Realizing the Potential of Automatically Scoring Three-dimensional Assessments

Xiaoming Zhai[1], Joe Krajcik[2], Yue Yin[3], Christopher J. Harris[4], Namsoo Shin[2], Ehsan Latif[1], David McKinney[4]
[1]University of Georgia; [2]Michigan State University; [3]University of Illinois Chicago; [4]WestEd

Project Overview

- Recent advances in technology-enhanced assessments, including automated scoring and artificial intelligence (AI), have the potential to support teachers in interpreting and using assessment information.
- Realizing the *Potential of Automatically Scored Three-Dimensional Assessment*, or PASTA, has been the central goal of our multi-institutional effort to develop an automated scoring system within an online platform that will analyze and display performance results and provide instructional guidance for teachers.
- This project aims to assist middle school science teachers in using technology-enhanced assessment tasks in an instructionally supportive role where assessment is integrated seamlessly into science lessons.

Our Approach

Our main approach is to design with the needs of teachers and students in mind so that the advances in AI can support teachers in bringing assessment and instruction together for improved learning outcomes. We achieve this by:

- Developing automatically generated student reports (AutoRs) for three-dimensional (3D) science performance that enables noticing, attending to, and interpreting student response information in an instructionally useful way
- Developing next-step instructional strategies to improve teachers' use of student response data that comes from the assessment and improve student learning
- Iteratively studying the effectiveness of AutoRs and the instructional strategies to support teachers' decision-making and student 3D learning
- Field testing the PASTA System in a variety of classrooms to study its impact.

AutoR Design

We used a conceptual framework for teacher-centered AutoR-based decision-making. The components of the framework include:

Cognitive demand – the amount and level of integration of the information in the AutoR, which includes three elements:

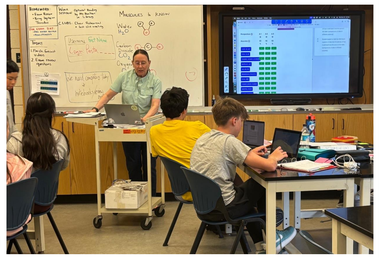
- The content presented in the report
- The synthesis level of presented information
- The depth of data mining.

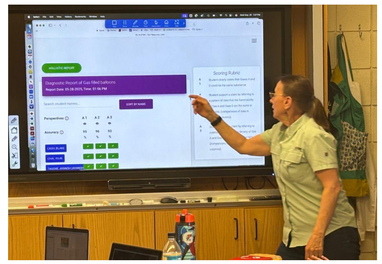
Human-centered design support – assist teachers in interpreting and using the information. We identify two elements:

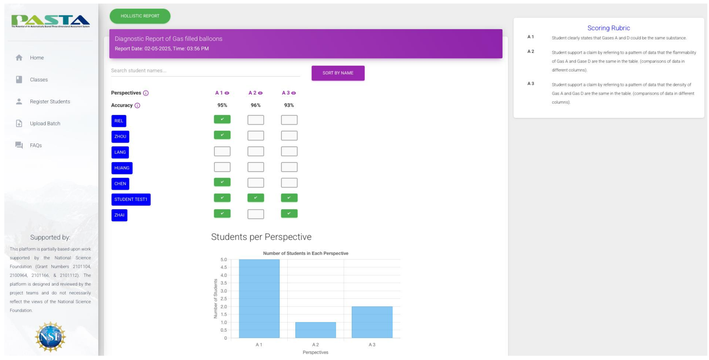
- User functionality
- Information presentation.

Teachers provided feedback on the AutoRs that was used to revise the interface. The interface presents three types of information:

- Individual student performance
- Grouping information
- Instructional strategies suggestions.







Instructional Strategies to Support Students' Learning

Based on student performance, teachers can select a variety of strategies to support students' learning. Dr. Peng He, Western Washington University, oversaw the design of the instructional strategies.

First-Hand Experiences

- Provide a meaningful direct context for students to link new ideas about characteristic properties from the experience to their prior knowledge. Use when most students do not show proficiency in understanding characteristic properties.

Second-Hand Experiences

- Provide a meaningful indirect context for students to find patterns in data and draw conclusions. Use when most students need support with CCC and SEP, lab materials are unavailable, or limited instructional time to do first-hand experiences

Simulations

- Allow students to manipulate objects and explore multiple variables' interactions, which may help them discover various data patterns depending on variables' interactions. Use when most students need support with the CCC and SEP, and direct activities that are not possible in real life because they are high-risk and challenging, or limited time.

Multimodal Experiences

- Provide opportunities for students to access diverse content through various means of learning. Use when students need more in-depth information about the DCI or SEP, access information not found in textbooks or labs or assign homework when instructional time is limited.

Case Study

Use of PASTA Platform

- Teachers aligned the tasks to lessons and then used them as part of their instruction. Based on the AutoRs, they then selected a follow-up instructional strategy.
- They found the AutoRs helpful for understanding students' difficulties with learning goals. For example, one teacher noted, "We had been reinforcing the ideas of claim, evidence, and reasoning all year. With the PASTA tasks, I noticed my students still struggle with reasoning..."

Value of Holistic Group Report

- Teachers responded that they found the holistic group report useful because it clearly shows overall class performance, helping them understand their students' strengths and weaknesses in relation to the learning goals.

Perceptions about the Instructional Strategies

- Teachers found the strategies provide clear instructions, were appropriate for the grade level, and aligned with the Standards. Most teachers used hands-on activities or teacher demonstrations to help students visualize the concepts addressed in the tasks. Some teachers found it helpful to modify the instructional strategies and suggested providing the strategies as editable Google Docs for easier use.

Automatic Scoring Development

Comparing types of natural sugars

Amy wants to know if the sugar found in honey, milk, sugarcane, and apples is the same. To find this out, she chemically removed the sugar from each food and recorded the properties of the sugar, as given in the Table 1 below:

Table1 . Characteristic properties of sugar in different foods

Source of Sugar Sample	Density	Solubility in Water	Melting Point
Honey 	1.69 g/cm^3	Yes	103°C
Milk 	1.53 g/cm^3	Yes	202°C
Sugar Cane 	1.59 g/cm^3	Yes	186°C
Apple 	1.70 g/cm^3	Yes	103°C

1: What similarities or differences would you look for in Table 1 to tell whether any of the foods have the same sugar? Explain why

2: What conclusion could you make based on the pattern (similarities or differences) that you found "if any" of the foods have the same sugar?

We developed 3D-based analytic rubrics for nine chemistry tasks from NGSA and are using four tasks/rubrics for classroom testing. Each task has 5 -10 scoring aspects. We used a pre-trained natural language processing model – BERT on scientific content to achieve SciEdBERT. We further fine-tune SciEdBERT using students' written responses for each task to get a binary text-classification task for automatic scoring. Here, we show an example of an assessment task, its analytical rubric and trained Model's accuracy.

Example Student Response		Sugar from the honey and the apple could be the same substance, since they have almost the same density, are both soluble in water, and have the same boiling point.	
Dimensions	**ID**	**Elements**	**Score**
DCI	E1	Student clearly indicates that the sugar from the apple and honey could be the same substance/type of sugar.	1
SEP+CCC	E2	Student supports a claim by referring to a pattern of data that the density of sugar in honey and apple are the same in the table. (The difference in density might due to measurement error)	1
SEP+CCC	E3	Student supports a claim by referring to a pattern of data that the melting points of sugar in honey and apple are the same in the table.	0
DCI	E4	Student indicates density is one of the characteristic properties to identify substances.	0
DCI	E5	Student indicates the melting point is one of the characteristic properties to identify substances.	0

Question		
Comparing types of natural sugars (ID #5157)		
Aspects	**% of student's correct human scoring**	**% Accuracy**
E1	31.93	84.61
E2	30.47	87.23
E3	44.61	92.42
E4	9.42	85.63
E5	10.1	83.41

Lessons Learned

- Iterative design with teachers-in-the-loop helps improve models, rubrics, and strategies
- Auto-analysis of students' scientific reasoning is hard but there is no easy solution
  - Precise rubrics and accurate human scoring are essential
  - Balanced datasets are crucial
  - Iteratively check models, rubrics, tasks, and student response analyses to improve models
- Scaling is difficult; no universal model works across tasks and populations.
  - Large, balanced datasets are needed across different sites
- Fine-tuning ChatGPT allows model testing but doesn't provide downloadable models for future use
- AI's predictive nature means perfect accuracy remains elusive
- Teachers are highly willing to use such systems.

This project is made possible by National Science Foundation grant Awards: 2101104 (PI Zhai), 2100964 (PI Krajcik), 2101112 (PI Harris), 2101166 (PI Yin). Any findings, conclusions, or opinions herein represent the views of the authors and do not necessarily represent the views of personnel affiliated with the National Science Foundation.