




ARTICLE



<https://doi.org/10.1057/s41599-024-03499-z>

OPEN

FEW questions, many answers: using machine learning to assess how students connect food-energy-water (FEW) concepts

Emily A. Royse¹, Amanda D. Manzanares², Heqiao Wang³, Kevin C. Haudek⁴, Caterina Belle Azzarello², Lydia R. Horne⁵, Daniel L. Druckenbrod⁶, Megan Shiroda⁷, Sol R. Adams⁸, Ennea Fairchild⁹, Shirley Vincent¹⁰, Steven W. Anderson¹¹ & Chelsie Romulo ¹²✉

There is growing support and interest in postsecondary interdisciplinary environmental education, which integrates concepts and disciplines in addition to providing varied perspectives. There is a need to assess student learning in these programs as well as rigorous evaluation of educational practices, especially of complex synthesis concepts. This work tests a text classification machine learning model as a tool to assess student systems thinking capabilities using two questions anchored by the Food-Energy-Water (FEW) Nexus phenomena by answering two questions (1) Can machine learning models be used to identify instructor-determined important concepts in student responses? (2) What do college students know about the interconnections between food, energy, and water, and how have students assimilated systems thinking into their constructed responses about FEW? Reported here is a broad range of model performances across 26 text classification models associated with two different assessment items, with model accuracy ranging from 0.755 to 0.992. Expert-like responses were infrequent in our dataset compared to responses providing simpler, incomplete explanations of the systems presented in the question. For those students moving from describing individual effects to multiple effects, their reasoning about the mechanism behind the system indicates advanced systems thinking ability. Specifically, students exhibit higher expertise in explaining changing water usage than discussing trade-offs for such changing usage. This research represents one of the first attempts to assess the links between foundational, discipline-specific concepts and systems thinking ability. These text classification approaches to scoring student FEW Nexus Constructed Responses (CR) indicate how these approaches can be used, in addition to several future research priorities for interdisciplinary, practice-based education research. Development of further complex question items using machine learning would allow evaluation of the relationship between foundational concept understanding and integration of those concepts as well as a more nuanced understanding of student comprehension of complex interdisciplinary concepts.

A full list of author affiliations appears at the end of the paper.

Introduction

Many global problems are considered “wicked” in that they integrate complex systems that are often studied in distinct disciplines (Balint et al., 2011). To solve these 21st-century socio-ecological problems, students must instead learn cross-cutting concepts across disciplines within interdisciplinary programs. The Next Generation Science Standards (NGSS) identify crosscutting concepts as a framework to link different science disciplines, providing a means for students to link knowledge across fields to establish a cogent, scientifically-based way of interpreting the world (National Research Council, 2012; NGSS Lead States, 2013). In environmental programs within higher education, recent efforts are defining the key disciplinary ideas, concepts, practices, and skills embedded in complex meaningful learning and implementing new curricula with interdisciplinary frameworks (*Global Council for Science and the Environment*, n.d.; Vincent et al., 2013). Frameworks that link concepts across disciplines can include the Sustainable Development Goals (SDGs; Education for Sustainable Development), Resilience Thinking, the United Nations Principles for Responsible Management (UN-PRME), and the Food-Energy-Water Nexus (FEW Nexus) (Leah Filho et al., 2001; Martins et al., 2022). Interdisciplinary approaches to curricula and course design rely on content mastery and skill development to understand systems interactions and higher-order thinking. With this shift toward higher-level learning across interdisciplinary environmental and sustainability (IES) programs, we need new assessments that elicit complex student thinking and can be used to identify and categorize different levels of understanding, not just memorization of facts (Laverty et al., 2016; J. W. Pellegrino et al., 2013; Underwood et al., 2018).

Assessing crosscutting, interdisciplinary learning is challenging, and often constructed responses (CR) (i.e., open-ended questions) are used for assessing interdisciplinary connections because student thinking and reasoning are more explicit compared to multiple choice type questions; however, these CR assessment items are challenging and time-consuming to design and grade. One rapidly developing tool with the potential to support this kind of assessment is text classification models, which are machine learning (ML) algorithms and statistical models that learn from and analyze data patterns. Due to the challenges of assessing interdisciplinary learning, IES programs provide a useful context for education research on the application of these types of ML for studying complex CR assessment items. Further, technology such as ML may help us evaluate these complex formative assessments and provide an opportunity to improve science teaching and learning (Harrison et al., 2023). Often in science assessment, each individual model is specifically developed for each question and response set in an iterative process using human coding and model development and selection methods, making this process potentially very time-intensive (Brew and Leacock, 2013). However, once a model is constructed, it can be used to score many responses very quickly, thus addressing the labor and time-intensive aspects of evaluated CR questions to allow for big data research using those specific questions and associated models. While this trade-off between model development and model use is an important consideration, the process of model development itself can be aided by several considerations, which may speed development time and improve the validity of the final model (Rupp, 2018). Thus a well designed process for developing and evaluating both questions and models is essential, although iterations throughout the process will always be necessary.

Here, we report on a process of using human-scored responses to construct ML-based text classification models for assessing CR questions focused on the food-energy-water (FEW) Nexus. As

part of this focus on the model development process, we address two research questions (1) Can machine learning models be used to identify instructor-determined important concepts in student responses? (2) What do our students know about the interconnections between food, energy, and water, and how have students assimilated “systems thinking” into their constructed responses about FEW?

Systems thinking as an example of cross-cutting concepts.

Systems thinking involves understanding the interdisciplinary connections and relationships between associated components within a system, rather than simply focusing on discrete concepts (Meadows, 2008). The Global Council for Science and the Environment’s (GCSE) draft proposal for key competencies in sustainability higher education identifies systems thinking as a core skill and includes increasing complexity across scales in their definition as the foundation for strategic solution development and future thinking (Brundiers et al., 2023). This level of understanding typically falls on the higher levels of Bloom’s Taxonomy of knowledge that include categories such as “apply”, “analyze” and “evaluate” (Bloom and Krathwohl, 1956; Krathwohl, 2002). Systems thinking is a key competency in STEM education, both in discipline-specific and interdisciplinary reasoning (Blatti et al., 2019; Hmelo-Silver et al., 2007; Mambrey et al., 2020; Momsen et al., 2022; Ravi et al., 2021; Redman et al., 2021; Redman and Wiek, 2021), and is recognized as a core competency by the National Science Foundation (NSF), the National Academies of Sciences, Engineering, and Medicine (National Science Foundation, 2020), and the US Next Generation Science Standards for K12 education (NGSS Lead States, 2013). Systems thinking was recently identified as a key competency by IES educators in higher education (Vincent et al., 2013), where understanding complex natural and social systems is applied and evaluated using systems thinking (Clark and Wallace, 2015; Varela-Losada et al., 2016).

While fostering systems thinking remains challenging, many potential strategies exist to help anchor student learning. Assessment of systems thinking is challenging and typically is approached from the context of the subject matter (see Randle and Stroink, 2018; Grohs et al., 2018; Gray et al., 2019; Bustamante et al., 2021; Liu, 2023; Dugan et al., 2022 and references within), which means there is not one agreed upon definition or assessment for systems thinking. For example, Soltis and McNeil (2022) have developed a systems thinking concept inventory specific to Earth Science, but valid and reliable approaches for measuring learning gains associated with systems thinking more broadly or in other applications are currently lacking. However, within the field of interdisciplinary environmental programs, there is a widely accepted definition of systems thinking from Wiek et al. (2016) for complex problem-solving for sustainability and commonly accepted concepts associated with systems thinking from Redman and Wiek (2021) (Box 1) and the 2021 NAS report on Strengthening Sustainability Education. Assessing systems thinking can be thus understood in the context of how it is integrated within a particular concept or set of concepts.

The FEW Nexus provides a concrete concept integration framework for developing the skill of systems thinking that applies across many interdisciplinary environmental programs as it connects complex environmental processes, management, policy, and socioeconomics of FEW resources (Smajgl et al., 2016). The FEW Nexus is a coupled systems approach to research and global development that accounts for synergies and trade-offs across FEW resource systems (D’Odorico et al., 2018; Leck et al.,

Box 1. | Systems thinking in the context of sustainability (Redman and Wiek, 2021)

Ability to collectively apply modeling and complex analytical approaches: (1) to analyze complex systems and sustainability problems across different domains (environmental, social, economic) and across different scales (local to global), including cascading effects, inertia, feedback loops, and other system dynamics; (2) to analyze the impacts of sustainability action plans (strategies) and interventions (how they change systems and problems).

2015; Simpson and Jewitt, 2019). For teaching and learning contexts, the FEW Nexus provides a scaffold for incorporating systems thinking and sustainability concepts into courses and across curricula. With global resource consumption outpacing supply, the FEW Nexus is a global priority area for research (Katz et al., 2020; Simpson and Jewitt, 2019). Understanding the FEW Nexus and the global focus on FEW research and decision-making makes it an ideal concept for exploring complex systems content in introductory IES courses, as FEW resource systems are visible to learners in their daily lives. Students need to develop their systems thinking to fully grasp the importance of the FEW Nexus and how it is impacted and impacts other systems, e.g., climate change, resource scarcity (Brandstädter et al., 2012).

The need for tools to assess interdisciplinary systems thinking.

Given the complexity of the relationships within the FEW Nexus and the relatively recent expansion of college-level IESs that incorporate FEW Nexus concepts, assessments that target these more advanced systems-level relationships are lacking. Assessing student conceptual understanding typically requires constructing valid and reliable tests, such as concept inventories (CIs) (Hestenes et al., 1992; Libarkin and Anderson, 2005; Libarkin and Geraghty Ward, 2011; Soltis and McNeal, 2022; Stone et al., 2003; Tornabee et al., 2016). Disciplinary CIs are traditionally used to assess learning using close-response questions (i.e., multiple choice). Existing CIs are inappropriate to assess complex skill development in IESs for two reasons: (1) IESs are interdisciplinary, and existing CIs do not capture the range of concepts typically covered in IES curricula, and (2) Close-ended questions (multiple choice) limit the ability to dissect higher level learning, such as systems thinking. An interdisciplinary, open-ended environmental CI could address these challenges; however, CR or open-ended assessments are labor-intensive to evaluate and can be very subjective for instructors to score. Artificial intelligence (AI) attempts to mimic human intelligent actions, including understanding language via Natural language processing (NLP) and classifying artifacts via ML. In the case of CIs, these approaches (NLP and ML) have been used to classify student written assessments and show promise for use with the first interdisciplinary environmental CI that enables assessment of deeper skill development (i.e., systems thinking, cause and effect, tradeoffs) while alleviating the burden of scoring CR questions. Few studies report on the use of interdisciplinary assessments in STEM (Gao et al., 2020), and this dearth of assessment tools also leads to little research about AI-based applications for such assessments (Zhai et al., 2020a, 2020b). The work presented here is a start towards developing assessments (like CIs) that use CR for more complex concepts, such as systems thinking and connecting concepts across disciplines. Here, we focus on FEW as it is a system that incorporates concept integration that connects environmental processes, management, policy, and socio-economics of FEW resources. There is a need for education research and collaboration in the FEW Nexus, as evidenced by the recently funded National Collaborative for Research on Food, Energy, and Water Education (NC-FEW), of which author Romulo is a member. FEW concepts are commonly covered in introductory environmental courses (Horne et al., 2023), and this

project will focus on IES introductory courses for this process of development.

Text classification: using machine learning processes for interdisciplinary assessment.

AI has been part of computer science for a number of decades, with the goal of having computers mimic human intelligence in performing complex tasks. AI utilizes approaches from several different computational subfields in computer science depending on the intended use or task performed. NLP is a branch of computer science that is interested in how computers can identify, understand, and support human language. NLP has become foundational for many AI applications, including speech recognition, language translation, and chatbots. NLP has been incorporated into education contexts in a variety of ways, including scoring of student texts, in both summative and formative uses (McNamara and Graesser, 2011; Shermis and Burstein, 2013), intelligent agents for interactive feedback (Chi et al., 2011), and customization of curricula materials and assessments (Mitkov et al., 2006). NLP has been applied in science assessment in a variety of ways. For example, NLP coupled with ML techniques has been used to develop predictive scoring models (Nehm et al., 2012), as an approach to explore sets of student responses (Zehner et al., 2015), and to assist in developing coding rubrics (Sripathi et al., 2023). Here, we focus on using NLP as part of text classification approaches to categorize student CR to assessment items (Dogra et al., 2022). Specifically, these text based CRs are short in length but rich in disciplinary content and common in STEM assessment practices (Liu et al., 2014). Using approaches from AI, these CRs can be automatically categorized according to coding rubrics that are developed with assessment items (Zhai et al., 2021a).

Machine learning has been described as a “computer program that improves its performance at some task through experience” (Mitchell, 1997). “Experience” here refers to some information (e.g., outcomes, labels) available to the program from which it can “learn.” Much of the recent work on automated scoring of student CR has utilized supervised ML approaches, which use text representations from NLP along with assigned human codes as input for text classification models (Zhai et al., 2020b). Generally, in supervised ML, these data are used to “train” ML algorithms in order to develop a scoring (or classification) model. Once the scoring model is developed, the model can be “tested” by comparing the consistency of human and machine-assigned codes on subsets of the same (or new) data (Jordan and Mitchell, 2015; Williamson et al., 2012). Various ML scoring approaches have been used to evaluate student CRs in science; these reports cover a range of grade levels and disciplinary topics (Jescovitch et al., 2021; Liu et al., 2014; Nehm et al., 2012; Wilson et al., 2023), such as the water cycle in secondary science (Lee et al., 2021). These studies and others have identified important considerations when designing assessment items, rubrics, and text classification models for evaluating responses to science CR assessments. Using these ML approaches in automated assessment scoring, important student ideas can be recognized by machines from authentic student work, as opposed to predefined answers. This is important to identify these key ideas as actually expressed by students. Thus a collection of student responses are

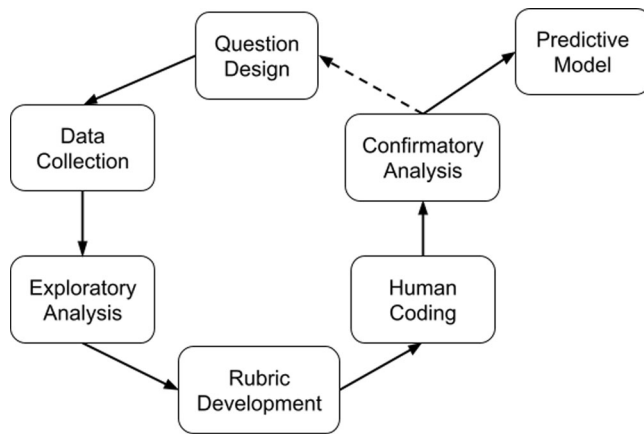


Fig. 1 A Question Development Cycle describes the general process of developing assessment items and associated predictive machine learning models. Adapted from Urban—Lurain et al. (2015), each box represents a stage of the process beginning with Question Design with outputs from one stage being used in subsequent stages, as indicated by solid arrows. A predictive model in the top, right corner is the ultimate goal of the cycle, in which a machine learning model can accurately predict classifications of new responses. A dashed arrow represents possible iteration(s) of the cycle depending on the outcomes of previous stages.

necessary to train the ML model and to represent the range of possible answers (Shiroda et al., 2022; Suresh and Guttag, 2021).

Methods

Overall, we follow a modified question development cycle (Urban—Lurain et al., 2015) (Fig. 1) that integrates question, rubric, and text classification model scoring as part of an integrative formative assessment development and validation process. Broadly, this approach uses linguistic feature-based NLP methods (Deane, 2006) to extract linguistic features from writing and then uses those extracted features as variables in supervised ML models that predict human raters' scores of student writing.

In the first stage of the cycle, we begin with Question Design (top) to target student thinking about important interdisciplinary constructs. Data Collection is typically done by administering the questions online to a wide range of students within appropriate courses and levels to collect a diverse range of responses. Exploratory analysis combines automated qualitative and quantitative approaches to the student-supplied text, including NLP, to explore the data corpus. For example, we use text analysis software to extract key terms and disciplinary concepts from the responses and look for patterns and themes among ideas. These terms, concepts and themes are used to assist Rubric Development. We use rubrics, both analytic and holistic, to code for key disciplinary ideas or emergent ideas in responses. These coding rubrics are subsequently used during the Human Coding of student responses in which one or more experts assign codes or scores to student responses. During Confirmatory Analysis, we develop text classification models by extracting text features from student responses using NLP approaches. These text features are subsequently used as independent variables in statistical classification and/or ML algorithms to predict expert human coding of responses, as part of supervised ML. In this stage, the performance of the ML model is measured by comparing the machine-assigned score to the human-assigned score. Once benchmarks for sufficient performance are achieved (Williamson et al., 2012), the model is saved and used as a Predictive Model. These Predictive Models can be used to completely automate the scoring of a new set of responses, predicting how experts would categorize or score the data. Often,

results from one or more stages of the cycle are used to refine the assessment question (dashed arrow), rubrics, and/or human coding. The overall process is highly iterative, with feedback from each stage informing the refinement of other components. Further, the iterative cycle allows considerations for automated scoring to be addressed throughout the cycle, providing opportunities to collect and examine valid evidence (Rupp, 2018).

Concept identification. In previous work, we performed content analysis on IES course materials collected from 30 institutions to identify shared learning objectives across IES courses and programs (Horne et al., 2023). We also conducted ~100 semi-structured interviews with undergraduates enrolled in the 10 IES programs used for data collection in this study. From these interviews, we found that students have a broad range of knowledge regarding FEW concepts (Horne et al., 2024; Manzanares et al. in review). We, therefore, sought to create assessment prompts that allowed us to explore a spectrum of student responses about the FEW Nexus. Informed by the previous results of the content analysis (Horne et al. 2023) and student interviews (Manzanares et al. in review), we identified two focal areas for assessment item development related to systems thinking (Box 1): (1) Identifying sources and Explaining Connections between FEW systems, and (2) Evaluating outcomes and Comparing Trade-offs between FEW systems (e.g., water used for food is water not used to create energy). We note that these assessment item topics align with NGSS standards of Systems & System Models (NGSS Lead States, 2013), since students must identify multiple boundaries, components, and connections between components, and they must predict outcomes from alterations in components or connections. We incorporated Bloom's Taxonomy, a classification system for identifying skills that we intend our students to learn (Krathwohl, 2002), to help us scaffold our questions. For example, we recognize that students first must be able to identify sources of FEW and make connections to their environment (Table 1: Sources of FEW and connections: reservoir) before they can understand the trade-offs of gaining a local energy source while losing land for crops (Table 2: Trade-offs systems: biomass energy production). As such we have created questions that align with varying levels of student knowledge regarding the FEW Nexus.

Assessment Items. We developed multiple assessment items targeting comprehension of *Identifying Sources of FEW and Connections and Trade-offs of FEW Systems* using different phenomena (e.g., dams, biomass energy) commonly encountered in IES courses (Table 1). Items about important phenomena in IES courses were presented in relevant disciplinary context and broadly focused on one of the three main foci identified previously. For example, the assessment item about reservoirs is designed to have students identify sources of water and energy usage, then explain how these usages may be connected (Identifying sources of FEW, Connections between FEW systems). Items were structured to contain several sub-parts or prompts to better elicit student thinking, each of which was designed to assess a specific construct. For example, in Table 1, parts A and B of the "Sources of FEW & Connections: Reservoir" item was designed to assess student's ability to identify relevant sources of energy and water resources, while the last sub-question assesses how students understand connections between these sources. Thus, many of these items are multi-dimensional, as they require students to integrate disciplinary knowledge and crosscutting concepts.

Data collection. Higher education institutions were invited to participate in this research from the existing connections of the

Table 1 Constructed response questions.	
Assessment Item (Question)	What the responses tell us about targeted learning outcomes
<p>Sources of FEW & Connections: Reservoir</p> <p>(a) A reservoir like the one pictured above is an artificial lake that stores water. What types of energy does the water in this reservoir possess?</p> <p>(b) Explain how the kinds of energy listed in your previous response could be used for food production by nearby farmers.</p> <p>Trade-offs of FEW systems: biomass energy production</p> <p>Biomass energy production involves growing certain crops and converting them to energy. Corn, in the form of ethanol, is a common source of biomass energy. To increase its energy independence from natural gas, your community decides to convert half of their existing agricultural bean fields to corn biomass crops that will eventually provide energy to the surrounding area. You have taken an environmental course and know that burning natural gas has a greater energy return than burning biomass (e.g., one unit of natural gas requires less energy to produce than one unit of ethanol biomass).</p> <p>(a) You realize that corn requires more water to produce than beans; much of your irrigation water comes from a nearby river and limited rainfall. How would you expect the area's water use to change as a result of this shift from agriculture to biomass production?</p> <p>(b) Trade-offs describe the compromise between positive and negative outcomes of a decision. A trade-off results in something decreasing in return for gains in something else. Describe the trade-offs to food, energy, and water systems in switching from bean farming to biomass (corn) farming</p>	<p>These responses tell us:</p> <p>(1) What students <i>understand</i> about how water can create energy</p> <p>(2) The <i>linkages</i> students <i>explain</i> between water, energy, and food resources</p> <p>These responses tell us:</p> <p>(1) Student capacity to <i>evaluate</i> predictions based on their <i>analysis</i> of cause and effect relationships</p> <p>(2) Student capacity to <i>understand</i> trade-offs in connection to their <i>analysis</i> and <i>evaluation</i> of cause and effect relationships</p>

Table 2 Analytic coding rubric for trade-offs systems: biomass energy production.			
Part A rubric		Part B rubric	
Bin	Brief description	Bin	Brief description
A	Increased water consumption	A1	Less food produced
B1	Water scarcity/not enough water total	A2	More land converted to meet food needs
B2	Generally, less water/decrease in water available	B1	More water use
B3	Less water for other things	B2	Less water available
B4	Change in human behavior in water use	C1	Energy produced (corn creates biomass energy)
C	Water prices change	C2	Energy return on investment (ethanol is less efficient/lower EROI than other sources)
D	Changes to river	C3	Renewable energy, more sustainable energy, lower environmental impact source of energy
E	Impacts to biodiversity/wildlife/ecosystem		

PIs and via an email to the Association of Environmental Studies and Sciences Listserv. Ten institutions were purposefully selected to represent the three primary categories of 4-year colleges according to the Carnegie Classifications of Institutions of Higher Education (Carnegie Foundation for the Advancement of Teaching, 2011) and the three approaches to IES curriculum design outlined by a representative survey of higher education institutions (see Vincent et al., 2013 for further description of the three curricular designs in IES program). The IES program curriculum research conducted by the NCSE found statistical alignment of all undergraduate degree programs in a large, nationally representative sample with one of the three broad approaches to curriculum design (Vincent et al., 2013). Our sample includes representation from baccalaureate colleges (4), master's college and universities (3), and doctoral/research universities (4) and programs/tracks representative of the three approaches to curriculum design—emphasis on natural systems (7), emphasis on societal systems (6), and emphasis on solutions development (4). By selecting programs that represent different types of four-year institutions and the three empirically determined curriculum design approaches, we ensured the inclusion of course materials representative of the diversity of the IES field. We focus on four-year programs for the development of the

NGCI due to resource constraints and the lack of equivalent research on community college IES curriculum design that would allow us to select representative programs. Additionally, community college IES degree programs are designed to either articulate with 4-year degree programs or to prepare students for immediate employment (Vincent et al., 2013).

Student responses ($n = 698$) were collected from introductory IES courses during Fall and Spring semesters from Spring 2022 through Spring 2023 by having students complete the assessment questions pre- and post-course discussion of the FEW Nexus. Demographic information revealed 57.45% identified as female, 4% as non-binary, and the remaining 38.55% as male. Racial and ethnic identities reported were 73.67% white, 5.3% Asian, 4.7% Hispanic/Latino/latinX, 1.78% black or african american, 1.38% american indian or Alaskan native and a majority choosing more than one identity (11.79%). We then added the items in a Qualtrics survey and administered the survey to over 400 IES undergraduates from seven post-secondary institutions across the United States to collect student responses (UNCO IRB#158867-1). Responses were then de-identified for coding to create training and testing data for machine learning.

We surveyed the eight IES instructors who had surveyed their students about the pilot items to collect content validity evidence

Table 3 Expertise level determination for Biomass Question Item Part (a).		
Expertise level	Components	Example
4	A + (B, D, or E)	While the turn to biomass is a more sustainable option the <i>use of fresh water is going to increase drastically</i> to be able to sustain such a change to the energy source. More than likely <i>the local river will have drastic impacts from such dependency</i> upon it especially if it is a dry season for rain.
3	A or B	There will be less water because farmers are needing it to sustain the corn.
2	C	Prices for water will go up.
1	None	I expect the area's water to not be used up as fast from the shift of agriculture.

and feedback on question structure. During this survey instructors were asked to both respond to the question item as if they were a student completing the assignment and then, in a separate survey, instructors were asked questions about the question items in the context of their courses. Instructors indicated that assessment phenomena (e.g., food vs. energy production, and energy flows) were typically covered in their introductory IES courses and the multi-part question structure was accessible to student learners.

Rubric development. Rubric development began by reviewing examples of previously published rubrics that were used in similar assessments and intended for use with automatic scoring (Jes-covitch et al., 2021; Sripathi et al., 2023). We agreed upon a scale that would best represent the students’ varying levels of knowl-edge (Table 2). We created each rubric by first analyzing the range of student answers we had received from the different participating institutions. During this initial review, we used an inductive approach and read student answers to identify common themes that revealed student knowledge regarding food, energy, and water systems and their relationships to each other, to other natural and to human systems. During this process we also re-examined our assessment items and the intended goal(s) of the item and also reviewed instructor responses as examples of “expert responses” and alignment with what types of responses students were providing. This was to ensure that students understood the questions the way we intended and to determine if our questions and, therefore, rubrics would need further alterations. To fully capture students’ knowledge, the majority of the NGCI questions needed to have separate rubrics for each sub-question, i.e., sub-questions A–C would each have their own rubrics. At this stage in rubric development, we relied on the previously acquired instructor responses to define an expert-level answer. Instructor responses were similar, and where there were divergences we identified commonalities across responses. We then compared instructor responses to student responses to create a range of scores reflecting novice to expert knowledge.

We designed dichotomous, analytic rubrics with parallel structures for each node of the FEW Nexus. Each response is categorized based on the ideas it contains, with each response receiving a zero or one score for each code based on the presence or absence of the targeted ideas. We provide an example rubric for parts of the *Trade-offs of FEW Systems: Biomass Energy Production* question item (hereafter referred to as “Biomass question item”) in Table 2, and the other rubrics are available in the Supplemental Methods file.

To determine the level of expertise a student displayed in their response, we defined a certain combination of bins to receive a holistic score of one through four (Table 3). For example, the following student response to Biomass Part A is considered an expert level response (coded as 4) because it contains the following ideas: water usage will increase (bin A) and there will be changes to the local river (bin D). The student therefore makes two connections between energy and water (water usage increasing, impacts to the river).

While the turn to biomass is a more sustainable option, the use of fresh water is going to increase drastically to be able to sustain such a change to the energy source. More than likely the local river will have drastic impacts from such dependency upon it especially if it is a dry season for rain.

Human coding. After the development of the initial rubrics, we iteratively refined the rubrics over several rounds of human coding. During each iteration, two or three researchers separately assigned scores to a set of 30 randomly selected student respon-ses. Each student response received a 0 or 1 for each bin in the rubric for the absence or presence of the corresponding theme in the response. After scoring the set of 30 responses separately, the researchers compared assigned scores and calculated percent agreement. A percent agreement of at least 85% per bin was considered the acceptable level of agreement between human coders to move forward with coding the rest of the dataset independently (80% agreement is acceptable per Hartmann, 1977). The scorers met to discuss agreement for each code; in cases of high percent disagreement, the rubric was revised to improve clarity on those codes. During these discussions, deci-sions about removing or revising codes with very low agreement or low frequency in the dataset were also made. For example, a reservoir code, “Energy needed for food production or irrigation” originally lacked clarification. It was then further described for coders with the addition of, “Irrigation minimum: POWERING the transport/pumping of water, but not implied movement of water without tying to energy. When to code with machinery: machinery + either harvest, produce, or process food.” Specifi-cities like this helped improve coder agreement.

After revising the rubric, a new sample of 30 student responses was compiled, which were independently scored by two to three researchers against the bins with previously high disagreement. This iteration of separate scoring, calculating percent agreement, and revising the rubric continued until scorers either reached 85% agreement for each code in the rubric or resolved remaining disagreements through discussion until consensus was reached (5 iterations for the Biomass question item and 4 iterations for the Reservoir question item). After reaching a consensus for the rubric, all student responses were divided between two members of the research team and were scored independently. A total of 346 responses were scored for the Reservoir question item and 483 responses for the Biomass question item (Supplemental Tables 5 and 6, respectively).

Text classification model development. We employed a super-vised ML text classification approach to assign student written responses a score. During our ML process, each individual stu-dent response was treated as a document and the bins in each scoring rubric were treated as classes (Aggarwal and Zhai, 2012). The predicted output of a ML model is a dichotomous outcome of whether a response would be categorized in each rubric bin or not. We decided to combine student responses for both parts of the *Sources of FEW & Connections: Reservoir* question item

(hereafter referred to as “Reservoir question item”) into a single text response for text classification model development for two reasons. First, the final coding rubric for each part of the question was identical, although certain ideas/bins were expected to be more frequent in one part than the other. Second, the human coding team adopted a similar approach when assigning codes: regardless of in which response part the student included the idea, the human coders marked the code as “present” for the response as a whole. For the Biomass question item, student responses to each part of the question were kept separated during model development since different parts have different coding rubrics (see Table 2).

Text features (single or strings of words) were extracted as n-grams from each response using NLP methods. We used a default set of extraction settings and processing, including stemming, stop word removal, and number removal, to generate a set of text n-grams. The computerized scoring system then generated predictions on whether each given document was a member of each class (i.e., rubric bin) using the extracted n-grams in a bag-of-words approach as input variables in a series of ML classification algorithms. To generate these predictions, we used an ensemble of eight individual machine-learning algorithms (Jurka et al., 2013) to score responses to each question. The predictions of the set of individual algorithms are then combined to produce a single class membership prediction for each response and rubric bin. The text classification, including the ensemble ML model, was generated using a 10-fold cross-validation approach using the Constructed Response Classifier (CRC) tool (Noyes et al., 2020). The CRC has been used previously to score short, concept-based CR even in complex disciplinary contexts and is described in more detail elsewhere (Jescovitch et al., 2021). For evaluation, we compared the machine-predicted score from the ensemble for each response in each rubric category to the human-assigned score for each response.

For each of the models developed in this study, we optimized model performance based on the training set by starting with a default set of extraction parameters, then adjusted several other common model parameters (e.g., n-gram length, digit removal) and retrained classification models to evaluate model performance. This is what we describe as exploratory, basic feature engineering, and we applied a similar approach to every model for each rubric bin. We used the human-coded data for Reservoir and Biomass questions, and we removed several responses with a missing value for a human-assigned score. We used 345 coded student responses for the Reservoir question item, 480 for the Biomass question item Part A, and 466 for Part B as our initial training and testing sets. During this and further iterative rounds, we used common benchmarks of Cohen’s kappa as our targets (kappa > 0.6 as substantial; kappa > 0.8 as “almost perfect” (Nehm et al., 2012). Cohen’s kappa is a measure of agreement between raters (in this case, human and machine) that takes into account chance agreement and is frequently reported in evaluating the overall performance of ML applications to science assessments (Zhai et al., 2021b). We further considered evaluation metrics of accuracy, sensitivity, specificity, F1 score and Cohen’s kappa to guide iterations of model development and to evaluate the overall performance of models once a benchmark was achieved (Rupp, 2018). It is noted that while Cohen’s kappa serves as the primary metric for reporting our model’s overall performance, we also routinely consider other evaluation metrics during model building and evaluation. The assessment of these metrics should not be construed as an all-encompassing validation, as their effectiveness is contingent upon the distribution of scores assigned by humans and the quality of those human scores (Williamson et al., 2012). In our specific context, we encountered

a challenge with the disproportionate representation of certain score points, particularly in some specific analytic rubric bins where cases scored as 1 were significantly fewer than 0. Such low cases of positive occurrences in the training set led to decreased sensitivity metrics for those rubric bins. In some cases (e.g. Reservoir B3), the overall model still exhibited an acceptable overall performance metric and an acceptable F1 score.

Analysis of model outputs and iterative model development.

After performing the initial model development and examining the basic feature turning settings, we examined the outputs of the model for low-performing rubric bins, including model evaluation metrics and groups of responses that showed disagreement between human and machine-assigned scores. We hoped to find possible ways to adjust the model parameters and/or training set of data to improve model performance in subsequent iterations. For example, we collected responses with disagreement in assigned human and machine scores. We examined the false negative and false positive predicted responses (compared to the human coding) in a rubric bin and performed conventional content analysis to try to identify words, phrases, or ideas that were common among these mis-scored responses (Hsieh and Shannon, 2005). We also reexamined the criteria of coding rubrics with low-performing models to ensure the criteria clearly identify important disciplinary ideas and to confirm the original assigned human codes to responses (Sripathi et al., 2023). The coding team met to discuss the results of miscode analysis and changes to target during iterative cycles, including possible changes to the rubric, best approaches to tuning model parameters consistent with assessment items and student ideas, and/or adjusting training sets.

The insufficiency of educational data (Crossley et al., 2016; Wang and Troia, 2023), which often suffers from limited availability of data for training ML models as compared to other sectors, and the observed lack of diversity in undergraduates’ CR (Jescovitch et al., 2021) have long posed challenges for educational researchers. These issues present difficulties for ML algorithms in discerning patterns effectively and reliably identifying a broad range of student ideas. To address these challenges, we have adopted a set of extended model tuning strategies, which have been both theoretically and empirically validated (Bonthu et al., 2023; Jescovitch et al., 2021; Romero et al., 2008). We employed these extended strategies beyond our exploratory, basic parameter tuning (described above). The extended strategies we employed are:

Additional feature engineering. In certain instances, we implemented two advanced feature engineering techniques, often arising to address patterns identified during our miscode analysis. These techniques encompassed (1) substituting specific words with synonyms and (2) extending N-gram analysis to more complex levels, including trigrams (three words combined into one feature) and quadgrams (four words combined into one feature).

Data rebalancing. Training sets that heavily represent only certain types of responses can impede model training; therefore, we applied data rebalancing strategies to address situations where the dichotomous coding significantly favored one category (over three times). When our dataset exhibited such imbalances, we implemented data rebalancing techniques by removing responses associated with the most frequently occurring codes to achieve a more equal distribution of the dichotomous codes. In our data set, cases coded as 0 often outnumbered those coded as 1. Since cases coded as 0 sometimes failed to provide meaningful patterns

for ML algorithms to learn from, we selectively removed excess cases coded as 0 to equalize or enhance the distribution (i.e., reducing the ratio to equal to or less than two times difference).

Dummy responses. For datasets characterized by a balanced distribution of dichotomous scoring codes, yet still yielding low performance metrics, another extended strategy was devised. In this strategy, we initially ensured dataset balance, saved cases with human rater scores, and ML-predicted scores and outputs of the CRC tool after the initial round of analysis. Subsequently, we filtered out responses that were incorrectly classified, identified by a misalignment between human and ML predicted scores. These misclassified cases underwent further qualitative examination, with notes indicating which phrases and segments included in (or absent from) the response were indicative of the critical concept targeted by the rubric. We then generated new cases (i.e., dummy responses), which only replaced the identified segments of responses with new words or phrases, without altering the sentence's underlying meaning. This procedure offers advantages, including mitigating overfitting concerns in which the model is only effective on responses very similar to the training set and augmenting the training dataset's size. The dummy responses were integrated into the overall dataset solely for model training purposes. To derive the final performance metrics of the classifier model, the dummy responses were subsequently removed for model evaluation calculations.

Merging rubric bins. In some instances, despite the explicit indication in the original rubric descriptors that certain ideas are intended to be scored separately as they are designed as mutually exclusive during rubric development, some machine models faced challenges in effectively identifying these subtle textual patterns. Collaborative discussions with expert raters led to a consensus among researchers to combine these rubric bins. This decision was informed by empirical investigation revealing overlapping content, and the re-coding of these bins to a single code / score to enhance the model's performance, aligning with practical considerations in the procedure.

It is important to note that these strategies can be combined or used consecutively as needed. Nevertheless, the initial round of analysis consistently adhered to the default and basic settings of the CRC tool, utilizing the parameter options provided therein. Further details on the application of these approaches to individual items and rubric bins, along with illustrative examples of dummy response creation, can be found in the supplementary materials.

Results

Here, we report on the use of ML-based text classification models to assess CR questions focused on the FEW Nexus. This section is organized by the research question, beginning by describing the successes and challenges in applying ML to score student CR to questions about sources of FEW resources and trade-offs associated with biomass energy production. We then examine the two questions related to reservoirs and biomass to describe FEW connections in student CR and co-occurrences across responses to understand student system thinking capacities. Co-occurrence suggests evidence of systems thinking as multiple FEW systems are interacting simultaneously in student responses.

Research Question (1): can natural language processing be used to identify instructor-determined important concepts in student responses? We developed a total of 11 text classification models for the Reservoir item, one each for the 11 "bins" contained in the coding rubric (Table 1). These eleven models had a

range of overall performance metrics (Table 4), ranging from Cohen's kappa of 0 to 0.957 and accuracies ranging from 0.892 to 0.992. Only one model (D2) failed to detect positive cases, which resulted in an overall Cohen's kappa = 0.000. This was due to a severe data imbalance in the human-assigned codes in this rubric bin, meaning that there were very few positive codes to responses assigned by humans in this bin. All other ten models met acceptable performance levels as measured by Cohen's Kappa values (kappa > 0.6 as substantial; kappa > 0.8 as "almost perfect" (Nehm et al., 2012)), with many models exhibiting "almost perfect" agreement with human assigned codes. We note that most models were tuned to this performance using only basic feature engineering manipulations, as described in the methods. There were also a few bins that met our target threshold of 0.6 only after employing extended strategies (e.g., employing dummy responses for A2), and for one bin (B4), we employed data rebalancing in tuning the model. The model for D2 showed high accuracy but decreased performance on other model metrics due to a severe imbalance of human code occurrence.

One result that emerged from discussions during iterative model development for the Reservoir question item was the similarity of codes A2 (producing hydropower) and B4 (energy transformations). Although we were successful in developing text classification for each code separately, the two models did require slightly different tuning strategies. When examining miscoded responses, the coding team noticed similar patterns in the groups of correctly and miscoded responses for each bin. Human coders reflected that during the coding of the responses, students expressed these ideas similarly, and it was, therefore, sometimes difficult to distinguish when students were explaining hydro-power versus describing transformations of energy (e.g., moving water turning turbines). Thus, these two codes (A2, B4), which were initially intended to capture a specific understanding of hydropower and a more general description of energy transformations, ended up being more similar than intended in the context of this item. One potential way forward for text classification is to combine the A2 and B4 codes into a single code and redevelop a text classification model to recognize the single code.

We developed 15 text classification models (eight models for Part A; seven models for Part B) to detect student ideas in response to the Biomass question item (Table 5). Overall, for this item, models demonstrated lower performance metrics than models for the Reservoir question item. For the Biomass question item, no model achieved a level of almost perfect agreement (as measured by Cohen's kappa value of >0.8), although the majority still achieved acceptable agreement with human scores. Due to the reduced maximal performance, these fifteen models had a narrower range of overall performance metrics than models for the Reservoir question item, ranging from Cohen's kappa of 0–0.674 and accuracies ranging from 0.755 to 0.991. Correspondingly, these Biomass models had a much broader range of sensitivity, specificity, and F1 score metrics too. This reduction in performance metrics is likely due in part to the target of the Biomass item: trade-offs around FEW. Although the item still centers on the FEW Nexus, this item allows students to respond in numerous ways about any number of possible trade-offs between any of the vertices. Thus, this item allows for a much wider possible answer space. As a result, a few models failed to reach the benchmark performance metrics (e.g., B2 in Part A), despite having frequent occurrences of both codes. This also suggests that the text complexity of expressing these ideas or the range of possible ideas in these responses is difficult for these text classification models to reliably identify. Although we attempted extended strategies on models for many of the Biomass models, we report on a few of the bins and attempts as exemplars of this

Table 4 Performance measures of automated classification models for reservoir question item.						
Measures	A1 ^a (N = 345) n (0, 1) = 133, 212	A2 ^b (N = 417) n (0, 1) = 259, 158	B1 ^a (N = 345) n (0, 1) = 225, 120	B2 ^a (N = 345) n (0, 1) = 270, 75	B3 ^a (N = 345) n (0, 1) = 321, 24	B4 ^c (N = 208) n (0, 1) = 132, 76
Accuracy [95% CI]	0.919 [0.885, 0.945]	0.940 [0.913, 0.961]	0.975 [0.951, 0.988]	0.986 [0.967, 0.995]	0.965 [0.940, 0.982]	0.947 [0.907, 0.973]
Cohen's Kappa	0.829	0.838	0.943	0.957	0.687	0.825
Specificity	0.895	0.981	0.978	0.996	0.997	0.977
Sensitivity	0.934	0.873	0.967	0.947	0.542	0.894
F1 score	0.895	0.953	0.980	0.991	0.982	0.959
Measures	C1 ^b (N = 433) n (0, 1) = 230, 203	C2 ^b (N = 363) n (0, 1) = 241, 122	C3 ^a (N = 345) n (0, 1) = 232, 113	D1 ^b (N = 365) n (0, 1) = 256, 109	D2 ^a (N = 345) n (0, 1) = 342, 3	
Accuracy [95% CI]	0.917 [0.887, 0.941]	0.857 [0.816, 0.891]	0.959 [0.933, 0.978]	0.829 [0.785, 0.867]	0.992 [0.975, 0.998]	
Cohen's Kappa	0.833	0.652	0.906	0.906	0	
Specificity	0.935	0.971	0.987	0.953	1	
Sensitivity	0.897	0.631	0.903	0.472	0	
F1 score	0.923	0.900	0.970	0.892	0.996	
The sample size and subsample size in this table pertain to the training data utilized for machine learning training. The original dataset comprises a total of 345 responses. Variations in sample size and subsample size correspond to different rubric bins, reflecting distinct data manipulation strategies employed to enhance model performance.						
- Rubric bins denoted by "a" indicate the use of basic feature engineering settings with no extended strategies.						
- Rubric bins marked "b" signify the introduction of an extended strategy of dummy responses.						
- Rubric bins marked "c" indicate the extended strategy of data rebalancing before introducing dummy responses.						

work, or findings that were similar between different bins. We provide more detail on applied strategies for each model in the Supplemental Materials.

The model for B3 code in Biomass question item Part A showed very low-performance metrics despite having a fair number of positive cases. The poor model performance is likely reflective of the range of student ideas covered by this rubric bin: a decrease in water availability for other uses (here, “other uses” means outside the context of bean and corn agriculture given in the question). As such, there is a wide range of possible other uses students could suggest, such as drinking water, home water use, and water for other crops. The broad range of acceptable answers was easy for humans to code, but difficult for the model to detect the underlying similarity. Although we tried some extended strategies for model iterations, these had little effect on overall model performance. During iterative rounds of model development, we decided to merge two codes, B3 and B4, since they both identified similar ideas, about less water available for other things and changes in human behavior due to less water. During our review of miscoded responses by the model, we noticed a number of miscoded responses were somewhat borderline cases of human code assignment between the two bins B3 and B4, with responses often implying or vaguely mentioning effects on community usage of water, without being explicit the change in use or behavior. For example, the response, “Since this is a place of limited rainfall, and the source of water is coming from the river I would expect that water use for the community may need to be diverted more towards the crops, and less towards other measures such as household use.” was coded positive for B3 by human coders but miscoded as missing B3 by the model. After merging these two codes into a single code and model, the performance of the overall model for the merged code was significantly improved for B3 and slightly decreased for B4 (see Table 5). After merging these bins into a single model, borderline responses, such as the example, were correctly classified by the model.

Similarly, the initial classification model for C1 in Biomass question item Part B failed to meet performance benchmarks even though student responses were nearly equally distributed between positive and negative cases, and we tried several extended strategies to improve model performance. However, the re-examination of coding rubrics for C1 and C2 presented an opportunity to recombine coding criteria as part of the iterative process of using model outputs to iterate on items and rubrics. The rubric was originally designed to identify student ideas about

the production of energy (C1), but not when used in conjunction with trade-offs with other energy sources or energy return on investment (C2). After several rounds of model iteration and discussion with the coding team, we decided to recode the original dichotomous rubric bins C1 and C2 as a single, multi-class code (i.e., a holistic coding rubric, with levels as 0, 1, or 2). This preserved the exclusivity of these two codes (C1 and C2 were intended to be mutually exclusive) while encoding the exclusive classes in the model training set. Making this a single, multi-class prediction increased the overall performance of the model, above the performance for the separate, binary models made for the original rubrics.

Research Question (2): what do our students know about the interconnections between food, energy and water, and how have students assimilated “systems thinking” into their constructed responses about FEW? Here we apply two different strategies for defining and evaluating student responses as novice to expert. To evaluate student knowledge about interconnections and how they have assimilated “systems thinking” into their constructed responses about FEW, we calculated the co-occurrence of codes. Level of expertise for the Reservoir question item is approximated by co-occurrence of the codes, and level of expertise for the Biomass question item is calculated by the code combination provided in Table 3.

Sources of FEW and connections: reservoir question item. We examined the predicted codes for each response to the Reservoir item to look for co-occurrence of codes in student responses. This can help identify connections students are making between FEW vertices, since the item prompts students to make these connections. For this analysis, we collapsed individual bins in Table 6 for the Reservoir rubric by grouping letter codes (e.g., A1 and A2 together as A bin), since these groupings indicate similar themes (A codes refer to hydroelectricity, B codes refer to energy production, C codes refer to use of energy; Table 1 in Supplemental Methods).

Responses frequently included ideas from A codes with ideas from C codes, indicating the same response connected generating hydropower to uses of energy for agriculture or infrastructure. The C codes also commonly occurred with the B codes, showing students explained connections between types of energy and uses of energy in agriculture or community resource use. D codes (uses of water) were the least frequently coded;

Table 5 Performance measures of automated classification models for biomass question item.**Part A**

Measures	A ^a (N = 480) n (0, 1) = 227, 253	B1 ^c (N = 330) n (0, 1) = 223, 107	B2 ^c (N = 307) n (0, 1) = 130, 177	B3 ^b (N = 264) n (0, 1) = 198, 66
Accuracy [95% CI]	0.819 [0.781, 0.852]	0.842 [0.799, 0.880]	0.759 [0.707, 0.806]	0.814 [0.762, 0.859]
Cohen's Kappa	0.637	0.602	0.507	0.359
Specificity	0.824	0.973	0.707	0.990
Sensitivity	0.814	0.570	0.799	0.288
F1 score	0.811	0.893	0.718	0.889

Measures	B4 ^{b,d} (N = 505) n (0, 1) = 381, 124	C ^a (N = 480) n (0, 1) = 384, 12	D ^a (N = 480) n (0, 1) = 412, 68	E ^a (N = 480) n (0, 1) = 455, 25
Accuracy [95% CI]	0.877 [0.845, 0.905]	0.977 [0.959, 0.989]	0.921 [0.893, 0.943]	0.973 [0.954, 0.986]
Cohen's Kappa	0.667	0.151	0.639	0.636
Specificity	0.971	1	0.973	1
Sensitivity	0.589	0.083	0.603	0.480
F1 scores	0.923	0.988	0.955	0.986

Part B

Measures	A1 ^a (N = 466) n (0, 1) = 258, 208	A2 ^a (N = 466) n (0, 1) = 462, 4	B1 ^a (N = 466) n (0, 1) = 254, 212	B2 ^a (N = 466) n (0, 1) = 271, 195
Accuracy [95% CI]	0.805 [0.766, 0.840]	0.991 [0.978, 0.998]	0.839 [0.802, 0.871]	0.779 [0.739, 0.816]
Cohen's Kappa	0.601	0	0.674	0.538
Specificity	0.868	1	0.882	0.672
Sensitivity	0.726	0	0.788	0.856
F1 score	0.832	0.996	0.857	0.718

Measures	C1 ^d (N = 515) n (0, 1) = 246, 269	C2 ^e (N = 180) n (0, 1) = 121, 59	C3 ^a (N = 466) n (0, 1) = 377, 89
Accuracy [95% CI]	0.755 [0.717, 0.792]	0.839 [0.777, 0.889]	0.903 [0.873, 0.929]
Cohen's Kappa	0.512	0.609	0.636
Specificity	0.793	0.942	0.984
Sensitivity	0.721	0.627	0.562
F1 scores	0.756	0.887	0.943

Combined bins

Measures	B1&B2f (N = 480) n (0, 1) = 296, 184	B3&B4f (N = 480) n (0, 1) = 344, 136	C1&C2g (N = 469) n (0, 1, 2) = 184, 219, 66
Accuracy [95% CI]	0.779 [0.739, 0.816]	0.844 [0.808, 0.875]	0.772 [0.731, 0.809]
Cohen's Kappa	0.520	0.572	0.615
Specificity	0.865	0.956	0.772, 0.831, 0.576
Sensitivity	0.641	0.559	0.832, 0.784, 0.988
F1 score	0.828	0.898	0.759, 0.900, 0.697

The sample size and subsample size in this table pertains to the training data utilized for machine learning algorithms employed in the development of the classification model. The original dataset consists of a total of 480 responses in Part A and 466 responses in Part B. Within this table, variations in sample size and subsample size correspond to different rubric bins, reflecting distinct data manipulation strategies adopted to enhance model performance. Strategies are described in more detail in the Methods under "Text Classification Model Development".

- Rubric bins denoted by "a" indicate basic feature engineering only, with no incorporation of extended strategies.
- Rubric bins denoted by "b" indicate the utilization of an extended strategy of advanced feature engineering by defining synonym sets.
- Rubric bins denoted by "c" indicate the extended strategies of data rebalancing before introducing dummy responses.
- Rubric bins marked as "d" signify the extended strategy of dummy responses.
- Rubric bins marked as "e" indicate that we employed the extended strategy of data rebalancing.
- Rubric bins marked as "f" indicate that we employed the extended strategy of merging rubric bins for Part A.
- Rubric bins marked as "g" indicate that we employed the extended strategy of merging rubric bins for Part B.

however, when D was coded, these responses were very frequently connected to hydropower (A codes). Co-occurrence within A codes and D codes suggests that students understand that hydropower is powered by water and is needed to create electricity. A–C codes were the most likely to occur together when students were making connections between FEW systems (59 responses). Only 12 students made connections between A–D codes, suggesting that water use beyond hydropower is not as commonly associated with energy use and production in this scenario despite water providing the primary source of energy in the reservoir.

Co-occurrence is how we can approximate the level of understanding of the respondee from novice to expert for the Reservoir question item. The assumption is that the quantity of co-occurrences indicates students have an understanding that there is some sort of connection between Food, Energy, and Water. For example, student responses could be coded in a number of bins regarding the type of energy, and what the energy is used for, e.g., irrigation or powering homes. We assume that students' answers indicating a greater understanding of the relationships between Food, Energy, and Water will include bin codes for hydroelectricity, irrigation for food, energy for

machinery, and energy for housing/farm bins (Table 7). Novice responses show they know that the dam is used to create hydropower, but they do not have any further knowledge about how this energy can be used and how it relates to food (Table 7).

Trade-offs systems: biomass energy production question item. Overall, students perform at a higher level for explaining changing water usage (Part A) than discussing trade-offs (Part B) (Table 8). The large majority of students discuss at least one trade-off in their response for Part B and, therefore, are placed in level 1 or higher (see Table 9 for an example of student responses). Due to the ML model performance for this question item, we have also included the number of responses for each Novice to Expert Level as Supplemental Methods Table 9.

For the Biomass question item Part A, slightly over half of the responses scored a level 3, with over 20% as Level 4 and about 23% as Level 1, and no level 2 responses (example responses provided in Table 9). For Part B, about half of the responses were grouped in level 2 and roughly 20% in level 1; both of these levels) had similar numbers of responses in those levels by ML and human-assigned codes. A small percentage of responses (~11%) were placed in level 3 by the ML model, while human codes had

slightly more responses (13.5%) in that level. There were no student responses predicted for level 4 by the ML model and only one response in that level based on human-assigned codes.

The lack of level 2 responses for Part A is due to having only one positive ML predicted for the C component. However, this response was scored to level 3 response because the student response also included one of the other codes. Since this was a poor-performing ML model for the C code (meaning that the model did not recognize any responses for this code), we explored using human scores for this code; even so, only three responses from the data set end up at level 2. Most responses in the dataset which are categorized in code C end up at levels 3 and 4, since these responses tend to incorporate water price increase as an effect of increased water use or water scarcity within their explanation (Table 9 provides an example student response).

For Biomass Part B, we found no level 4 responses in our data set, which was driven by the lack of ML predictions for category A2, which is a requirement for obtaining this level. A2 is an infrequent category in the dataset with only 4 positive cases assigned by human coders. Even when we explored using human scores in place of ML-predicted scores for this specific rubric bin, we observed only a single response in Level 4. About one-third of the student responses score at Level 2, which demonstrates an ability to connect at least two FEW vertices when discussing trade-offs. The largest group of students (~40%) end up at Level 1, which is a trade-off focused on a single vertice of the nexus (food, energy, or water).

Of the 138 responses that do not fit the other patterns in Part A and the 77 responses categorized as Level 0 in Part B, most were a combination of derivations of “I don’t know” or trivial responses such as “it will go up” or “You need all food, energy, and water in this situation.” However, there were also responses that the ML model did not predict any expertise level, but would be considered one of the expertise levels by human coders. For example, this student’s response that was not predicted to achieve an expertise level includes concepts that occurred infrequently and, as such, was not provided a code—reduced water availability means that water would need to come from someplace else, and require more labor and cost for transportation:

Table 6 Co-occurrence of predicted codes for reservoir responses.

Code	Total code instances	Number of responses which co-occur with other code			
		A	B	C	D
A—Hydroelectricity	241	N/A	80	157	47
B—Energy production	148	80	N/A	92	30
C—Use of energy	206	157	92	N/A	29
D—Water	69	47	30	29	N/A

Total instances may be less than the sum of the values in a row, since a response can be categorized in any number of rubric bins.

Table 7 Example student responses for reservoir question item co-occurrence.

Novice to expert	Example student response
Expert	(a) I guess potential energy. My assumption is that the water will be turned into hydroelectric energy though once passed through the dam. (b) The energy produced by this might be used by farmers to move other water for irrigation to their crops. It may also be used to power some machinery or even just their homes and facilities.
Expert	(a) Hydroelectric power and as a water resource. (b) The water helps farmers grow their crops but also helps power their farms and machinery.
Novice	(a) hydraulic (b) Don't know

Table 8 ML predicted novice to expert distribution of responses for Biomass.

Level	Components	# of Part A responses	Components	# of Part B responses
4	A + (B, D, or E)	103 (21.4%)	(A2) + (B2) + (C2 or C3)	0
3	A or B	267 (55.6%)	(A1) + (B1) + (C1)	51 (10.9%)
2	C	0 (0%)	Two of the following: (A1 or A2), (B1 or B2), and/or (C1 or C2 or C3)	229 (49.2%)
1	None of the above	110 (22.9%)	One of the following: (A1 or A2), (B1 or B2), or (C1 or C2 or C3)	104 (22.3%)
0		N/A	None of the above	81 (17.4%)
Total		480		465

Due to data rebalancing, not all responses had predicted scores in all categories. In these cases, we used the human score for the given category for a response.

Table 9 Example student responses for the biomass question item using calculations from Table 3.		
Novice to expert level	Example student response to Part A	Example student response to Part B
4	The people living in the area would have to take serious water cuts in order for this plan to work. Watering grass lawns, for example, would likely be banned. We would use more of the river water even still, and in drought years, we might not be fully able to produce biomass for fuel. We would ask people to take shorter showers, water lawns less, if at all, and to not leave faucets running unless essential.	No predicted responses
3	Water usage will likely shoot up as a result of the increased use of corn crops in order to produce more energy via biomass compared to food production for beans that could be facilitated by the same fields.	The switch from beans to corn in order to create more energy independent from natural gas means less food production for the community because of that shift. The switch also requires more water to involved in order to produce more of the corn. The more corn that is produced the more energy that is produced. This energy could help with food in other ways.
2	No predicted responses	More water will be diverted to the corn, and because of that, there will be limited water for other uses.
1	What I would expect the area's water use to change as a result of this shift from agriculture to biomass production by using a non-renewable resource for the process	I think the trade-off in this situation would be an increase[d] in corn production and a decrease in water and energy systems.

“A shift from agriculture to biomass production means the community will need to pay for excess water. If there is very minimal rainfall during a year, the community will need to gain a water supply from the surrounding neighborhoods. Buying water, transporting it, and ensuring the corn is watered requires extra labor, which requires extra pay.”

Discussion

The application of ML for assessing interdisciplinary learning involves both the development of the process as well as using that process to understand student thinking and learning. The ML process here shows promise for use in evaluating complex constructed responses for systems thinking, especially as part of formative assessment practice, and we also report on the evaluation itself. Here we discuss findings in the context of our research questions and results, including limitations pertaining to each topic within each section.

Use of ML to uncover student understanding of FEW Nexus. Considerations for future assessments of student CRs, particularly in the context of science-related items, demand significant attention. Despite the relative success of current applications, there are remaining challenges to using ML approaches to score a broader range of assessment constructs and response types (Zhai et al., 2020a). These challenges can be characterized by limitations such as insufficient data (or specific types of responses/ideas in CR), subjectivity, imbalances, and the prevalence of noise, and these all present substantial obstacles within the iterative ML training process (Maestrales et al., 2021). These challenges, if not effectively addressed, have the potential to compromise the achievement of optimal model accuracy, thereby raising questions about the validity and reliability of ML applications in educational evaluation settings (Suresh and Guttag, 2021). Another challenge is the complexity of the assessment target (i.e., what you are trying to measure), and the complexity of expected student responses can pose challenges to such AI-based evaluation (Zhai et al., 2020a). Others have suggested that features of the assessment item itself, such as the subject domain or scenarios used in the assessment, might impact the accuracy of ML models (Lottridge et al., 2018; Zhai et al., 2021b). To address these challenges, we utilized automated scoring approaches for text classification, which examine complex systems integration. In this study, our technical strategies have introduced a practical solution

through data augmentation to help address insufficient data and data imbalance, yielding promising implications. This approach involves generating dummy responses that are subsequently revised with identified synonym sets, thus facilitating the measurement of responses with similar structures and content while preserving the overall meaning and essence. Notably, we found that this approach effectively improved model performance, particularly when dealing with specific descriptors in Reservoir and Biomass question items.

One complexity of assessing complex CRs in postsecondary education is that there are varying disciplinary requirements and usages of student literacy compared to the more consistent expectations of K-12 education. As such, student responses considered holistically consist of a range of literacy abilities, which can impact the “understanding” of natural language processing and text classification models. For this research, student responses were collected from across the United States at different institution types (baccalaureate colleges, master’s colleges, and doctoral universities) to provide a wider range of student responses from which to develop the ML models. The resulting models are thus trained on the many ways that people may write about the question item concepts. High variation in the responses, which can be the result of variation in literacy, language, and understanding, result in more complexity, and are thus more difficult items for model development. Some of this difficulty may be addressed with a larger sample size, but if student responses are too varied or certain types of responses are too infrequent in the sample, then accurate ML models may not be easily achievable. Further, although we refer to the scoring of responses into rubric bins, we posit another important outcome of this work is characterizing students’ thinking about FEW concepts. The inclusion of automated text scoring systems into formative assessment evaluation isn’t only for “scoring” but provides a way for instructors to use open-response items and identify complex student ideas, or potential barriers to student learning (Harris et al., 2023). This is a critical aspect of formative assessment practice, allowing instructors a richer, more nuanced view of how students’ think about complex systems like the FEW nexus.

Defining criteria for developing text classification models. During the course of our iterative process, models exhibited superior performance in certain rubric categories characterized by well-defined criteria and a robust explanatory framework outlining the expected content under each rubric category. This finding aligns with prior research that underscored the efficacy of ML algorithms in

successfully discerning the quality of student responses using fine-grained analytic scoring methodologies (Ariely et al., 2023). Conversely, challenges become apparent in scenarios where substantial overlap exists between rubric categories, leading to redundancy and a lack of clarity (Liu et al., 2014). In such instances, the Kappa value frequently falls short of the desired threshold (Zhai et al., 2021). These insightful observations underscore the imperative need for the refinement of rubric definitions within future assessments. This refinement should be guided by a comprehensive and quantitative delineation of assessment criteria, aimed at mitigating the issues of overlap and ambiguity that our study and prior research have duly highlighted. For example, we revised closely related yet exclusive rubric bins to a single, multi-class prediction after attempting multiple model improvement strategies, yet failing to meet threshold performance metrics. Changing the structure of the rubric maintained the coding criteria of individual bins, now as “levels”, but provided additional information about exclusivity which resulted in better overall model performance. Alternatively, other coding bins with overlapping criteria or developed with too fine-grained of categories than needed to differentiate student ideas, can be merged into a single code. Conversely, other rubric codes that are too broad initially may need to be split or have better-defined coding criteria to better categorize cases (Sripathi et al., 2023).

We also note different levels of successful performance metrics for text classification models for the Reservoir versus Biomass question items. Indeed, most models for the Reservoir question item rubric bins achieved very good performance (i.e., “almost perfect” Cohen’s kappa measures), but most models for the Biomass question item rubric bins achieved only “acceptable” performance. This is despite both assessment items being in the Environmental Science domain, being centered on the FEW Nexus as context, and undergoing similar iterations in ML development. We interpret these findings to provide further evidence that the underlying construct of the assessment items and/or the expected complexity in student response can influence ML model performance, as noted by others (Haudek and Zhai, 2023; Lottridge et al., 2018). Thus, a practical implication of this work is that more complex assessment targets (e.g., trade-offs in socio-ecological systems), or assessment items that encompass larger systems will need additional feature engineering or more advanced ML techniques for accurate response evaluation (Wiley et al., 2017; Zhai et al., 2020a). Further, this highlights the need for an iterative approach in these research efforts. Although we lay out our approach as a “cycle” (see Fig. 1), in practice, it is highly iterative, with results from all stages informing the work of other stages, often in feedback loops. To improve final model outcomes, all stages of item development, data collection, and rubric alignment should be revisited, not only tuning specific model features. Following principled item design procedures (e.g. Harris et al., 2019) and incorporating automated scoring systems into the methodological pipeline (Rupp, 2018) are important considerations. Nevertheless, successful item/rubric/model development often takes multiple iterative rounds, which we continue to do, and models should be updated and expanded.

Such challenges to using NLP for short answer scoring are well reported and exist for assessments across science domains (Shermis, 2015; Liu et al., 2016). This leads to a broad range of scoring model performances (see Zhai et al., 2021b). These iterative cycles of revision do require an investment of human effort with an outcome of having automated classification models that can predict categories for any number of new responses and for any number of new users. Further, researchers also learn about student thinking about the targeted key concepts (see section “Student understanding of systems thinking in the FEW nexus” below) as they work to design items, rubrics, and models (e.g. Sripathi et al., 2023).

Scoring novice to expert levels. Scoring through levels [Level 1–Level 4] allows us to see the real distribution of knowledge for students in introductory courses. Level 4 responses were least frequent, most likely due to this level’s creation being based on an instructor’s expert response. Although, level 4 responses were seldom seen in students, it allows us to set a growth goal and see students who have previous knowledge at the expert level. As seen previously, only one student was able to achieve that level, which suggests that the task at hand is indicative of student ability. The level 4 response level is a baseline for exemplary understanding, it can also be used in the future to see if senior level or graduate students are performing at the expected level or to evaluate different strategies for achieving higher learning outcomes. Another We did have one student be within the level, which suggests that it is possible that students can strive to that level at a beginning level course. In addition, this supports student learning and growth, as we can expect as learning improves the FEW system understanding and could be a good baseline for growth as more students learn better. Additionally, no students fell within the pre-established Level 2 for responses that only included C codes (responses that only addressed a change in water prices), however students who were rated in Level 3 and Level 4 did include that content in their response. This particular content seems to be closely connected with the higher level responses rather than being a piece of information distinct from the other content and future research may delve into the content mapping of student responses.

We report on using the computer predicted scores to place student response in expertise levels. Overall, the computer placement may slightly underpredict student performance on these items as compared to human assignment, especially in the mid-level. This is more notable in the Biomass item, especially for Part A, indicating that the difficulty of this item may affect level assignment. However, although the classifications result from individual models with varying degrees of accuracy, the overall distribution of responses across all levels approximates the distribution from human assigned placements (see Table 9 of the Supplemental Methods for human assigned placements as comparison with the ML predictions of Table 8). This supports the use of these automated classification models to evaluate group or large class performance as part of formative assessment practice, even though individual response placement in specific levels may vary. That is, a reasonable approximation of the distribution of a large number of responses collected in an introductory course can be generated in seconds to minutes using the developed classification models, as opposed to the effort of human reading and assigning levels to all collected responses.

Future prospects of generative AI. The recent advancements in generative AI have raised additional considerations about assessment in education, among a host of many different possible applications (Kasneci et al., 2023). Although many of these issues are common to uses of classroom assessment in many contexts, some issues are particularly overlapping with the process of assessment development and automated scoring presented here. Recent explorations in using large language models for automated scoring of essays and short responses show great promise (e.g. Cochran et al., 2023; Mizumoto and Eguchi, 2023; Latif and Zhai, 2024). Using such an approach would simplify and expedite the automated scoring process, thus permitting automated scoring for different assessment prompts (Weegar and Idestam-Almquist, 2024) and could contribute to generalizability of models (Mayfield and Black, 2020). One promising application of generative AI is to do pattern finding, contextualized representation of information, and clustering of collections of student responses to open-ended tasks in support of formative assessment practice (Wang et al., in review; Wulff et al., 2022). This may assist instructors to easily find patterns and capture

token-level representations in student responses based on the linguistic context, thus allowing them to attend to student ideas and thinking as exhibited in their classroom, without reading and sorting individual responses.

On the other hand, the use of generative AI in education raises many concerns about academic integrity and students easily finding or asking AI to generate answers to assessments (Chan, 2023). Some studies have found that generative AI models still perform less well for producing more complex assessment tasks and tend to do better on quantitative tasks as compared to explanatory (Nguyen Thanh et al., 2023). Additionally, regarding the language attributes, current AI-generated responses, when compared to human-authored counterparts, typically manifest a discernible deficiency in cohesive and coherent elements, accompanied by a writing style characterized by uniformity and repetition (Wang et al., 2023). It is very likely these shortcomings of these AI models will not last long. Instead, educators should re-evaluate the purposes of assessment (Chan, 2023), including how and what content and practices are necessary for students to be “skilled” in a discipline. Therefore, focusing teaching and learning on foundational principles within the discipline, which allows students to see science across contexts and define problem boundaries, like systems and systems models, maybe one such approach. Educators should also consider the purpose learning activities that students engage with, both in the classroom and outside of the classroom. The application of generative AI represents a frontier in the use of technology in support of formative assessment in the classroom (Harris et al., 2023).

Student understanding of systems thinking in the FEW Nexus.

Systems thinking involves understanding the interdisciplinary connections and relationships between associated components within a system, rather than simply focusing on discrete concepts (Meadows, 2008). For teaching and learning contexts, the FEW Nexus provides a scaffold for incorporating systems thinking and sustainability concepts into courses and across curricula. A primary advantage of the NGCI is the potential to capture a student’s understanding of relationships within the FEW Nexus. While the analytic rubrics were developed to score student understanding of FEW isolated discrete parts of the systems in the scenario presented by each item, by examining the constellation of scores a student response achieved across criteria, we quantified student patterns of explanations about these systems.

What our students know about the Food-Energy-Water Nexus. Both the Reservoir and the Biomass question items present students with scenarios about the FEW Nexus relationship centering water with connections to energy production and agriculture. The codes described in the analytic rubric represent the most common concepts students included when presented with these scenarios. The frequency of these concepts may indicate that these ideas are foundational as introductory students construct knowledge about FEW systems. Many of the most common codes could be classified as demonstrating basic knowledge, which is the simplest cognitive task presented in Bloom’s taxonomy model (Bloom and Krathwohl, 1956; Krathwohl, 2002). For example, in the Reservoir question item rubric, the A codes were the most commonly found in our dataset (Table 6), and indicated responses identifying that a dam could be related to hydropower. While this type of statement is reasonable for introductory level courses where students are developing new understanding and aligns with the content presented in introductory IES courses (Horne et al., 2023), knowledge statements alone do not achieve the competency goals for IES students (Wiek et al., 2011). More complex student responses in our study contained combinations of codes, however exceptionally creative explanations or

concepts were not always frequent enough to be included in the analytic rubric or be captured reliably in the ML models.

The Biomass item presents students with an opportunity to consider directionality within trade-offs, and directionality concepts are thus frequent in the associated rubric. In the Biomass sub-questions, students often included at least one statement about directionality of the quantity of food, energy, or water, but responses including predictions across these three ideas were infrequent. Making a statement about change or directionality, such as describing the quantity of food or water decreasing, is a relatively simple task in systems thinking, but is foundational to more complex tasks that consider changes over time (Sweeney and Sterman, 2007). Students who described trade-offs in their responses to this question sometimes went beyond discussing the cause-and-effect components of the system and discussed concepts not immediately asked by the question, such as the impact of this scenario on water pricing. However, these types of responses did not always register in the ML models, and some were too infrequent to be included in the rubric. The frequency of simpler codes describing FEW *concepts* in comparison to codes describing FEW *consequences* presents a challenge, given that IES curricula prioritizes FEW in relation to socio-environmental topics (Horne et al., 2023). The emergence of these concepts in student responses provides insight into what students will need to do with these ideas after the classroom and how students may move from identifying FEW concepts to applying predictions about FEW impacts on people, land, and communities.

How students assimilate systems thinking into their constructed responses. The frequency of co-occurrences in our analysis can serve as a proxy for gauging the level of understanding among respondents, ranging from novice to expert, regarding the relationships between food, energy, and water. The pattern of responses students gave sheds light on the connections between the concepts of food, energy, and water within the context of our study (see Table 6). In responses to the Reservoir item, we observed in our data that students frequently combined ideas under A codes (descriptions of hydropower) with C codes (uses of energy). This combination of concepts aligns well with the task presented in the item, and this pattern suggests a moderate association between generating hydropower and its applications in agriculture or infrastructure. Additionally, we noticed a prevalent co-occurrence of connections between C codes and B codes, signifying that students can connect the production of various energy types and their local utilization in agriculture or community resource management. These types of responses represent a robust understanding among students that hydropower is harnessed from water sources and plays a role in electricity generation. When students’ responses were coded into various categories such as the type of energy and its intended purposes (e.g., irrigation or powering homes), we found that responses indicating a more comprehensive understanding tended to include bin codes related to hydroelectricity, energy for irrigation in food production, energy for machinery, and energy for residential or agricultural purposes (see Table 7). In contrast, D codes represented facets of student explanations that centered on the use of *water*, and not necessarily energy, from the reservoir in the prompt. While overall these codes were less frequent than codes describing the use of energy, they were associated with more novice responses that co-occurred with A codes (e.g., stating that hydropower is related to the prompt) but not as frequently with explanations of how energy is produced and used for agriculture. Students commonly linked the concepts of energy generation, energy applications in agriculture, and broader infrastructure. In contrast, novice responses included the basic concept that dams were related to hydropower but lacked further knowledge about how energy is generated, how energy could be employed, or its relevance to food production, and instead offered

how water from the reservoir could be used for agricultural purposes (see Table 7).

While examining co-occurrences between codes within the Reservoir item explores how students characterize the components of a FEW system, examining co-occurring codes in responses to the Biomass item offers a way to model how students describe trade-offs. The combinations of co-occurring codes reflect the complexity of a students' response, which serves as the basis for the logic of the Novice to Expert scale (Table 8). Without including at least two of the facets of the FEW Nexus, a response to the Biomass item would not describe a trade-off. For example, a Level 1 response to the Biomass item would only include one facet of FEW, while a response including more specific details and more than one FEW element would be more expert-like. Further, moving from describing individual effects to multiple effects may also indicate a student is reasoning about the mechanism behind the system, which is a more expert-like approach to systems thinking (Hmelo-Silver and Pfeffer, 2004). However, aligned with previous research in science education indicating the challenge of developing expert-like systems thinking (Hmelo-Silver and Pfeffer, 2004; Jacobson and Wilensky, 2006; Sweeney and Sterman, 2007), expert-like Level 3 and Level 4 responses were infrequent in our dataset compared to responses providing simpler, incomplete explanations of the systems presented in the question.

Conclusion

There is growing support and interest in establishing interdisciplinary environmental education in higher education that integrate concepts and disciplines in addition to providing varied perspectives (Christie et al., 2015; Cooke and Vermaire, 2015; Wallace and Clark, 2018). Most of these IESs do not incorporate systematic evaluation and assessment, and especially non-summative evaluations, with one of the main challenges to developing evaluation being the diversity of content and fields (Vincent et al., 2017). There is a need to assess student learning in IESs as well as rigorous evaluation of IES educational practices, especially of complex synthesis concepts. Here, we described initial steps in developing ML text classification models as a tool to assess student systems thinking capabilities using two questions anchored by FEW Nexus phenomena (i.e., water-energy connections, biomass trade-offs). Our two questions are first steps to fulfilling a much-needed gap in educational assessment by providing a means to analyze complex concept integration related to the FEW Nexus using ML. Successes and challenges to ML approaches to scoring student FEW Nexus CR indicate several future research priorities for interdisciplinary, practice-based education research: further development of human scoring methods to specifically prepare training and test data for ML models; developing evaluation systems for student responses on novice to expert scales; developing assessment instruments using multiple CR question items; and examining how students incorporate social competencies and human factors into their explanations of FEW topics. Some of these research priorities address the critical issue of time investment in developing text classification models. Data collection, in the form of hundreds of student responses to the same question, rubric development (an iterative process), human scoring of student responses for training and test data, and model development (also an iterative process), all require a large amount of person-hours. This particular project has included the collaboration of 10 institutions for data collections, as well as two research labs at two additional institutions for scoring and model development with multiple postdoctoral scholars and graduate students. This investment is a severe limitation in the development of such models, and the process information presented here is intended to support other scholars in their model development through in-depth discussion of strategies for model

improvement and likely outcomes. However, once a model is developed and has achieved acceptable evaluation metrics, it can be used to very quickly assess large numbers of students' responses and conduct research on large datasets. This trade-off in investment is also offset by research that makes available resulting models to the scholarly community, as with questions and models presented in this paper (see supplemental information access).

Development of these question items using text classification models and CR assessment items allows evaluation of the relationship between foundational concept understanding and integration of those concepts as well as more nuanced understanding of student comprehension of complex interdisciplinary concepts. This proposed research represents one of the first attempts to assess the links between foundational, discipline-specific concepts and systems thinking and learning. We have been able to engage a range of institutions in all phases of the project thus far. Institutions were chosen as a representative sample of EPs across the US and include baccalaureate colleges (4), master's colleges (3), and doctoral universities (3). This is critical to ensure that findings and outcomes are applicable to undergraduates across the US. We anticipate that the information gleaned from reviewing environmental curricula across the United States, combined with concept inventory results showing student learning, will better inform those making curricular and staffing decisions regarding college environmental science and studies programs. Thus, students enrolled in IES programs will benefit by having courses and programs evaluated with a valid and reliable instrument. Additionally, combining discipline-specific ideas and phenomena within a new set of CR assessment items focused on complex system thinking will provide faculty with a valid and reliable instrument for evaluating learning. Our instrument development methodology is also applicable to other multidisciplinary assessments. For instance, the Next Generation Science Standards places a strong emphasis on using three-dimensional learning—how science practices, content knowledge, and crosscutting concepts interconnect (Douglas et al., 2020). Lastly, environmental and sustainability objectives are becoming commonplace among university mission and vision statements. Providing shared EP objectives with aligned assessments that can inform instruction and student learning helps meet these objectives of undergraduate education.

Data availability

Scored student response data is available through contact with the corresponding author. Source code for the text classification tools used in this study is available at <https://github.com/BeyondMultipleChoice/AACRAutoReport>. Assessment items are available at <https://beyondmultiplechoice.org/>. Text classification models will be saved and published to the public in subsequent papers at <https://beyondmultiplechoice.org/>.

Received: 22 September 2023; Accepted: 22 July 2024;
Published online: 13 August 2024

References

- Aggarwal CC, Zhai C (eds) (2012) Mining text data. Springer US
- Ariely M, Nazaretsky T, Alexandron G (2023) Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *Int J Artif Intell Educ* 33(1):1–34. <https://doi.org/10.1007/s40593-021-00283-x>
- Balint PJ, Stewart RE, Desai A, Walters LC (2011) Wicked environmental problems. Island Press/Center for Resource Economics
- Blatti JL, Garcia J, Cave D, Monge F, Cuccinello A, Portillo J, Juarez B, Chan E, Schwebel F (2019) Systems thinking in science education and outreach toward a sustainable future. *J Chem Educ* 96(12):2852–2862. <https://doi.org/10.1021/acs.jchemed.9b00318>

- Bloom B, Krathwohl D (1956) Taxonomy of educational objectives; the classification of educational goals by a committee of college and university examiners. Handbook I: Cognitive Domain. Longmans, Green, New York, NY
- Bonthu S, Rama Sree S, Krishna Prasad MHM (2023) Improving the performance of automatic short answer grading using transfer learning and augmentation. *Eng Appl Artif Intell* 123:106292. <https://doi.org/10.1016/j.engappai.2023.106292>
- Brandstädter K, Harms U, Großschädl J (2012) Assessing system thinking through different concept-mapping practices. *Int J Sci Educ* 34(14):2147–2170. <https://doi.org/10.1080/09500693.2012.716549>
- Brew C, Leacock C (2013) Automated short answer scoring: principles and prospects. In: Shermis MD, Burstein J (eds) *Handbook of automated essay evaluation*. Routledge
- Brundiers K, King J, Parnell R, Hiser K (2023) A GCSE proposal statement on key competencies in sustainability: guidance on the accreditation of sustainability and sustainability-related programs in higher education. Global Council for Science and the Environment, p. 40
- Bustamante M, Videira P, Baker L (2021) Systems thinking and complexity science-informed evaluation frameworks: assessment of the economics of ecosystems and biodiversity for agriculture and food. *N Dir Eval* 2021(170):81–100
- Carnegie Foundation for the Advancement of Teaching (2011) *The Carnegie Classification of Institutions of Higher Education*, 2010 edition. The Carnegie Classification of Institutions of Higher Education
- Chan CKY (2023) A comprehensive AI policy education framework for university teaching and learning. *Int J Educ Technol High Educ* 20(1):38. <https://doi.org/10.1186/s41239-023-00408-3>
- Chi M, VanLehn K, Litman D, Jordan P (2011) An evaluation of pedagogical tutorial tactics for a natural language tutoring system: a reinforcement learning approach. *Int J Artif Intell Educ* 21(1–2):83–113. <https://doi.org/10.3233/JAI-2011-014>
- Christie BA, Miller KK, Cooke R, White JG (2015) Environmental sustainability in higher education: what do academics think? *Environ Educ Res* 21(5):655–686. <https://doi.org/10.1080/13504622.2013.879697>
- Clark SG, Wallace RL (2015) Integration and interdisciplinarity: concepts, frameworks, and education. *Policy Sci* 48(2):233–255. <https://doi.org/10.1007/s11077-015-9210-4>
- Cochran K, Cohn C, Hastings P, Tomuro N, Hughes S (2023) Using BERT to identify causal structure in students' scientific explanations. *Int J Artif Intell Educ*. <https://doi.org/10.1007/s40593-023-00373-y>
- Cooke SJ, Vermaire JC (2015) Environmental studies and environmental science today: inevitable mission creep and integration in action-oriented transdisciplinary areas of inquiry, training and practice. *J Environ Stud Sci* 5(1):70–78. <https://doi.org/10.1007/s13412-014-0220-x>
- Crossley SA, Allen LK, Snow EL, McNamara DS (2016) Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *J Educ Data Min* 8(2):1–19
- D'Odorico P, Davis KF, Rosa L, Carr JA, Chiarelli D, Dell'Angelo J, Gephart J, MacDonald GK, Seekell DA, Suweis S, Rulli MC (2018) The global food–energy–water nexus. *Rev Geophys* 56(3):456–531. <https://doi.org/10.1029/2017RG000591>
- Deane P (2006) Strategies for evidence identification through linguistic assessment of textual responses. In: Williamson DM, Bejar II, Mislevy RJ (eds) *Automated scoring of complex tasks in computer-based testing*. Lawrence Erlbaum Associates, pp. 313–372
- Dogra V, Verma S, Kavita, Chatterjee P, Shafi J, Choi J, Ijaz MF (2022) A complete process of text classification system using state-of-the-art NLP models. *Comput Intell Neurosci* 2022:1883698. <https://doi.org/10.1155/2022/1883698>
- Douglas KA, Gane BD, Neumann K, Pellegrino JW. (2020) Contemporary methods of assessing integrated STEM competencies. In: Johnson CC, Mohr-Schroeder MJ, Moore TJ, English LD (eds) *Handbook of Research on STEM Education*. 1st ed. Routledge: 234–254
- Dugan KE, Mosyjowski EA, Daly SR, Lattuca LR (2022) Systems thinking assessments in engineering: a systematic literature review. *Syst Res Behav Sci* 39(4):840–866
- Gao X, Li P, Shen J, Sun H (2020) Reviewing assessment of student learning in interdisciplinary STEM education. *Int J STEM Educ* 7(1):24. <https://doi.org/10.1186/s40594-020-00225-4>
- Global Council for Science and the Environment (n.d.) Sustainability in Higher Education: Accreditation. <https://www.gcseglobal.org/pathways-to-accreditation>. Accessed 5 Sept 2023
- Gray S, Sterling EJ, Aminpour P, Goralnik L, Singer A, Wei C, Akabas S, Jordan RC, Giabbanelli PJ, Hodhod J, Betley E (2019) Assessing (social-ecological) systems thinking by evaluating cognitive maps. *Sustainability* 11(20):5753
- Grohs JR, Kirk GR, Soledad MM, Knight DB (2018) Assessing systems thinking: a tool to measure complex reasoning through ill-structured problems. *Think Skills Creat* 28:110–130
- Harris CJ, Krajcik JS, Pellegrino JW, DeBarger AH (2019) Designing Knowledge-In-Use assessments to promote deeper learning. *Educ Meas Issues Pra* 38(2):53–67. <https://doi.org/10.1111/emip.12253>
- Harris CJ, Weibe E, Grover S, Pellegrino JW (eds) (2023) *Classroom-based STEM assessment: Contemporary issues and perspectives*. Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc. <https://cadrek12.org/resources/classroom-based-stem-assessment-contemporary-issues-and-perspectives>
- Harrison JS, Josefy MA, Kalm M, Krause R (2023) Using supervised machine learning to scale human-coded data: a method and dataset in the board leadership context. *Strateg Manag J* 44(7):1780–1802. <https://doi.org/10.1002/smj.3480>
- Hartmann DP (1977) Considerations in the choice of interobserver reliability estimates. *J Appl Behav Anal* 10(1):103–116. <https://doi.org/10.1901/jaba.1977.10-103>
- Haudek KC, Zhai X (2023) Examining the effect of assessment construct characteristics on machine learning scoring of scientific argumentation. *Int J Artif Intell Educ*. <https://doi.org/10.1007/s40593-023-00385-8>
- Hestenes D, Wells M, Swackhamer G. (1992) Force concept inventory. *Phys Teach* 30(3):141–158. <https://doi.org/10.1119/1.2343497>
- Hmelo-Silver CE, Pfeffer MG (2004) Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cogn Sci* 28(1):127–138. https://doi.org/10.1207/s15516709cog2801_7
- Hmelo-Silver CE, Marathe S, Liu L (2007) Fish swim, rocks sit, and lungs breathe: expert-novice understanding of complex systems. *J Learn Sci* 16(3):307–331. <https://doi.org/10.1080/10580400701413401>
- Horne L, Manzanares A, Babin N, Royse EA, Arakawa L, Blavascunas E, Doner L, Druckenbrod D, Fairchild E, Jarchow M, Muchnick BR, Panday P, Perry D, Thomas R, Toomey A, Tucker BH, Washington-Ottombre C, Vincent S, Anderson SW, Romulo C (2023) Alignment among environmental programs in higher education: what Food–Energy–Water Nexus concepts are covered in introductory courses? *J Geosci Educ* 1–18. <https://doi.org/10.1080/10899995.2023.2187680>
- Horne L, Manzanares AD, Atalan-Helick N, Vincent S, Anderson SW, Romulo C (2024) An exploratory study of drawings as a tool to evaluate student understanding of the Food–Energy–Water (FEW) Nexus. *J Environ Stud Sci*. <https://doi.org/10.1007/s13412-024-00929-x>
- Hsieh H-F, Shannon SE (2005) Three approaches to qualitative content analysis. *Qual Health Res* 15(9):1277–1288. <https://doi.org/10.1177/1049732305276687>
- Jacobson MJ, Wilensky U (2006) Complex systems in education: scientific and educational importance and implications for the learning sciences. *J Learn Sci* 15(1):11–34. https://doi.org/10.1207/s15327809jls1501_4
- Jescovitch LN, Scott EE, Cerchiara JA, Merrill J, Urban-Lurain M, Doherty JH, Haudek KC (2021) Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *J Sci Educ Technol* 30(2):150–167. <https://doi.org/10.1007/s10956-020-09858-0>
- Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. *Science* 349(6245):255–260. <https://doi.org/10.1126/science.aaa8415>
- Jurka TP, Collingwood L, Boydston AE, Grossman E, Van Atteveldt W (2013) RTextTools: a supervised learning package for text classification. *R J* 5(1):6–12
- Kasneci E, Sessler K, Kuchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Gönemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, ... Kasneci G (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ* 103:102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Katz SL, Padowski JC, Goldsby M, Brady MP, Hampton SE (2020) Defining the nature of the nexus: specialization, connectedness, scarcity, and scale in Food–Energy–Water management. *Water* 12(4):972. <https://doi.org/10.3390/w12040972>
- Krathwohl DR (2002) A revision of bloom's taxonomy: an overview. *Theory Into Pract* 41(4):212–218. https://doi.org/10.1207/s15430421tip4104_2
- Latif E, Zhai X (2024) Fine-tuning ChatGPT for automatic scoring. *Comput Educ: Artif Intell* 6:100210. <https://doi.org/10.1016/j.caeai.2024.100210>
- Laverty JT, Underwood SM, Matz RL, Posey LA, Carmel JH, Caballero MD, Fata-Hartley CL, Ebert-May D, Jardeleza SE, Cooper MM (2016) Characterizing college science assessments: the three-dimensional learning assessment Protocol. *PLoS ONE* 11(9):e0162333. <https://doi.org/10.1371/journal.pone.0162333>
- Leal Filho W, Levesque VR, Salvia AL, Paço A, Fritzen B, Frankenberger F, Damke LI, Brandli LL, Ávila LV, Mifsud M, Will M, Pace P, Azeiteiro UM, Lovren VO (2021) University teaching staff and sustainable development: An assessment of competences. *Sustain Sci* 16(1):101–116. <https://doi.org/10.1007/s11625-020-00868-w>
- Leck H, Conway D, Bradshaw M, Rees J (2015) Tracing the Water–Energy–Food Nexus: description, theory and practice. *Geogr Compass* 9(8):445–460. <https://doi.org/10.1111/gec3.12222>
- Lee H-S, Gweon G-H, Lord T, Paessel N, Pallant A, Pryputniewicz S (2021) Machine learning-enabled automated feedback: supporting students' revision of scientific arguments based on data drawn from simulation. *J Sci Educ Technol* 30(2):168–192. <https://doi.org/10.1007/s10956-020-09889-7>

- Libarkin JC, Anderson SW (2005) Assessment of learning in entry-level geoscience courses: results from the geoscience concept inventory. *J Geosci Educ* 53(4):394–401
- Libarkin JC, Geraghty Ward EM (2011) The qualitative underpinnings of quantitative concept inventory questions. *Geological Society of America Special Papers*, vol 474. Geological Society of America, pp. 37–48
- Liu OL, Rios JA, Heilman M, Gerard L, Linn MC (2016) Validation of automated scoring of science assessments. *J Res Sci Teach* 53(2):215–233. <https://doi.org/10.1002/tea.21299>
- Liu OL, Brew C, Blackmore J, Gerard L, Madhok J, Linn MC (2014) Automated scoring of constructed-response science items: prospects and obstacles. *Educ Meas: Issues Pract* 33(2):19–28. <https://doi.org/10.1111/emip.12028>
- Liu SC (2023) Examining undergraduate students' systems thinking competency through a problem scenario in the context of climate change education. *Environ Educ Res* 29(12):1780–1795
- Lottridge S, Wood S, Shaw D (2018) The effectiveness of machine score-ability ratings in predicting automated scoring performance. *Appl Meas Educ* 31(3):215–232. <https://doi.org/10.1080/08957347.2018.1464452>
- Maestrales S, Zhai X, Touitou I, Baker Q, Schneider B, Krajcik J (2021) Using machine learning to score multi-dimensional assessments of chemistry and physics. *J Sci Educ Technol* 30(2):239–254. <https://doi.org/10.1007/s10956-020-09895-9>
- Mambrey S, Timm J, Landskron JJ, Schmiemann P (2020) The impact of system specifics on systems thinking. *J Res Sci Teach* 57(10):1632–1651. <https://doi.org/10.1002/tea.21649>
- Manzanares AD, Horne L, Royse EA, Azzarello CB, Jarchow M, Druckenbrod D, Babin N, Atalan-Helick N, Vincent S, Anderson SW, Romulo C (in review). Undergraduate students' knowledge about the relationships between climate change and the Food–Energy–Water Nexus. *Int J Sustain High Educ*
- Martins FP, Cezarino LO, Liboni LB, Botelho Junior AB, Hunter T (2022) Interdisciplinarity-based sustainability framework for management education. *Sustainability* 14(19):12289. <https://doi.org/10.3390/su141912289>
- Mayfield E, Black AW (2020) should you fine-tune BERT for automated essay scoring? In: Burstein J, Kochmar E, Leacock C, Madnani N, Pilán I, Yannakoudakis H, Zesch T (eds) proceedings of the fifteenth workshop on innovative use of NLP for building educational applications. Association for Computational Linguistics. 151–162. <https://doi.org/10.18653/v1/2020.bea-1.15>
- McNamara D, Graesser AC (2011) Coh-Metrix: an automated tool for theoretical and applied natural language processing. In *Applied Natural Language Processing*. IGI Global, pp. 188–205
- Meadows DH (2008) Thinking in systems: a primer. Chelsea Green Publishing
- Mitchell TM (1997) Machine learning, vol 1. McGraw-hill New York
- Mitkov R, Le An H, Karamanis N (2006) A computer-aided environment for generating multiple-choice test items. *Nat Lang Eng* 12(2):177
- Mizumoto A, Eguchi M (2023) Exploring the potential of using an AI language model for automated essay scoring. *Res Methods Appl Linguist* 2(2):100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Momsen J, Speth EB, Wyse S, Long T (2022) Using systems and systems thinking to unify biology education. *CBE—Life Sci Educ* 21(2):es3. <https://doi.org/10.1187/cbe.21-05-0118>
- National Research Council (2012) A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: The National Academies Press. <https://doi.org/10.17226/13165>
- National Science Foundation (2020) STEM Education for the Future—2020 Visioning Report.pdf. <https://www.nsf.gov/ehf/Materials/STEM%20Education%20for%20the%20Future%20-%202020%20Visioning%20Report.pdf>
- Nehm RH, Ha M, Mayfield E (2012) Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *J Sci Educ Technol* 21(1):183–196. <https://doi.org/10.1007/s10956-011-9300-9>
- NGSS Lead States (2013) Next generation science standards: for states, by states. The National Academies Press
- Nguyen Thanh B, Vo DTH, Nguyen Nhat M, Pham TTT, Thai Trung H, Ha Xuan S (2023) Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australas J Educ Technol* 39(5):59–81. <https://doi.org/10.14742/ajet.8902>
- Noyes K, McKay RL, Neumann M, Haudek KC, Cooper, MM (2020) Developing computer resources to automate analysis of students' explanations of London dispersion forces. *J Chem Educ* 97(11):3923–3936. <https://doi.org/10.1021/acs.jchemed.0c00445>
- Pellegrino JW, Wilson MR, Koenig JA, Beatty AS (2013) Developing assessments for the next generation science standards. National Academies Press
- Randle JM, Stroink ML (2018) The development and initial validation of the paradigm of systems thinking: development and validation of systems thinking. *Syst Res Behav Sci* 35(5):645–657
- Ravi M, Puente-Urbina A, van Bokhoven JA (2021) Identifying opportunities to promote systems thinking in catalysis education. *J Chem Educ* 98(5):1583–1593. <https://doi.org/10.1021/acs.jchemed.1c00005>
- Redman A, Wiek A, Barth M (2021) Current practice of assessing students' sustainability competencies: a review of tools. *Sustain Sci* 16(1):117–135. <https://doi.org/10.1007/s11625-020-00855-1>
- Redman A, Wiek A (2021) Competencies for advancing transformations towards sustainability. *Front Educ* 6. <https://www.frontiersin.org/articles/10.3389/educ.2021.785163>
- Romero C, Ventura S, Espejo PG, Hervás C (2008) Data mining algorithms to classify students. In: Baker RSJd, Barnes T, Beck JE (eds) Educational data mining 2008. The 1st International Conference on Educational Data Mining, Proceedings. Montréal, Québec, Canada
- Rupp AA (2018) Designing, evaluating, and deploying automated scoring systems with validity in mind: methodological design decisions. *Appl Meas Educ* 31(3):191–214. <https://doi.org/10.1080/08957347.2018.1464448>
- Shermis MD (2015) Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educ Assess* 20(1):46–65. <https://doi.org/10.1080/10627197.2015.997617>
- Shermis MD, Burstein J (2013) Handbook of automated essay evaluation: Current applications and new directions. Routledge
- Shiroda M, Uhl JD, Urban-Lurain M, Haudek KC (2022) Comparison of computer scoring model performance for short text responses across undergraduate institutional types. *J Sci Educ Technol* 31(1):117–128. <https://doi.org/10.1007/s10956-021-09935-y>
- Simpson GB, Jewitt GPW (2019) The development of the Water-Energy-Food Nexus as a framework for achieving resource security: a review. *Front Environ Sci* 7:8. <https://doi.org/10.3389/fenvs.2019.00008>
- Smajgl A, Ward J, Pluschke L (2016) The water–food–energy Nexus – Realising a new paradigm. *J Hydrol* 533:533–540. <https://doi.org/10.1016/j.jhydrol.2015.12.033>
- Soltis NA, McNeal KS (2022) Development and validation of a concept inventory for earth system thinking skills. *J STEM Educ Res* 5(1):28–52. <https://doi.org/10.1007/s41979-021-00065-z>
- Sripathi KN, Moscarella RA, Steele M, Yoho R, You H, Prevost LB, Urban-Lurain M, Merrill J, Haudek KC (2024) Machine learning mixed methods text analysis: an illustration from automated scoring models of student writing in biology education. *J Mixed Methods Res* 18(1):48–70. <https://doi.org/10.1177/15586898231153946>
- Stone A, Allen K, Rhoads TR, Murphy TJ, Shehab RL, Saha C (2003) The statistics concept inventory: a pilot study. 33rd annual frontiers in education, 2003. FIE 2003. T3D_1-T3D_6. <https://doi.org/10.1109/FIE.2003.1263336>
- Suresh H, Guttaj J (2021) A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. Association for Computing Machinery. <https://doi.org/10.1145/3465416.3483305>
- Sweeney LB, Sterman JD (2007) Thinking about systems: Student and teacher conceptions of natural and social systems. *Syst Dyn Rev* 23(2–3):285–311. <https://doi.org/10.1002/sdr.366>
- Tornabee R, Lavington E, Nehm RH (2016) Testing validity inferences for genetic drift concept inventory scores using Rasch and item order analyses [Conference paper]. National Association for Research in Science Teaching (NARST) Annual International Conference
- Underwood SM, Posey LA, Herrington DG, Carmel JH, Cooper MM (2018) Adapting assessment tasks to support three-dimensional learning. *J Chem Educ* 95(2):207–217. <https://doi.org/10.1021/acs.jchemed.7b00645>
- Urban-Lurain M, Merrill J, Haudek K, Nehm R, Moscarella R, Steele M, Park M (2015) Automated analysis of constructed responses: what are we modeling? [Conference paper]. National Meeting for the Society for the Advancement of Biology Education Research (SABER)
- Varela-Losada M, Vega-Marcote P, Pérez-Rodríguez U, Álvarez-Lires M (2016) Going to action? A literature review on educational proposals in formal environmental education. *Environ Educ Res* 22(3):390–421. <https://doi.org/10.1080/13504622.2015.1101751>
- Vincent S, Bunn S, Sloane S (2013) Interdisciplinary environmental and sustainability education on the nation's campuses: curriculum design. National Council for Science and the Environment, Washington, DC
- Vincent S, Rao S, Fu Q, Gu K, Huang X, Lindaman K, Mittleman E, Nguyen K, Rosenstein R, Suh Y (2017) Scope of interdisciplinary environmental, sustainability, and energy baccalaureate and graduate education in the United States. National Council for Science and the Environment: Washington DC
- Wallace RL, Clark SG (2018) Environmental studies and sciences in a time of chaos: problems, contexts, and recommendations. *J Environ Stud Sci* 8(1):110–113. <https://doi.org/10.1007/s13412-018-0469-6>
- Wang H, Troia GA (2023) Writing quality predictive modeling: integrating register-related factors. *Writ Commun* 40(4):1070–1112. <https://doi.org/10.1177/07410883231185287>
- Wang H, Li T, Haudek K, Royse EA, Manzanares M, Adams S, Horne L, Romulo C (2023) Is ChatGPT a threat to formative assessment in college-level science? an analysis of linguistic and content-level features to classify response types. In: Schlippe T, Cheng ECK, Wang T (eds) artificial intelligence in education technologies: new development and innovative practices. AIET 2023. Lecture Notes on Data Engineering and Communications Technologies, vol 190. Springer, Singapore. https://doi.org/10.1007/978-981-99-7947-9_13

- Wang H, Haudek KC, Manzanara AD, Romulo CL, Roysse EA (in review) extending a pretrained language model (BERT) using an ontological perspective to classify students' scientific expertise level from written responses
- Weegar R, Idestam-Almqvist P (2024) reducing workload in short answer grading using machine learning. *Int J Artif Intell in Educ* 34(2):247–273. <https://doi.org/10.1007/s40593-022-00322-1>
- Wiek A, Withycombe L, Redman CL (2011) Key competencies in sustainability: A reference framework for academic program development. *Sustain Sci* 6(2):203–218. <https://doi.org/10.1007/s11625-011-0132-6>
- Wiek A, Bernstein MJ, Rider WF, Cohen M, Forrest N, Kuzdas C, et al. (2016) operationalising competencies in higher education for sustainable Development. In Barth M, Michelsen G, Rieckmann M, Thomas I (eds) handbook of higher education for sustainable development. (London: Routledge), 297–317
- Wiley J, Hastings P, Blaum D, Jaeger AJ, Hughes S, Wallace P, Griffin TD, Britt MA (2017) Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *Int J Artif Intell Educ* 27(4):758–790. <https://doi.org/10.1007/s40593-017-0138-z>
- Williamson DM, Xi X, Breyer FJ (2012) A framework for evaluation and use of automated scoring. *Educ Meas: Issues Pract* 31(1):2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Wilson CD, Haudek KC, Osborne JF, Buck Bracey ZE, Cheuk T, Donovan BM, Stuhlsatz MAM, Santiago MM, Zhai X (2023) Using automated analysis to assess middle school students' competence with scientific argumentation. *J Res Sci Teach* 61:38–69. <https://doi.org/10.1002/tea.21864>
- Wulff P, Buschhüter D, Westphal A, Mientus L, Nowak A, Borowski A (2022) bridging the gap between qualitative and quantitative assessment in science education research with machine learning—a case for pretrained language models-based clustering. *J Sci Educ Technol* 31(4):490–513. <https://doi.org/10.1007/s10956-022-09969-w>
- Zehner F, Sälzer C, Goldhammer F (2015) Automatic coding of short text responses via clustering in educational assessment. *Educ Psychol Meas* 76(2):280–303. <https://doi.org/10.1177/0013164415590022>
- Zhai X, Krajcik J, Pellegrino JW (2021a) On the validity of machine learning-based next generation science assessments: a validity inferential network. *J Sci Educ Technol* 30(2):298–312. <https://doi.org/10.1007/s10956-020-09879-9>
- Zhai X, Shi L, Nehm RH (2021b) A meta-analysis of machine learning-based science assessments: factors impacting machine-human score agreements. *J Sci Educ Technol* 30(3):361–379. <https://doi.org/10.1007/s10956-020-09875-z>
- Zhai X, Haudek K, Shi L, H. Nehm R, Urban-Lurain M (2020a) From substitution to redefinition: a framework of machine learning-based science assessment. *J Res Sci Teach* 57(9):1430–1459. <https://doi.org/10.1002/tea.21658>
- Zhai X, Yin Y, Pellegrino JW, Haudek KC, Shi L (2020b) Applying machine learning in science assessment: a systematic review. *Stud Sci Educ* 56(1):111–151. <https://doi.org/10.1080/03057267.2020.1735757>

Acknowledgements

Our team would like to acknowledge and thank the students and faculty who contributed materials, time, and responses to this research. This material is based upon work supported by the National Science Foundation under Grant Nos. 2013373 and 2013359.

Author contributions

CLR, SWA, KCH, and SV conceived the original project; all authors contributed to conceptual design and provided background perspectives; EAR, ADM, LRH, CBA, SRA,

EF, and CLR, completed human coding for training and test data and HW, created the machine learning natural language processing algorithms; All authors contributed to the Introduction, Results and Discussion. All authors discussed results and interpretation, as well as reviewed and edited the manuscript at all stages.

Competing interests

The authors declare no competing interests.

Ethical approval

This work uses human subject matter data in the form of a survey that collected responses from over 400 IES undergraduates at seven post-secondary institutions across the United States. The protocol was reviewed by the lead institution (University of Northern Colorado) Institutional Review Board (IRB) and approved by all other participating institutions (UNCO IRB#158867-1). Responses were de-identified prior to human coding and for machine learning model development.

Informed consent

Prior to completing the IRB-approved survey, students consented to the research study by agreeing to the consent form provided in supplemental materials.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-024-03499-z>.

Correspondence and requests for materials should be addressed to Chelsie Romulo.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

¹Aims Community College, Greeley, USA. ²School of Psychological Sciences at the University of Northern Colorado, Greeley, CO, USA.

³Department of Counseling, Educational Psychology and Special Education at Michigan State University, East Lansing, MI, US. ⁴Department of Biochemistry and Molecular Biology at Michigan State University, East Lansing, MI, USA. ⁵Unity Environmental University, Unity, ME, USA.

⁶Department of Earth and Chemical Sciences at Rider University, Lawrenceville, NJ, USA. ⁷Department of Human Biology at Michigan State University, East Lansing, MI, USA. ⁸Metropolitan State University of Denver, Denver, CO, USA. ⁹Pacific Northwest National Laboratory, Richland, WA, USA. ¹⁰Vincent Evaluation Consulting LLC, Tulsa, OK, USA. ¹¹Department of Earth and Atmospheric Sciences, University of Northern Colorado, Greeley, CO, USA. ¹²Department of Geography, GIS and Sustainability at University of Northern Colorado, Greeley, CO, USA.

✉email: Chelsie.Romulo@unco.edu