambleton, 2004; Zapata-Rivera and Katz, 2014). Consequently, the effectiveness of an assessment depends not only on the quality of the measurement itself, but also on whether assessment results are communicated in ways that support appropriate interpretation and use.

Yet, teachers often struggle to interpret information presented in score reports. Studies have documented challenges related to statistical terminology, graphical displays, and representations of uncertainty, such as measurement error (Zwick et al., 2014; Heltne et al., 2023). These findings suggest teachers may benefit from additional supports that help them understand assessment information and translate results into instructional actions.

To address these challenges, researchers have explored a variety of approaches for supporting score interpretation, including alternative visualizations, explanatory materials, tutorials, and interactive reporting systems (Zwick et al., 2014; Zapata-Rivera et al., 2016). A recurring theme across this work is that effective score reporting should be designed around the needs of intended users rather than solely around psychometric considerations. Drawing on principles of audience analysis, Zapata-Rivera and Katz (Zapata-Rivera and Katz, 2014) argue that users differ in their knowledge, goals, attitudes, and decision-making contexts, and that these characteristics should inform the design of score reporting systems.

This perspective has motivated growing interest in interactive score reporting systems to better communicate students' assessment data to teachers and others who may have limited assessment under-

standing. These systems can provide explanations, contextual information, and opportunities to explore assessment data in ways that support decision-making. For example, Forsyth et al. (Forsyth et al., 2017) demonstrated the potential of conversational technologies to support teachers' understanding of score reports. Similarly, recent work has shown that generative AI-enhanced dashboards can support users in interpreting complex analytics visualizations by enabling conversational interactions, providing contextualized explanations, and facilitating personalized access to data insights (Jin et al., 2025). More recently, Zapata-Rivera et al. (Zapata-Rivera et al., in press) examined AI-enhanced interactive reports and found that teachers valued opportunities to ask questions, receive explanations, obtain instructional recommendations, and interact with assessment information in more flexible ways. These findings suggest that the value of interactive reporting lies in its ability to personalize the reporting experience, providing teachers with the information and support most relevant to their needs.

As score reporting evolves beyond static presentations, a key question is how interactive supports can be designed to address teachers' informational needs while maintaining the validity of the inferences and decisions drawn from assessment results. AI-enhanced reporting systems offer one promising approach by providing personalized explanations, recommendations and opportunities for exploration. To investigate this potential, this study introduces an AI-enhanced score reporting system that provides personalized conversational support for interpreting assessment results. We conducted a usability study with teachers to examine the system's usefulness, usability, and potential to support instructional decision-making.

3 Methods

3.1 System Architecture

Figure 1 illustrates the architecture and interaction flow of the teacher-LLM score reporting system using GPT-4o. This system combines interactive score reports, retrieval-augmented generation, and LLM support to help teachers interpret assessment results and identify instructional actions. As shown in Figure 1, information provided through the Teacher Profile Interface is incorporated into a persistent background context that includes system instructions, assessment documen-

tation, score reports, and audience-analysis information. During interaction, teacher prompts and report-specific keywords are used to retrieve relevant supporting documents from an external knowledge base. The retrieved information, together with the current report context, conversation history, and teacher prompt, is sent to the LLM to generate responses. All interactions are stored in the conversation history, enabling coherent multi-turn conversations and allowing the system to maintain context throughout a session.

3.2 Prompt Construction and Context Management

In addition to the overall system architecture, the design of the prompt sent to the LLM plays a critical role in determining the accuracy, alignment, and actionality of generated responses. Each request sent to the LLM consists of three major components: Background Content, Conversation History, and the Last Message. Background Content contains persistent information that remains fixed throughout a session, including the system prompt, teacher profile, assessment documentation, score reports, and theory-based audience analysis. Conversation History stores previous teacher prompts and AI responses to support coherent multi-turn interactions. The Last Message contains the current report context, retrieval-augmented information, and the teacher's most recent query. Overall, these components are assembled into a single structured prompt and transmitted to the LLM whenever a new request is made, enabling the model to generate contextually informed responses tailored to both the teacher and the currently selected score report.

Figure 2 presents the user interfaces of the AI-enhanced score reporting system. The system was designed to support teachers' interpretation of assessment results across classroom-, group-, and individual-level score reports while providing AI-assistance. Teachers first complete a Teacher Profile Interface (Figures 2a–2b) to specify their goals, preferences, prior experience, and instructional context. This information is subsequently used to personalize AI responses. After profile submission, teachers can navigate among classroom-, group-, and individual-level reports and interact with the AI assistant through the chat interface. The system dynamically presents report-specific suggested prompts based on the currently selected report level (Figure 2c), allowing teachers to quickly access

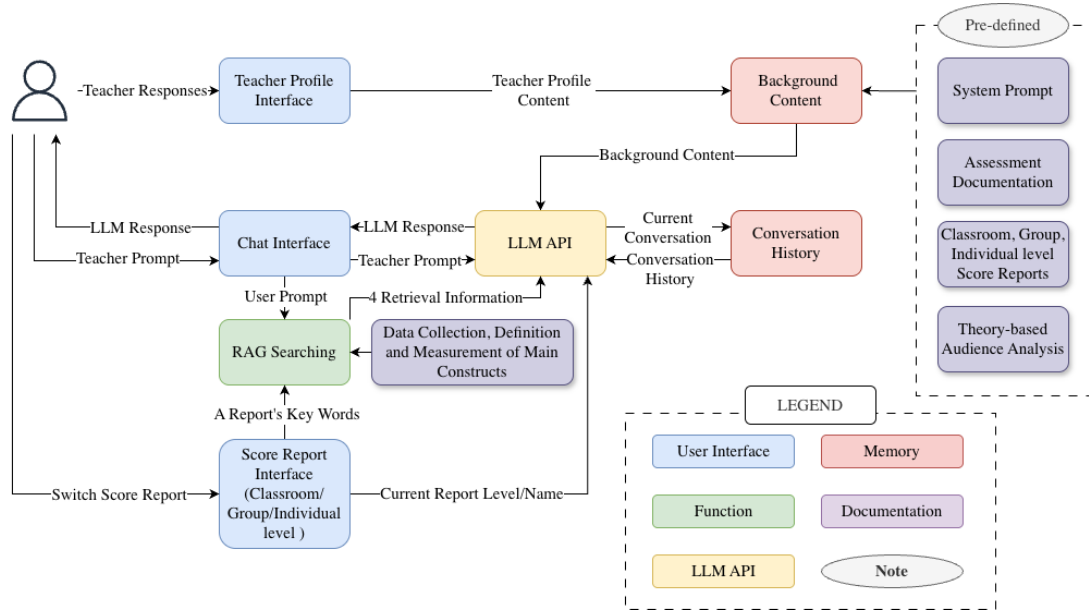


Figure 1: Architecture of the teacher-LLM interaction system for score reporting.

relevant summaries, insights, and instructional recommendations. Teachers may also enter their own questions. During interaction, conversation history can be saved, previously selected prompts are visually tracked, and AI-generated responses are displayed alongside the score reports to support data interpretation and instructional decision making (Figure 2d).

Before the study, the first author conducted 20 test interactions using four different teacher personas to evaluate the system's responses. The testing focused on three criteria: accuracy, whether the LLM's responses were correct based on the information provided to the system, such as measure definitions and scoring scales from RAG or actual students' information presented in the score reports; alignment, whether responses appropriately reflected the requested reporting level (classroom, group, or individual); and personalization, whether responses were tailored to the teacher profile represented by each persona. After the system passed this internal testing process, it was considered ready to be presented to participants in the usability study.

3.3 Participants and Procedure

We recruited 11 participants with teaching or assessment development experience to evaluate an LLM-augmented score reporting system. Each participant completed a one-hour study consisting of a pre-survey, a usability task, and a post-survey. The study protocol was reviewed and approved by the

Institutional Review Board at the authors' institution.

The pre-survey collected participants' background information, including their teaching experience, grade levels taught, teaching subject areas, prior experience with assessment score reports, comfort with interpreting score reports, and preferences for score-report features and functionalities. Prior to the usability tasks, participants were instructed to imagine that the students presented in the reports were their own students and were asked to complete a teacher profile within the system.

The usability study employed a concurrent think-aloud protocol in which participants verbalized their thoughts while interacting with the system. This approach required the participants to speak their thoughts aloud as they performed a task (Cooke, 2010). Participants completed three tasks corresponding to classroom-, group-, and individual-level score reports. In the classroom-level task, participants explored class-wide performance trends and identified groups that might require additional instructional support. In the group-level task, participants analyzed student characteristics, including performance, prior knowledge, interest, self-efficacy, and technical issues, and considered instructional strategies for supporting the group. In the individual-level task, participants reviewed a student's performance and engagement profile to determine appropriate targeted interventions. Throughout the usability study, participants were encouraged to interact with the AI assistant



Figure 2: User interfaces of the AI-enhanced score reporting system. (a) Teacher Profile Interface, (b) Completed teacher profile, (c) Report-specific menu options and suggested prompts, and (d) Example interaction between a teacher and the AI assistant.

by selecting suggested prompts or posing their own questions. Following each task, participants responded to task-specific questions regarding the usefulness, relevance, actionability, and instructional value of the AI-generated responses.

After the usability tasks, participants completed a semi-structured interview and post-survey. The interview explored participants' perceptions of the accuracy, personalization, usefulness, helpfulness, and potential improvements of the system. The post-survey included the System Usability Scale (SUS) (Brooke et al., 1996) to assess perceived usability and ease of use. Details about study materials including score reports are available on our Open Science Framework repository¹.

3.4 Data Analysis

Descriptive statistics were used to summarize participants' responses to the pre-survey and post-survey. Qualitative data from the concurrent think-aloud protocols and interview questions were analyzed using inductive thematic analysis (Clarke and Braun, 2017). All usability sessions were audio-recorded and transcribed. Two researchers inde-

pendently coded approximately 20% of the transcripts using an inductive coding approach. Following discussion and refinement of the coding framework, the coders achieved an agreement rate of 81% with a Cohen's $\kappa = .74$. One researcher then applied the finalized coding scheme to the remaining transcripts. Codes were subsequently grouped into broader themes to identify recurring patterns related to participants' perceptions of the AI-enhanced score reporting system, including perceived benefits, limitations, and suggestions for improvement.

4 Results

4.1 Pre-survey

A total of 11 participants completed the pre-survey. Most participants had prior teaching experience ($n = 10$), while one participant reported no teaching experience.

Among participants with teaching experience, the largest group reported 6–10 years of teaching experience ($n = 5$), followed by 3–5 years ($n = 3$) and 0–2 years ($n = 2$). Participants represented a diverse range of teaching areas, including English Language Arts ($n = 2$), Science ($n = 2$), Social

¹https://osf.io/n7msj/overview?view_only=e814189caf3d4666838ab6f1b23813e5

Studies (n = 2), and English as a Second Language (ESL; n = 2). Additional teaching areas included Mathematics, Special Education, Assessment, Educational Technology, Elementary Multi-subject Education, Senior Honors Thesis Seminar instruction, and Cognitive Science for Decision Making (n = 1). Teaching experience spanned multiple grade levels, with Postsecondary/Adult Education being the most frequently reported level (n = 7), followed by Grades 3–5 (n = 4), Grades 9–12 (n = 4), Grades 6–8 (n = 2), and Grades K–2 (n = 1).

Participants generally reported substantial experience with educational score reports including with Individual Student Reports (n = 7, 63.6%), Group/Classroom-Level Reports (n = 6, 54.5%), followed by Paper-Based Reports (n = 5, 45.5%). One participant reported experience with Interactive Dashboards, and one participant reported experience using spreadsheets of student data and learning management system summary reports. Three participants (27.3%) indicated that they had not previously interacted with assessment score reports as teachers.

Participants also reported relatively high levels of confidence in interpreting score reports. Five participants (45.5%) indicated they were *Very Comfortable*, four (36.4%) were *Comfortable*, and two (18.2%) selected *Neither Comfortable nor Uncomfortable*. No participants reported being uncomfortable with score report interpretation.

When asked about preferred score report features (Figure 3), participants most frequently selected visualizations and the ability to examine performance across classroom-, group-, and individual-level reports. Interpretive text explanations and text-based summaries were also commonly valued. Overall, participants expressed a preference for score reports that combine visual representations of assessment results with explanatory information that supports interpretation and instructional decision-making. See Figure 3 for the complete distribution of responses.

Finally, six participants (54.5%) reported using assessment score reports to inform instructional decision-making, whereas five participants (45.5%) reported not doing so. As shown in Figure 4, participants who used score reports described using them to identify learning gaps, adjust instruction, modify lesson plans, monitor student progress, determine concepts requiring additional instruction, support differentiated instruction, identify students for intervention, and review item-level patterns of

student performance.

4.2 Usability Study

Thematic analysis of the think-aloud protocols and interview responses revealed several themes related to participants' perceptions of the AI-enhanced score reporting system. Overall, participants viewed the system positively and highlighted its potential to support data interpretation and instructional decision-making. At the same time, they identified several areas for improvement related to recommendation specificity, transparency, and the appropriate role of AI in educational contexts.

Theme 1: Teachers valued actionable instructional recommendations (n = 10). Across the classroom-, group-, and individual-level tasks, participants consistently described the AI-generated recommendations as useful because they translated assessment information into potential instructional actions. Rather than simply summarizing student performance, the AI assistant helped teachers identify intervention strategies and instructional next steps, and provided additional support through links to external resources.

Theme 2: Contextual learner information made the reports more meaningful (n = 8). Participants valued the ability to consider assessment results alongside other indicators of student learning and experience such as WIDA scores and metrics on prior knowledge, topic interest, self-efficacy, engagement, and technical issues. Integrating these sources of evidence helped teachers develop a more holistic understanding of learner performance and consider why students may have struggled. Participants particularly appreciated that the AI assistant could incorporate contextual measures such as self-efficacy and motivation into its recommendations rather than relying solely on performance outcomes. Several participants also highlighted the system's ability to synthesize information across multiple learner characteristics simultaneously, helping them connect different data points and see the whole picture of student learning.

Theme 3: AI responses generally aligned with teacher reasoning, but teachers still relied on professional judgment (n = 10). Participants frequently reported that AI-generated suggestions aligned with what they would typically do in their classrooms. For many teachers, the AI served as a useful second opinion that confirmed, extended, or organized their own thinking. However, participants consistently emphasized that AI should

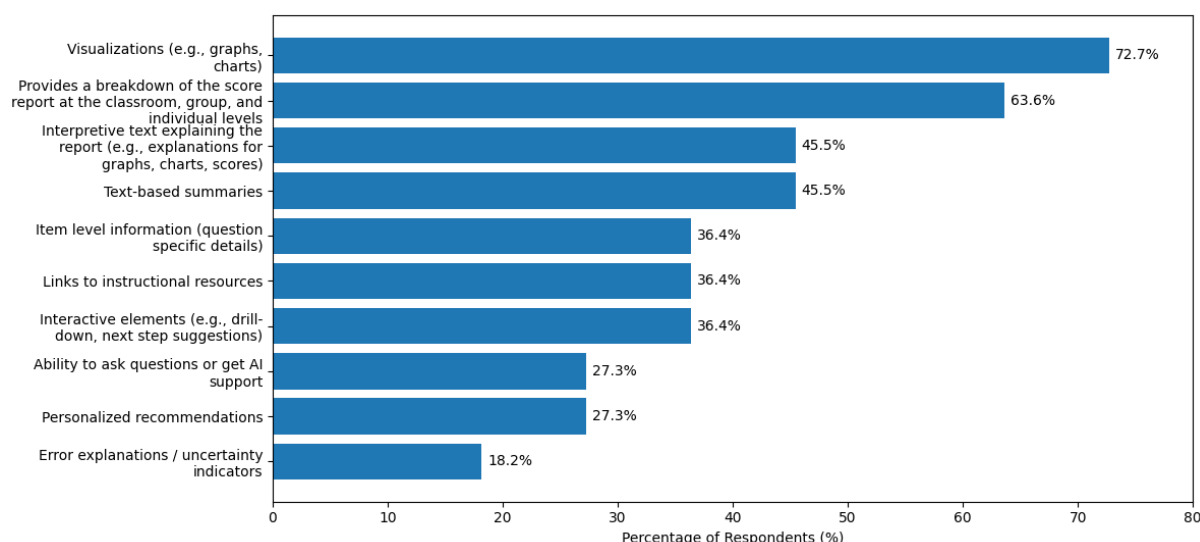


Figure 3: Participants' preferred features in student score reports.

support rather than replace teacher expertise. Participants evaluated AI recommendations against their own knowledge of students, classroom context, and instructional goals before adopting.

Theme 4: Teacher Profile information supported personalization (n = 7). Several participants noted that the AI-generated responses appeared to account for their prior knowledge, assessment experience, and attitudes toward score reports as captured in the Teacher Profile. Participants also reported that the recommendations aligned with the communication preferences they had specified in the Teacher Profile, which contributed to a greater sense of personalization and relevance. However, participants suggested that future versions should tailor recommendations more explicitly based on teacher experience, instructional goals, classroom contexts/constraints, and available resources.

Theme 5: Teachers wanted richer evidence and clearer explanations (n = 9). Participants saw a need for more transparent connections between AI recommendations and underlying assessment evidence. Participants frequently requested item-level information, task details, student responses, construct definitions, and explanations of how learner characteristics were measured. They also expressed interest in comparing multiple constructs simultaneously and understanding how specific learner characteristics contributed to AI-generated recommendations. In addition to summaries, participants wanted explanations that clarified why students struggled and what evidence supported each recommendation.

Theme 6: Recommendations were useful but sometimes too generic or impractical (n = 3).

Although participants generally viewed recommendations positively, some suggestions were seen as overly broad or lacking actionable detail. Recommendations such as using visual supports, scaffolding, or small-group instruction were considered appropriate but too general to guide classroom action. Participants requested more concrete examples, instructional resources, external links, lesson ideas, and implementation guidance. Some recommendations were also perceived as impractical. For example, several participants questioned suggestions involving pairing high-performing and low-performing students, noting that such approaches may not always align with classroom realities and established instructional practices. Several participants also reported that some AI-generated responses were overly long, and they suggested providing more concise summaries while retaining access to additional details when needed.

Theme 7: The system may be especially useful for novice teachers or teachers with lower assessment familiarity (n = 5).

Participants suggested that the AI assistant may be particularly valuable for novice teachers, educators with limited assessment literacy, or teachers unfamiliar with the assessed content. More experienced teachers often viewed the recommendations as confirming strategies they already employed, whereas less experienced teachers may benefit from explicit guidance on interpreting reports and planning interventions. Participants noted that the system could help re-

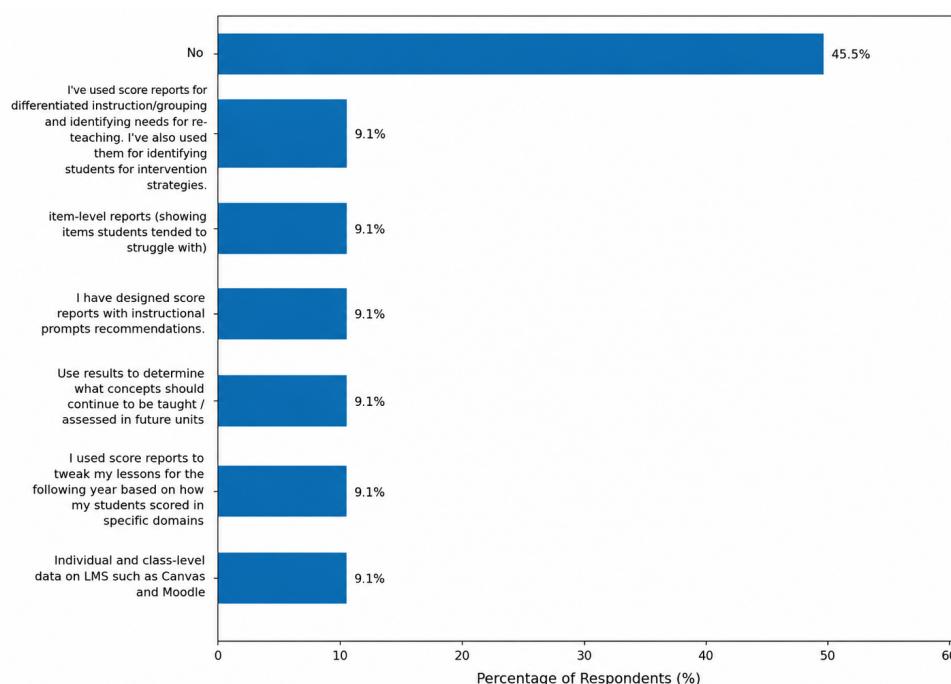


Figure 4: Use of score reports for instructional decision-making.

duce the expertise required to interpret complex assessment information.

Theme 8: Participants perceived the system as a tool for improving efficiency and reducing cognitive workload (n = 9). Participants described the AI assistant as a time-saving tool that could reduce the effort required to interpret complex score reports. Participants commented positively on the accuracy of the AI-generated responses, noting that the recommendations were generally well grounded in the information provided within the score reports. Several teachers further observed that the suggestions appropriately reflected the level of analysis being examined, with recommendations adapting to classroom-, group-, and individual-level reports in ways that aligned with their instructional decision-making needs. Several participants also suggested that the system would be particularly valuable when applied to assessment reports spanning an entire semester or school year. In such contexts, teachers envisioned using the AI assistant to review progress across multiple students, identify trends over time, monitor growth relative to standards, and support instructional planning across a larger body of assessment data.

Overall, participants viewed the AI-enhanced score reporting system as a promising tool for supporting assessment interpretation and instructional planning. However, they also emphasized the im-

portance of maintaining teacher agency, providing transparent evidence-based recommendations, and ensuring that AI-generated suggestions are appropriately aligned with classroom contexts.

4.3 Post Survey

On a scale from 0 to 100, teachers reported generally favorable usability of the Score Report AI Assistant, with an average SUS score of 74.77 (SD = 14.85, Median = 75, range = 42.5–95). Specifically, most participants agreed or strongly agreed that the system was easy to use (90.9%) and that they felt confident using it (81.8%). Similarly, 90.9% indicated that they would like to use the system frequently, while 63.6% agreed that most teachers would learn to use the system quickly (Figure 5).

Participants also responded favorably to items on system functionality and integration. Nearly three-quarters of participants (72.7%) agreed or strongly agreed that the various functions of the system were well integrated. In contrast, most participants disagreed or strongly disagreed that the system was unnecessarily complex (72.7%), cumbersome to use (81.8%), or required learning many things before getting started (72.7%). Furthermore, 90.9% disagreed or strongly disagreed that they would need technical support to use the system, and that the system exhibited excessive inconsistency. These findings suggest that participants perceived

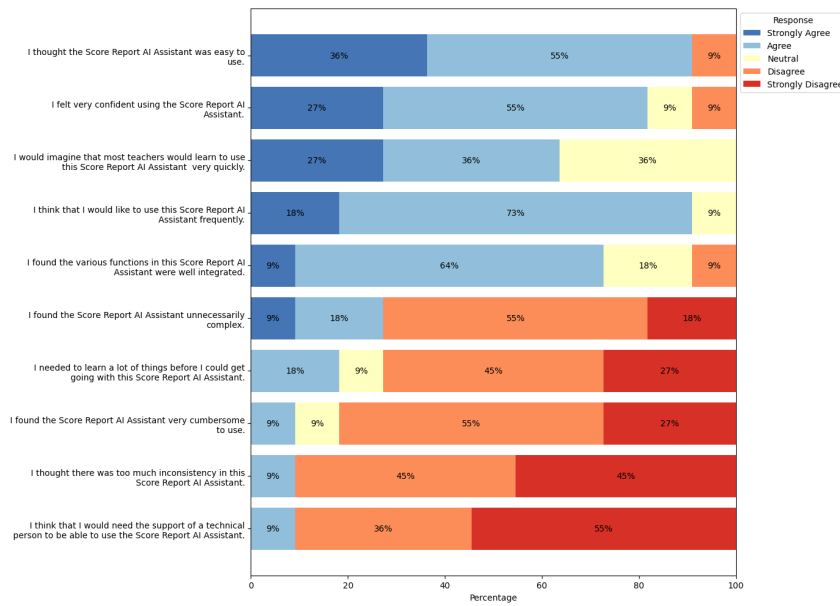


Figure 5: Participants’ responses to the System Usability Scale (SUS) items.

the AI-enhanced score reporting system to be intuitive and easy to use, while requiring minimal external support.

5 Discussion

Participants in this study rarely treated AI-generated recommendations as final answers. Instead, they used the system to explore possible explanations, examine supporting evidence, and consider alternative interpretations before reaching their own conclusions. This pattern aligns with recent work on teacher agency, which emphasizes teachers’ capacity to critically engage with AI-generated outputs while maintaining professional judgment and autonomy in decision-making (Barrot, 2026). The findings suggest that teachers view AI as a resource for reflection and sensemaking rather than a substitute for instructional expertise.

Several teachers noted that understanding the non-cognitive data presented on students was often as important as understanding their assessment performance itself, particularly when identifying appropriate supports for students. Participants envisioned the system as a tool that could help them understand how such contextual information interplays with patterns of performance across longer periods of time, such as a semester or academic year, and across multiple assessments. Future reporting systems could support teachers not only in reviewing assessment results but also in helping them to build their understanding of students’ development over time.

Recommendations from the system were often viewed as helpful, but a takeaway is that participants want more specific recommendations backed by stronger connections to the evidence available in the reports. Teachers appear to value recommendations made by AI-enhanced reporting systems, yet they expect those recommendations to be transparent, evidence-based, and grounded in classroom realities.

Overall, these findings suggest that the design of AI-enhanced reporting systems is not simply a user-interface problem, but also a measurement challenge. Decisions about which learner characteristics to include, how they are measured, and what forms of analysis are supported directly shape the inferences that users may draw. Consequently, future work must consider not only how to provide richer contextual information, but also how to ensure that the relationships, trends and recommendations generated by reporting systems in response to user needs are psychometrically justified and support valid score use.

6 Conclusion

This study explored the use of conversational AI to support teachers’ interpretation of assessment results. Our findings identify design considerations for AI-enhanced reporting systems, including the importance of contextual learner information, transparency, and teacher agency. Future research should examine the necessity of retrieval-augmented generation within the system and eval-

uate whether AI-enhanced reporting can meaningfully improve assessment interpretation and instructional decision-making, particularly among teachers with lower confidence in understanding measurement and assessment information.

References

- Jessie S Barrot. 2026. Generative artificial intelligence for automated essay scoring: Exploring teacher agency through an ecological perspective. *Assessing Writing*, 67:100990.
- John Brooke and 1 others. 1996. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.
- Dante Cisterna, Jamie N Mikeska, Katherine E Castellano, and Jennifer Lentini. 2025. Score reports for teacher educators: Informing assessment results for content knowledge for teaching. *Educational assessment*, 30(1):41–65.
- Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology*, 12(3):297–298.
- Lynne Cooke. 2010. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE transactions on professional communication*, 53(3):202–215.
- Steven M Downing. 2003. Validity: on the meaningful interpretation of assessment data. *Medical education*, 37(9):830–837.
- Carol M Forsyth, Stephanie Peters, Diego Zapata-Rivera, Jennifer Lentini, Art Graesser, and Zhiqiang Cai. 2017. Interactive score reporting: an autotutor-based system for teachers. In *International Conference on Artificial Intelligence in Education*, pages 506–509. Springer.
- Dean P Goodman and Ronald K Hambleton. 2004. Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17(2):145–220.
- Aleksander Heltne, Niek Frans, Benjamin Hummelen, Erik Falkum, Sara Germans Selvik, and Muirne CS Paap. 2023. A systematic review of measurement uncertainty visualizations in the context of standardized assessments. *Scandinavian Journal of Psychology*, 64(5):595–608.
- Yueqiao Jin, Kaixun Yang, Lixiang Yan, Vanessa Echeverria, Linxuan Zhao, Riordan Alfredo, Mikaela Milesi, Jie Xiang Fan, Xinyu Li, Dragan Gasevic, and 1 others. 2025. Chatting with a learning analytics dashboard: The role of generative ai literacy on learner interaction with conventional and scaffolding chatbots. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 579–590.
- Michael T Kane. 2013. Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1):1–73.
- Diego Zapata-Rivera. 2018. Introduction: Why is score reporting relevant? In *Score reporting research and applications*, pages 1–5. Routledge.
- Diego Zapata-Rivera, Reginald M. Gooch, Caitlin Tenison, and Robert Long. in press. Ai-enhanced interactive reports for teachers. In Alina von Davier and Duanli Yan, editors, *Artificial Intelligence for Measurement in Educational Learning and Assessment*.
- Diego Zapata-Rivera, Rebecca Zwick, and Margaret Vezzu. 2016. Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, 21(3):215–229.
- Juan Diego Zapata-Rivera and Irvin R Katz. 2014. Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, 21(4):442–463.
- Rebecca Zwick, Diego Zapata-Rivera, and Mary Hegarty. 2014. Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, 19(2):116–138.