Leveraging multi-AI agents for a teacher co-design

Hongwen Guo [*1], Matthew S. Johnson[†1], Luis Saldivia[‡1], Michelle Worthington[§1], and Kadrye Ercikan[¶1]

[1]ETS Research Institute, 660 Rosedale Rd, Princeton, NJ 08541

[*]**Paper prepared for 2025 AIME-con at Pittsburgh**

Abstract

This study uses multi-AI agents to accelerate teacher co-design efforts. It innovatively links student profiles obtained from numerical assessment data to AI agents in natural languages. The AI agents simulate human inquiry, enrich feedback and ground it in teachers' knowledge and practice, showing significant potential for transforming assessment practice and research.

Keywords: Human-centered AI, AI agents, large-scale assessment, response and process data, feedback

1 Background

1.1 Literature review

The existing work in learning analytics and educational data mining has provided a strong foundation for understanding student learning through data. Researchers have been seminal in leveraging fine-grained log data from digital learning environments to offer deep insights into complex learning processes (e.g., Baker and Yacef, 2009; Baker, 2021; Thomas et al., 2025; Darvishi et al., 2024). Work in these learning areas also demonstrates the practical application of learning process data in refining intelligent tutoring systems and prediction of student learning outcomes (e.g., Khan Academy, 2025; Ritter et al., 2013; Zheng et al., 2019).

While learning analytics leverages diverse student interaction data (e.g., from learning management systems) to provide feedback and improve instructional design, assessment analytics applies similar data mining and statistical techniques to interpret student performance on tests, evaluate item quality, ensure assessment validity and fairness, and develop measurement innovation (Ercikan et al., 2023; Ercikan and Pellegrino, 2017). Recently, process data collected from log data in large-scale assessments (LSAs) has been gaining momentum in educational measurement, largely due to data availability from NAEP, PISA, TIMSS, etc. (National Assessment Governing Board, 2020; Organisation for Economic Co-operation and Development, 2020; International Association for the Evaluation of Educational Achievement, 2020). Studies using process data in LSAs and other assessments can be found in areas such as test-taking strategies, score validity on the assessments, its relationship with performance (Ercikan et al., 2020; Guo and Ercikan, 2021; Pools and Monseur, 2021), and problem-solving patterns (Greiff et al., 2016; Zoanetti and Griffin, 2017).

Process/log data, as exhibited in the above studies, contain nuanced information about how students engaged with tasks and assessments. Such large and complex data from LSAs may pose challenges to traditional psychometric analysis but offer opportunities for using AI to discover data insights. Recent studies (e.g., Guo et al., 2024a,b) attempted to use NAEP multi-source data (i.e., response data and process data) and human-centered AI (HAI) frameworks to generate preliminary student profiles, which show promises in contextualizing a performance score and providing meaningful and actionable feedback to classroom teachers. These preliminary student profiles were created based on multi-source data when students interacted with LSAs digital platforms. The HAI approach helped to identify near a dozen preliminary profiles, many associated with low-performing students. For teaching, such profiles are intended to provide educators with rich, meaningful feedback, helping them understand how students engaged with the assessment beyond a performance score, which can shed light on students' learning skills to inform classroom teaching practices. Similar

[*]hguo@ets.org; Corresponding author
[†]msjohnson@ets.org
[‡]lsaldivia@ets.org
[§]mworthington@ets.org
[¶]kercikan@ets.org

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 25–34

to AI applications in learning systems, the AI applications on LSAs can help to drive significant innovation in LSA practice and research, informing both teaching and learning practices.

However, for such preliminary student profiles generated from LSA research to make a real impact on teaching, they need to be refined and improved and grounded in classroom practice with a teacher co-design.

1.2 Aims

As a stepping stone toward transforming LSA research to teaching practice, the primary goal of the current study is to leverage multiple AI agents and their reasoning capabilities to facilitate an effective teacher co-design for transforming assessment research into teaching professional development.

More specifically, in the current study, we propose to use AI multi-agents to find a common ground before we collaborate with real teachers. AI agents will act as experienced educators to understand the multi-source data, refine the preliminary student profiles, generate highlights of student strengths and needs, and suggest possible intervention. These AI-educator agents also communicate with an AI-researcher agent, so that these jointly-created feedback/narratives about a student will be better grounded in both teachers' classroom practice and assessment data for the later teacher co-design. This study addresses the following research questions:

RQ-1: How to extract explainable features that can be mapped into natural languages, so that AI agents can understand?

RQ-2: How to create a coherent crew of AI agents that produce feedback based on empirical data?

RQ-3: How to evaluate whether AI agents' outputs are consistent with research findings?

The project intersects with current advancements in AI and education technologies to give back more data insights to educators to bridge assessment outcomes and learning needs. Deep data insights from LSAs provide indicators of broader student attributes (time management, test navigation regulation, engagement, learning needs) beyond a performance score, which offers rich information for teachers to prepare for personalized intervention. The current study exemplifies an innovative AI application in measurement research. The use of distinct AI agent personas - representing teachers in varied contexts, a coach, and a researcher - demonstrates an attempt to model diverse expert reasoning and tackle the complexity of student data interpretation. The AI-crew-generated feedback will help accelerate and enrich the teacher co-design, so that AI-agents' results can be better communicated to and understood by real teachers, allowing them to endorse, reject, or revise the results to support their students.

In the following method section, we briefly introduce the data and insights produced from previous research, describe explainable feature creation to address RQ-1; describe a crew of multi-AI agents for our exploration to address RQ-2; and additional experiment with AI to address RQ-3. In the result section, we display examples of outputs from the AI crew, highlighting the diverse perspectives from the AI agents, the AI-refined student profiles, and other useful feedback to educators. In the result section, we also show the evaluation of outputs from AI crew. In the last section, we discuss the contributions of this study, its limitations, and the future directions.

2 Methods

2.1 Data

In this study, we used a subset of data that contained manually labeled preliminary student profiles produced from Guo et al.'s (2024b) using the National Assessment of Educational Progress (NAEP) Grade 8 Mathematics assessment. The NAEP multi-source data contain a student's item responses, item response times, number of item visits, digital tools uses, as well as the sequences of item navigation (i.e., how much time was spent on an item and in what order). For details, please refer to National Assessment Governing Board's (2020) for NAEP process data released for secondary analysis.

A human-centered AI (HAI) architecture was proposed for human experts and AI collaboration to produce preliminary profiles for over ten thousand students who took one NAEP math block. The proposed HAI framework (refer to Figure 1) is built on a three-step architecture. This structure underscores: firstly, the critical input of human knowledge in the data preprocessing; secondly, the application of AI algorithms (including machine learning, deep learning) to improve data analysis and identify patterns; and thirdly, the integration of AI's computational power (e.g. active learning) with human expert judgment to finalize the

profiles. Researchers and content experts investigated the extracted features and visualization of the multi-source student data and created students' preliminary profiles with AI for all students.

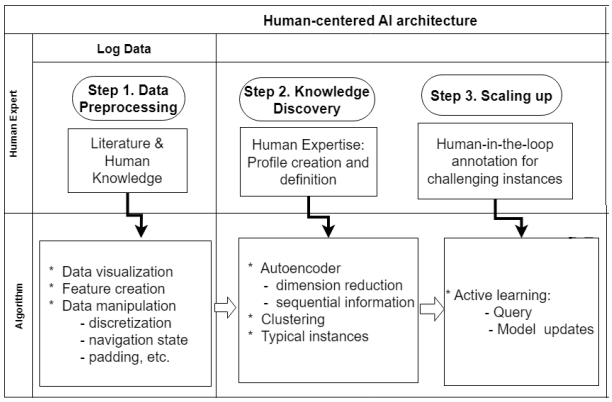

Figure 1: The human-centered AI architecture with three key steps from data preprocessing to scaling up to produce student profiles (Guo et al., 2024a,b).

2.2 Feature Creation and Mapping

In previous multi-source data studies (Guo et al., 2024a,b), deep learning models (i.e., autoencoders) were used to compress sequential data to produce latent features for profile prediction. Because of well-known challenges in latent feature interpretation, in the current study, we extracted features that were explainable from the multi-source data (refer to Table 1 for main features) to address RQ-1. These explainable features enabled mapping numerical values into natural languages for exploration of multi-AI agents. Features, not presented, also include mean and standard deviation of item visits and item scan, locations of longest bursts and longest jump, etc.

Among the explainable features, the new features, including navigation regularity, scan, scan burst, and jump, were created to address the challenge in describing a student's sequential navigation behaviors. Definitions of some of the new features are straightforward, as described in Table 1. Below we focus on the definition of the navigation regularity (Reg) feature which uses the concept of entropy to quantify and measure the unpredictability or randomness of a student's navigation behaviors when interacting with the test as a whole.

More specifically, let $Y = \{y_1, y_2, \cdots, y_{m+1}\}$ be the sequence of item numbers a student visited from the beginning of the test session to the end, where $m+1$ is the total number of item visits; let $X = \{x_1, x_2, \cdots, x_m\}$ be the lag difference (i.e., $x_t = y_{t+1} - y_t$. Let $X^* = \{x_1^*, x_2^*, \cdots, x_m^*\}$ be the absolute value of X , where $x_t^* = |x_t|$.[1]

The entropy H of the sequence $X^* = \{x_1^*, x_2^*, \cdots, x_m^*\}$ is

$$H(X) = -\sum_{1}^{m}[p(x_i) * \log(p(x_i))],$$

where $p(x_i)$ is the probability of x_i. The navigation regularity (or simply, Regularity) is defined as:

$$\text{Reg}(X) = \frac{1}{1 + H(X)}, \quad (1)$$

so that the upper bound of $Reg(X)$ is 1. That is, for students to have the value of $Reg(X) = 1$ on the test, they have to navigate the test very orderly (i.e., moving between adjacent items only). A value of $Reg(X)$ close to zero indicates unregulated navigation behaviors.

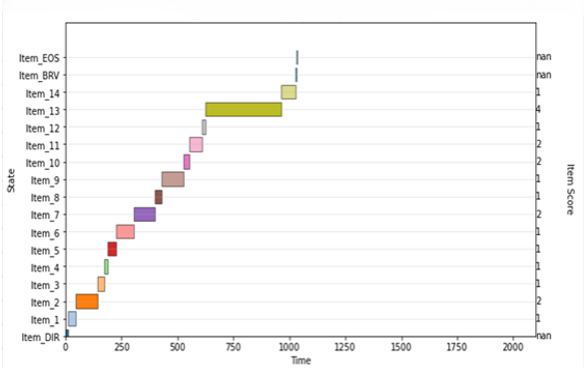

Figure 2: Navigation plots of student A. In the plot, the x-axis stands for the testing time, the y-axis on the left stands for the item state (i.e., what item the student was working on) and other navigation states, and the y-axis on the right stands for the item score the student obtained. Each colored rectangle shows the time spent on an individual navigation item state (Guo et al., 2024b). The plot shows Student A with $Reg(X) = 1$.

A navigation plot is the visualization of three sequences (navigation item state, time on the state, and score received (Guo et al., 2024b). Refer to two examples in Figures 2 and 3, respectively, which shows two students' navigation patterns. One student (Student A; $Reg(X) = 1$) worked linearly one item at a time following the item presentation order on the test; the other (Student B;

[1]Taking absolute value is to conveniently define the jump event. A jump event occurs when a $X_i^* \geq 2$. Readers can modify these definitions based on their circumstances.

$Reg(X) = 0.29$) exhibited an irregular navigation pattern, showing behaviors such as quick item scans, skipping items, and jumping among items.

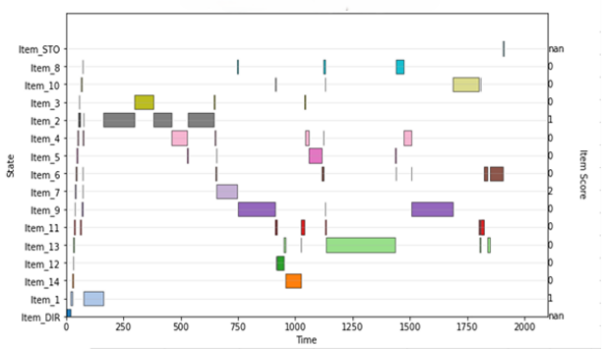

Figure 3: Navigation plots of Student B. In the plot, the x-axis stands for the testing time, the y-axis on the left stands for the item state (i.e., what item the student was working on) and other navigation states, and the y-axis on the right stands for the item score the student obtained. Each colored rectangle shows the time spent on an individual navigation item state (Guo et al., 2024b). The plot shows Student B with $Reg(X) = 0.29$.

2.3 AI Agents

In this exploration, to address RQ-2, we created a crew of five AI agents embodying "teacher", "professional coach", and "researcher" personas (refer to Figure 4). Their roles logically build upon each other sequentially.

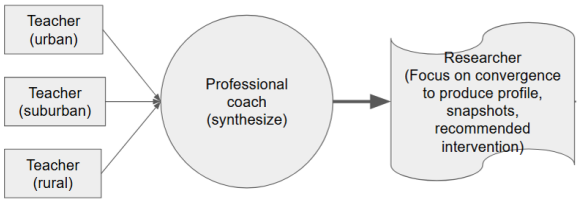

Figure 4: A crew of AI agents

The three math teacher AI-agents represent experienced teachers working in the urban, suburban, and rural settings, respectively, who help to pretest concepts, provide diverse contextual lenses, and react to an assessment idea from research results, flagging potential misunderstandings or concerns early on. Each teacher AI-agent is required to read the student data (i.e., features and preliminary profiles), reflect on their knowledge, and provide factual feedback to improve the preliminary profile, highlight students' strength and growth areas, and recommend potential intervention in a whole person approach.

The math coach AI-agent reads and synthesizes the three teacher AI-agents' results, including convergence and divergence in teachers' reports. The math coach's report may help to identify hypothetical points of friction or alignment in how each persona views data insights and practicalities from the assessment.

The scientist AI-agent reads the math coach's report, checks against the preliminary student profile, to ensure alignment and conciseness of the refined profiles, student's strength and needs, and recommended intervention.

This workflow (from multiple initial teacher analyses to coach synthesis, and then to research-informed summary), as shown in Figure 4, mimics a rigorous, collaborative human inquiry process, moving from divergent thinking to convergent, to provide meaningful collaboration in the teacher co-design.

2.4 Evaluation

To compare the AI-crew-refined profiles with the original manually-labeled preliminary profiles to address RQ-3, we conducted experiments using sentence embedding approaches and the GenAI agent approach. In the first approach, we cluster the sentence embedding results into ten clusters, and then evaluate the consistency between the embedding-generated clusters and preliminary profiles. In the second approach, we created an independent editing agent (with the persona of a meticulous editor and data analyst specializing in educational data). This AI-editor read and analyzed the refined profiles generated by the AI-crew, and then put them into ten clusters. Evaluation is carried out again on the consistency between the AI-Editor-generated clusters and preliminary profiles.

3 Results

Given the computation load of GenAI, in this study, we selected 50 students with manually-labeled preliminary profiles (five students in each profile) to explore the multi-AI agent application. Refer to Table 2 for the descriptions of the preliminary profiles, modified from those in (Guo et al., 2024b) based on explainable features introduced in this study.

To explore the multi-AI agent approach and prepare for the next-step teacher co-design, we used

Name	Description	Interpretation
Total score	Sum of item scores	Value range (discrete): [0, 21]. High value: good performance; low value: low performance. Most important feature, affecting interpretation of others.
Total time	Sum of item response times	Value range (continuous): (0, 1800]. Low value: less engaged; high value: issues in time management. Important feature, affecting interpretation of others
Total visit	Sum of item visits	Value range (discrete): [1, 113]. A student visiting all items just once has a value of 14. Low value: few item visits/less engaged; high value: issues in behavior regulation. Important feature, affecting interpretation of others
Not-reached (NR)	Number of not-reached items (no response time)	Value range (discrete): [0, 13]. High value: worked on few items; low value: worked on many items. Speeded: if non-zero NR & high time.
Rapid response (RR)	Sum of rapid-responded items	Value range (discrete): [0, 114]. Low value: less engaged; high value: issues in time management. RR: likely not spent adequate time to understand/work the item and associated with low effort.
Prolonged time(PL)	Sum of items with prolonged times (over 95 percentile)	Value range (discrete): [0, 4]. High value: likely struggling on high number of items. Non-zero value may indicate struggling, mostly due to lack of knowledge and skills to solve the problem(s), and subsequently likely led to NR items (i.e., test is speeded).
Navigation Regularity (Reg)	A measure to show whether a student mostly followed the order of item presentation on the test	Value range (continuous): [0.29, 1]. High value: orderly navigation through items (value of 1 indicates always moving forward or backward one item a time; no skipping around); low value: irregularly navigated through items. Also refer to Burst, scan, and jump related features for context.
Scan	Number of quick item scan behaviors in the entire session. Five seconds or less spent on an item is flagged as a scan behavior.	Value range (discrete): [0, 72]. High value: unregulated scan behaviors; low value: engaged with items (i.e, slow and steady win the race).
Longest scan burst	Number of longest scan behaviors in a burst	Value range (discrete): [0, 20]. High value may indicate global review, especially when its location is high; low value may indicate local review.

Table 1: Main features created and their interpretation.

Label	Description & Preliminary Profile
1	Attempted little to no items. Unengaged group
2	Very Low score, low/regular time, and regular visit behavior. Low engagement with very low performance, navigated through most items with low time
3	Low score, low/regular time, and regular visit behavior. Low engagement with low performance, navigated through most items with low time.
4	Low score, full/regular mixed time, and regular visit behavior. Engaged with low performance, navigated through most items, used mixed strategies
5	Low or very low score, unregulated and/or speeded, with high visit behavior. Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with without speededness
6	Low score, full/regular time with some prolonged item response times. Engaged with low performance, navigated through most items, spent a large amount of time on a small number of items, with or without speededness
7	Medium score, regular time and visit behavior, Medium performing group in all dimensions
8	Medium score, full/regular time with some prolonged item response times, and regular visit behavior. Medium performing, show strategic engagement behaviors (such as strategical response times)
9	High score, regular time and visit behavior. High performing group, expected navigation patterns.
10	Very high score, regular time and visit behavior. Highest performing group, expected navigation patterns

Table 2: The ten preliminary profiles, modified from (Guo et al., 2024b).

Azure OpenAI API (model: GPT-4o-mini; OpenAI, 2024) for its cost efficiency and CrewAI (Moura and contributors, 2024) for its easy implementation.

3.1 Example reports

To illustrate the work by the crew of the mulit-AI agents, we show the final outputs for two students (Student A and Student B).

Student A obtained a perfect score of 21 points, spent a total of 1029 seconds on the test, and visited all the 14 items linearly without any revisits, scan behaviors, or prolonged response times. The navigation regularity index is 1 (refer to the navigation plot in Figure 2). Student A's preliminary profile is 10 (i.e., *Very high score, regular time and visit behavior. Highest performing group, expected navigation patterns*).

Based on these facts of Student A, the AI crew provided the following feedback.

```
Key Strengths:
a. Strong understanding of mathematical concepts
and procedures.
b. Excellent problem-solving skills, reflected in high
scores and structured navigation.

Primary Growth Areas:
a. Continued exploration of advanced mathematical
concepts to prepare for high school.
b. Addressing potential gaps in resources or technology
access that may limit enrichment opportunities.

Critical Recommended Interventions:
a. Cognitive Support:
1. Implement differentiated instruction featuring
advanced math challenges to sustain engagement and
deepen understanding.
2. Introduce project-based learning opportunities that
integrate real-world applications of math.

b. Holistic Support (engagement, learning skills):
1. Establish mentorship programs connecting students
with advanced peers or professionals in STEM fields.
2. Organize community workshops focused on resilience,
stress management, and healthy coping strategies.
```

For Student A, the refined profile from the AI crew is

```
Very high cognitive performance with a perfect NAEP
score. Engages effectively with assessments and manages
time well. Exhibits strong self-regulation in navigation
behaviors. Resilient, supported by community influences.
Prepared for advanced mathematical challenges and
enrichment opportunities.
```

Student B obtained a total score of 2 out of the maximum 21 points, spent maximum total time allowed on the test; visited all 14 items but with 74 total visits. This student had one prolonged response time without any rapid responses. The longest scan burst is 39, and the largest jump is 10, with a navigation regularity of 0.29 (refer to the navigation plot in Figure 3). Student B's preliminary profile is 5 (i.e., *Low or very low score, unregulated and/or speeded, with high visit behavior. Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with/without speededness*).

Based on these facts of Student B, the crew provided the following feedback .

```
Key Strengths:
a. Demonstrated persistence and commitment to
engaging with all test items.
b. Ability to revisit questions, indicating a desire
for clarity and understanding.

Primary Growth Areas:
a. Need for strengthening foundational math skills
and conceptual understanding.
b. Development of effective time management
and test-taking strategies.

Critical Recommended Interventions:
a. Cognitive Support:
1. Implement targeted small group instruction focusing
on foundational math skills through real-world
applications.
2. Introduce structured practice sessions with timed
quizzes to improve pacing and time management skills.

b. Holistic Support (engagement, learning skills):
1. Foster a growth mindset by framing mistakes as
learning opportunities and encouraging reflective
discussions.
2. Create mentorship or peer tutoring programs to
provide emotional support and academic guidance.
```

For this student, the refined profile from the AI crew is

```
Very low cognitive performance in math,
high engagement with all test items, challenges in
time management and self-regulation, potential
struggles with anxiety, demonstrated resilience in
facing academic tasks, requires targeted support for
foundational skill development and emotional
resilience.
```

As shown in these examples, the outputs from the AI crew greatly enriched the interpretation of the student preliminary profile with depth, nuance, and the whole person learning perspective. These outputs will serve as a starting point for us to communicate with real teachers to collaborate on creating meaningful and actionable data insights for professional training.

3.2 Evaluation

We experimented several sentence embedding techniques, but results were unsatisfactory, mainly due to the fact that the available sentence embedding models in NLTK, without additional manipulations, did not differentiate the degree of importance for different features (e.g., Total Score has the upmost importance in profiling). Because of the space limit, results from the embedding approach are not presented.

Results from the AI-editor are presented in Table 3, that compares the clusters from the AI agent's analysis and the preliminary profile labels.

From Table 3, we observed that only one student's cluster was not consistent with the preliminary profile. That is, this student's preliminary

Table 3: Contingency Table for Preliminary Profile (Label) and AI-editor's Cluster (Cluster ID)

	Cluster ID										
Label	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**All**
1	0	0	0	0	0	0	0	0	0	5	5
2	0	0	0	0	0	0	0	0	5	0	5
3	0	0	0	0	0	0	0	5	0	0	5
4	0	0	0	0	0	0	5	0	0	0	5
5	0	0	0	0	0	5	0	0	0	0	5
6	0	0	0	0	4	1	0	0	0	0	5
7	0	0	0	5	0	0	0	0	0	0	5
8	0	0	5	0	0	0	0	0	0	0	5
9	0	5	0	0	0	0	0	0	0	0	5
10	5	0	0	0	0	0	0	0	0	0	5
All	5	5	5	5	4	6	5	5	5	5	50

profile (*6: Engaged with low performance, navigated through most items, spent a large amount of time on a small number of items, with or without speededness*) was classified into the adjacent preliminary profile by AI-Editor based on the AI Crew description (i.e. *5: Low or very low score, unregulated and/or speeded, with high visit behavior. Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with without speededness*). The major difference between these two preliminary profiles resides in navigation regularity, while students in Preliminary Profile 6 showed slightly better navigation behaviors (e.g., a higher value of Navigation Regularity Index). This discrepancy of one student's profile in Table 3 may indicate that it is challenging for AI agents to differentiate these two preliminary profiles.

4 Discussion and Conclusion

As AI continues to transform education and assessment practices, the current study explores the opportunity of using multi-AI agents to enhance, accelerate, and innovate measurement research to support education.

This multi-AI agents approach allows for rapid, low-cost exploration of diverse viewpoints, facilitating the identification of areas for deeper, evidence-based discussion with classroom teachers, as well as potential shortcomings in the research design. The outputs from the AI crew helps to develop a better teacher co-design study that aims at providing meaningful and actionable feedback to teachers from the big and rich LSAs' multi-source data.

In this multi-AI agent exploration, we found that AI crew (agents of teachers, math coach, and researcher) were able to enrich feedback, and their narratives were likely to be more grounded in teachers' knowledge and classroom practice than those preliminary profiles from research findings. This would help us to move one step closer to the teacher co-design to bridge the gap between assessment research and teacher practice.

There are a few observations worth mentioning in this exploration. First, even though we asked teacher agents in the crew to consider student data (features and preliminary profiles), we observed that final outputs still contain speculation, without data evidence, on why certain behaviors occurred on the assessment. For AI agents to generate factual profiles, we ended up with requiring every AI agent to refer back to student data, to ensure the final narratives were anchored in empirical data. That is, we used the empirical student data as a guardrail for AI agents to generate outputs. Another observation is the discrepancy between preliminary profiles and AI editor's analysis, which indicates that features, feature mapping, as well as AI agents' persona instructions in this study need to be improved. If AI agents have difficulties to differentiate some student profiles, they are likely to be challenging to real classroom teachers. This multi-AI agent exploration offers opportunities for us to improve our study design.

Overall, this study explored the use of multi-AI agents to prepare and accelerate the process of a teacher co-design for transforming research findings from LSAs to teaching practice. Based on data-driven student profiles obtained from NAEP multi-source data, we assembled a crew of AI agents that mimicked a rigorous human inquiry process to prepare for the teacher co-design. Built on previous studies, we proposed a few innovative approaches to link assessment-data-driven research that uses numerical features to AI agents that use natural languages. Among these innovations explored in this study, explainable feature creation is one of the key steps, which enables the mapping of numerical features into natural languages, and providing empirical data bases for AI agents to reason and produce factual feedback. Features associated with the visual navigation plot, particularly the navigation regularity index, will find wider applications in capturing a behavior process in assessment analytics, learning analytics, and other areas. Most importantly, this set of features will enable gen-

eralization of AI methodologies proposed in this study to other item blocks and even other tests in the future work.

Our exploration showed that the AI crew could enrich feedback and ground it in teachers' knowledge and practice, better preparing researchers for the real teacher co-design. Note that these AI-generated profiles are exploratory in the study. Given the increasing capabilities of GenAI, AI agent uses, evaluation, and validation need further research to empower researchers and educators. In addition, AI outputs in the current study need to be improved further by human experts and teachers. Meaningful understanding and valid insight still require direct engagement with actual teachers and researchers to capture genuine experiences and build trust for impactful assessment research and practice. Such AI systems, built on rich LSA data, research, and teacher co-designs, will be able to promote a more consistent and thorough initial analysis for all students' data, ensuring that feedback includes key factors (cognitive, engagement, learning skills) meaningful to guide teaching professional development with the evolving educational technologies. Collaboration between AI and human experts provides deeper analytical support at a larger scale than might be possible with human expertise alone for education innovation.

Acknowledgments

This work was supported, in part, by the Gates Foundation [INV-086180] and National Center for Education Statistics (NCES). Any opinions expressed in this work are those of the authors alone and shall not be attributed to the Gates Foundation, NCES, or ETS.

References

R. S. Baker and K. Yacef. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

Ryan Baker. 2021. *Artificial intelligence in education: Bringing it all together*, pages 43–54. OECD.

Ali Darvishi, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. Impact of ai assistance on student agency. *Computers & Education*, 210:104967.

Kadriye Ercikan, Hongwen Guo, and Qiwei He. 2020. Use of response process data to inform group comparisons and fairness research. *Educational assessment*, 25(3):179–197.

Kadriye Ercikan, Hongwen Guo, and Han-Hui Por. 2023. Uses of process data in advancing the practice and science of technology-rich assessments. In Natalie Foster and Mario Piacentini, editors, *Innovating Assessments to measure and support complex skills*, pages 211 – 228. OECD Publishing.

Kadriye Ercikan and James Pellegrino. 2017. *Validation of score meaning in the next generation of assessments: The use of response processes*. Routledge, New York, NY.

Samuel Greiff, Christoph Niepel, Ronny Scherer, and Romain Martin. 2016. Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61:36–46.

H. Guo, M. Johnson, K. Ercikan, L. Saldivia, and M. Worthington. 2024a. Large-scale assessments for learning: A human-centered AI approach to contextualize test performance. *Journal of Learning Analytics*, 11(2):229–245.

H. Guo, M. Johnson, L. Saldivia, M. Worthington, and K. Ercikan. 2024b. Human-centered ai for discovering student engagement profiles on large-scale educational assessments. *Journal of Measurement and Evaluation in Education and Psychology*, 30(12):282–301.

Hongwen Guo and Kadriye Ercikan. 2021. Differential rapid responding across language and cultural groups. *Educational Research and Evaluation*, 26(5-6):302–327.

International Association for the Evaluation of Educational Achievement. 2020. TIMSS 2023 international databasee.

Khan Academy. 2025. Keeping your streak alive: insights + tips from the last 6 months.

João Moura and contributors. 2024. CrewAI. Software library. Accessed: June 3, 2025. Please update year and version based on your usage.

National Assessment Governing Board. 2020. Response process data from the 2017 NAEP grade 8 mathematics assessment. Technical report, National Assessment Governing Board. Last accessed on June 2, 2025.

OpenAI. 2024. Generative pre-trained transformer 4 omni (gpt-4o). Model variant: GPT-4o mini. Accessed via Microsoft Azure.

Organisation for Economic Co-operation and Development. 2020. PISA 2018 database.

E. Pools and C. Monseur. 2021. Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assess Educ*, 9(10).

Steve Ritter, Ambarish Joshi, Stephen E. Fancsali, and Tristan Nixon. 2013. Predicting standardized test scores from cognitive tutor interactions. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, pages 169–176, Memphis, Tennessee, USA. International Educational Data Mining Society.

Danielle R Thomas, Conrad Borchers, Sanjit Kakarla, Jionghao Lin, Shambhavi Bhushan, Boyuan Guo, Erin Gatz, and Kenneth R Koedinger. 2025. Does multiple choice have a future in the age of generative ai? a posttest-only rct. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 494–504, New York, NY, USA. Association for Computing Machinery.

G. Zheng, S. E. Fancsali, S. Ritter, and S. Berman. 2019. Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics*, 6(2):153–174.

Nathan Zoanetti and Patrick Griffin. 2017. Log-file data as indicators for problem-solving processes. In Beno Csapo and Joachim Funke, editors, *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, chapter 11. OECD Publishing, Paris.