AI-Assisted Learning and Concept Mapping for SDG-Focused Cross-School Outdoor Education in Taiwan

Cheng-Hung Weng, National Taipei University of Technology, Taiwan
Wen-Liang Chang, National Taipei University of Technology, Taiwan
Po-Ching Lu, National Taipei University of Technology, Taiwan

The Asian Conference on Education & International Development 2025
Official Conference Proceedings

Abstract

This study employs a quasi-experimental design to investigate the differences in learning outcomes and performance of junior high school students from metropolitan areas in Taiwan under various regional environments when engaging with water resource topics integrated with Sustainable Development Goals (SDGs). Students engaged in diverse outdoor environmental education activities, followed by classroom instruction on related issues. Their learning was assessed through text reading comprehension tests and the creation of concept maps. AI-assisted tools were used to evaluate students' learning progress and provide recommendations for deconstructing textual content. The preliminary research outcomes include: 1. Establishing practical models for cross-school collaboration to explore SDG-related topics; 2. Developing a process record for AI-assisted teaching evaluation and feedback systems; 3. Creating an assessment model utilizing AI to support concept map analysis; 4. Offering recommendations for the development of issue-oriented, cross-regional outdoor education curricula for sustainable development.

Keywords: concept mapping, AI-assisted teaching evaluation, SDGs

Introduction

Integrating the United Nations Sustainable Development Goals (SDGs) into curricula has become a global educational trend. This approach enhances students' awareness of global issues and fosters responsibility and engagement as global citizens. By exploring topics like climate change, inequality, and sustainable energy, students develop interdisciplinary thinking and problem-solving skills. SDG-focused education promotes active learning, critical thinking, and collaboration, making it a key strategy for cultivating sustainability-minded future generations.

Taiwan's subtropical climate and diverse ecosystems make it ideal for outdoor learning, particularly for water-related education. The Ministry of Education promotes outdoor programs where students explore rivers, conduct field experiments, and apply scientific methods such as water testing and biodiversity monitoring. Research shows that experiential learning improves both ecological understanding and emotional connection to environmental issues (Liefländer et al., 2013), while encouraging systems thinking about the links between environment, health, and social equity.

Due to geographic constraints, not all students can directly access natural environments. In such cases, texts and recorded media serve as alternatives. This presents a challenge for educators: how to evaluate students' understanding of unfamiliar contexts. Traditional tests may not capture the depth of students' cognitive engagement.

Concept maps help assess students' grasp of content structure and their ability to organize information. Visualizing concept relationships reveals how well students connect and integrated ideas. Network-style maps often indicate deeper understanding. As both learning and diagnostic tools, concept maps offer valuable insight into students' thinking.

However, evaluating concept maps is time-consuming. With the advancement of artificial intelligence (AI), automated tools now offer potential solutions. AI can interpret user input and images, raising the question: can it reliably assess student-generated concept maps and provide individualized feedback?

Recent studies show strong alignment between AI and human scoring of concept maps. For example, Bleckmann and Friege (2023) found that AI scored student maps with 80% accuracy and a Cohen's κ of 0.73—comparable to human raters. Medical education research supports AI validity as well. Ho et al. (2018) found strong correlations between AI-scored maps and manually graded essay questions, highlighting the efficiency and consistency of automated assessment. With proper training and frameworks, AI can score concept maps with a reliability close to human judgment (Bleckmann & Friege, 2023; Ho et al., 2018). Yet, challenges remain. Hubal et al. (2020) noted that while AI captured conceptual complexity, it struggled with evaluating organizational structure, suggesting a need for further refinement. As AI evolves, scoring accuracy continues to improve, narrowing the gap with human evaluation. This progress opens promising possibilities at the intersection of educational assessment and AI.

Building upon the preceding discussion, this study seeks to investigate the following research questions:

1. What are the processes involved in establishing cross-school collaboration to explore water resource issues in the context of the Sustainable Development Goals (SDGs)?

2. In what ways can artificial intelligence (AI) support teachers in assessing and analyzing student-generated concept maps, and what are the critical stages of this process?
3. How do assessment outcomes differ between AI-based evaluations and those conducted by human teachers?

Research Design

Outdoor Education

A substantial body of research has demonstrated that when students engage in outdoor learning experiences facilitated by teachers, they tend to exhibit enhanced development across cognitive, physical, social, and emotional domains. Utilizing natural environments as an extension of the indoor classroom has been shown to improve both academic performance and lifelong learning behaviors (Ruether, 2018). Compared to traditional, monotonous classroom instruction, outdoor environments foster greater student engagement and active participation (Dettweiler et al., 2015). Furthermore, students involved in outdoor educational activities not only achieve more enduring learning outcomes but also benefit from interdisciplinary learning experiences gained through direct interaction with the natural world (Becker et al., 2017).

According to a comprehensive study by Ruether (2018), outdoor education typically encompasses the following six characteristics:

1. It enables students to form meaningful connections between acquired knowledge and real-life contexts;
2. It supports learning through multisensory experiences;
3. It enhances students' intrinsic motivation to learn;
4. It provides more opportunities for peer interaction through physical movement;
5. It facilitates the development of students' social skills;
6. It contributes to behavioral improvements, particularly among students with attention-related difficulties, such as those with ADHD.

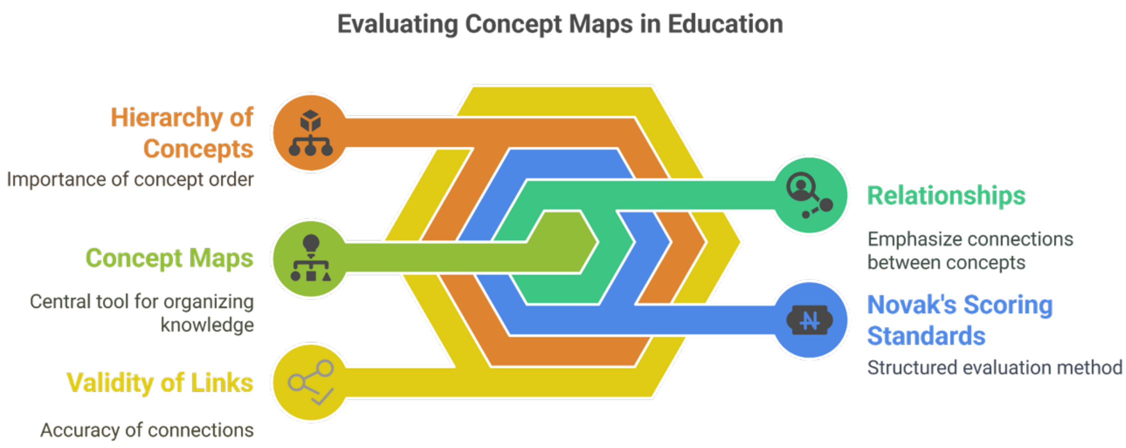
For students residing in urban environments, opportunities to engage with untouched natural settings are often limited. Therefore, educators can consider utilizing semi-natural environments, such as riversides, as alternative outdoor learning spaces. Such settings are likely to elicit distinct and potentially enhanced learning outcomes from students participating in these courses.

Concept Maps in Education (Concept Maps and Novak's Scoring Standards)

Concept maps are graphical tools used to organize and represent knowledge, emphasizing the relationships between different concepts. They are particularly effective in science education, where complex ideas and their interconnections are central to learning. Novak's scoring standards provide a structured approach to evaluating concept maps, focusing on the hierarchy of concepts, the validity of links, and the overall structure of the map. These standards have been widely adopted in educational research and practice, offering a reliable method for assessing the depth and accuracy of conceptual understanding (Adlaon, 2012; Lubberts, 2009; Novak & Gowin, 1984).

The scoring criteria for concept maps are mainly based on the N-G scoring method proposed by scholars Novak and Gowin (1984), which is based on Ausubel's learning theory and divides concept maps (Figure 1) into four items for scoring (Ausubel, 1968). These four items and their scoring criteria are listed below:

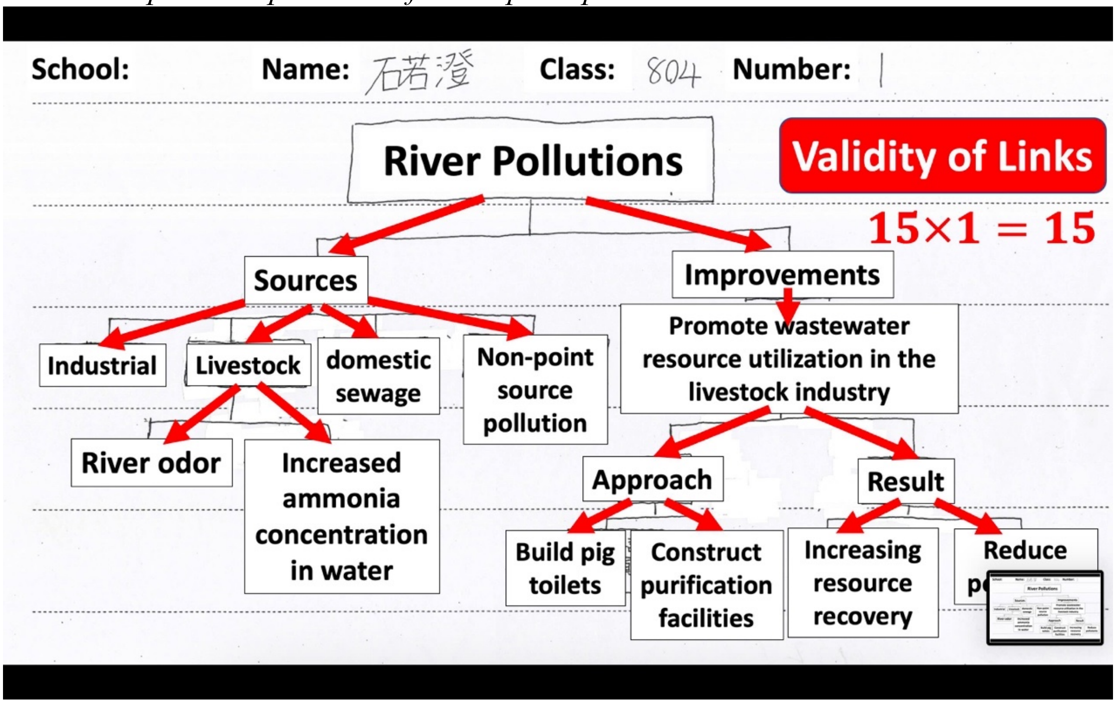
Figure 1
Concept Maps and Novak's Scoring Standards



Relationships or Propositions

One mark is awarded for each valid proposition (i.e. a meaningful link between two concepts). No marks will be awarded or deducted for vague or incorrect links, as shown in Figure 2.

Figure 2
Relationships or Propositions of Concept Map

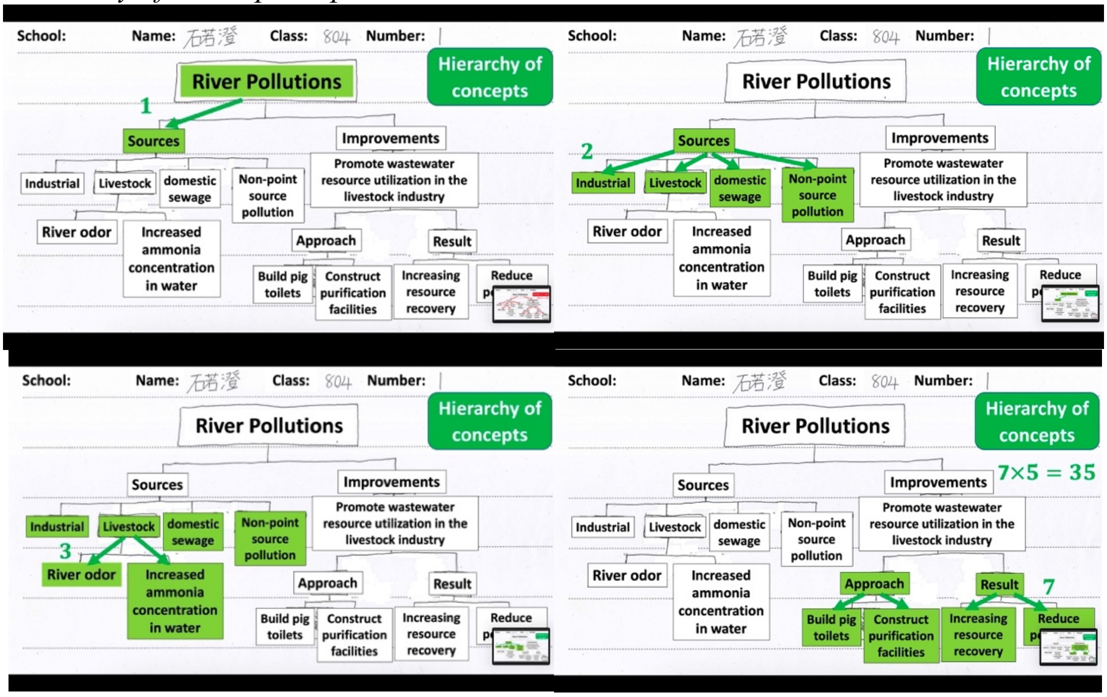


Hierarchy

The concept map is presented in a hierarchical format, with 5 points given for each valid hierarchical relationship. The hierarchical relationships indicate the organization of the concept, as shown in Figure 3.

Figure 3

Hierarchy of Concept Map



Cross-Links

Ten marks were awarded for effective cross-linking, indicating that students were able to make meaningful connections between different areas of knowledge and demonstrate creative thinking. If the cross-linking is effective but the conceptual propositions cannot be combined, 2 marks will be awarded.

Example

The student explains the meaning of the concept through a specific event or object, awarding 1 mark for an effective example.

In addition, there are other scoring methods such as Goldsmith and Johnson's Closeness index scoring method, which is assessed by comparing the structural similarity of the students' concept maps to the experts' concept maps.

Concept Map Scoring Formula

Total Score = (5 × number of valid strata) + (1 × number of valid propositions) + (10 × number of valid cross-links) + (1 × number of valid examples).

Benefits of Concept Maps

Visual organization of knowledge: Concept maps help students understand the relationships between concepts by presenting complex knowledge in a visual way.

Promoting Creative Thinking: Through cross-linking, students are able to make connections between different areas of knowledge and develop creative thinking skills.

Assessing Learning Outcomes: Concept maps can be used as a tool to assess student learning outcomes, reflecting students' knowledge organization and understanding.

AI in Educational Assessment

The advent of AI has revolutionized various aspects of education, including assessment. AI-powered tools are increasingly being used to automate grading, provide personalized feedback, and enhance the efficiency of educational evaluations. In the context of concept mapping, AI can potentially analyze the structure and content of maps, offering objective and consistent scoring based on predefined criteria. However, the consistency of AI-generated scores with those of human teachers remains a critical area of investigation (Ivanova et al., 2024; Ogunsakin, 2024; Owan et al., 2023).

The Prompt Framework in AI Research

The Prompt Framework refers to a structured approach used in AI systems to generate responses to specific tasks. In the context of educational assessment, this framework can be adapted to guide AI systems in evaluating concept maps. By incorporating Novak's scoring standards into the Prompt Framework, AI systems can be trained to assess concept maps with a high degree of accuracy and consistency. This integration has the potential to address some of the challenges associated with manual scoring, such as subjectivity and time constraints (Anohina & Grundspenkis, 2007; Ivanova et al., 2024).

Score Consistency Between AI Evaluation and Human Teachers

The consistency of scores between AI evaluation systems and human teachers is a crucial factor in the adoption of AI in educational assessment. Studies have shown that AI systems can achieve high levels of accuracy in scoring concept maps, particularly when trained on large datasets of human-scored maps. However, the consistency of AI-generated scores with those of human teachers can vary depending on the complexity of the maps and the specific criteria used for evaluation. For instance, AI systems may struggle with assessing the hierarchical structure of concepts, a key aspect of Novak's scoring standards, leading to discrepancies in scores (Anohina-Naumeca et al., 2010; Pramjeeth & Ramgovind, 2024; Zhao, 2024).

Research Design

Participants

In this study, a purposive sampling method was used to select two classes in each of the two junior high schools in Taipei City and New Taipei City, two administrative districts in Taiwan, for the teaching experiment (n = 110). Four classes with a total of 110 students were selected using the intentional sampling method with eighth-grade students. Students were given an instructional design that included a concept map and a text on water pollution. There was no significant difference in the pre-test scores of the four classes in their respective schools compared to their science scores in the previous semester.

Research Tools: AI Platform Usage

Claude

After testing on various AI platforms, Claude can currently read the image files, but not the images in the PDF.

NotebookLM

It is currently the only AI interface that can read PDFs of students' hand-drawn conceptual maps.

AI-Scored Content Checking

A scoring consistency test is conducted, and the prompts are then analyzed and optimized by natural science subject matter experts to determine the content of the scores that meets the needs of the subject matter.

We found that the only mainstream AIs that can read handwritten and hand-drawn messages in pictures and perform scoring accurately during the study period are Claude and notebooklm, but Claude can only read picture files.

However, the large number of images in the scanned PDF files can be easily read and accurately scored by notebooklm. So we chose notebooklm to build a fast and convenient AI personalized scoring system.

Verify AI Scoring Consistency

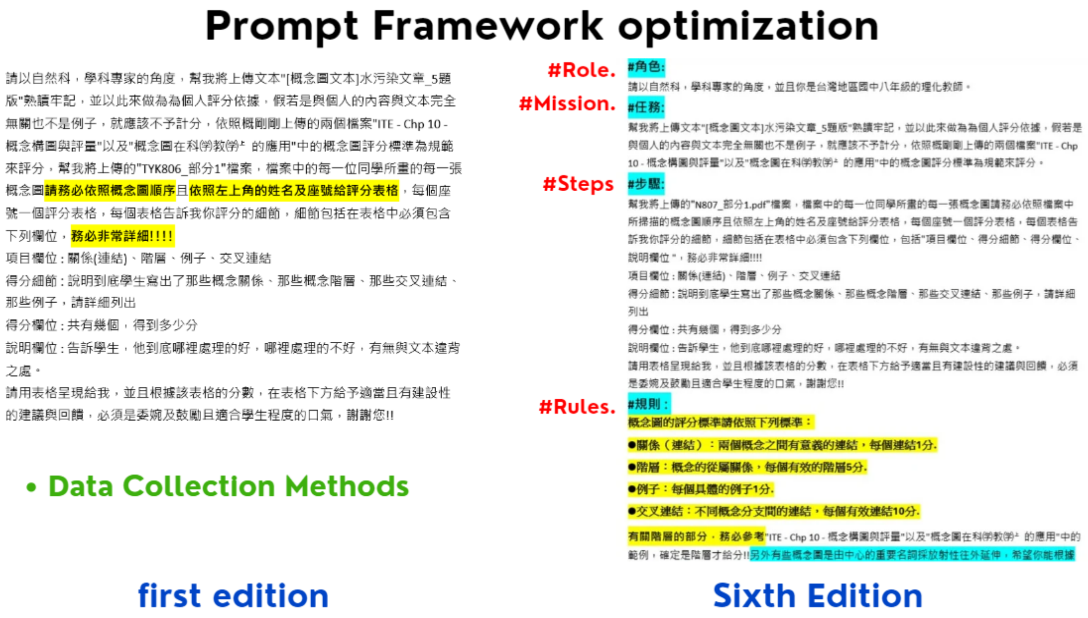
We were concerned about the impact of version updates and the time gap between ratings on the consistency of ratings, so we did a ratings consistency test, and the result was that after a week both the total score and the sub-items were perfectly consistent.

Prompt Framework Optimization

In the part of prompt Framework optimization, we went through six iterative optimization processes and finally reached a relatively stable state that is closer to the scoring of human teachers. It is mainly divided into four parts: roles, tasks, steps, and rules. Among them, the rules part is the focus of optimization and can correct more problems, as shown in Figure 4.

Figure 4

Prompt Framework Optimization



Differentiated Feedback

The section on differentiated feedback is outlined below. The red box on the left side of the picture is about how to ask AI to give feedback in a tactful way and it must be feedback that is appropriate for the students. The red square on the right is the personalized feedback message given by AI to give students positive reinforcement.

Data Collection Methods

In this study, students were provided with a text article discussing the issue of water pollution in Taiwan. The content covered key aspects such as the sources of pollution, relevant environmental regulations, the proportion of various pollutants, and the underlying causes associated with each source. Furthermore, the article introduced the River Pollution Index (RPI) as a framework for evaluating pollution levels, using the Bei-gang River as a case study. Following the reading, students were required to construct concept maps that visually represented the main ideas and relationships presented in the text.

Results

Grading Consistency Between Human Teachers and AI

From the results, three of the four classes in this study reached a "high degree of consistency", while the class of TYK 804 had a large difference in the scores between the AI and human teachers; therefore, it is an important reference for the optimization of cues in the following section.

This indicates a fair level of consistency, suggesting that the two grading mechanisms have limited consistency in classifying students' scores. (Table 1)

Table 1
Grading Consistency Between Human Teachers and AI

Group	Kappa	Interpretation	ICC	Interpretation	Overall Evaluation
N803	0.687	Substantial Agreement	0.835	High Agreement	Highly reliable
N807	0.78	Substantial Agreement	0.83	High Agreement	Highly reliable
TYK806	0.766	Substantial Agreement	0.923	Very High Agreement	Extremely reliable
TYK804	0.236	Fair Agreement	0.331	Moderate Agreement	Needs improvement

Teachers reviewing the concept maps drawn by students will give appropriate differentiated feedback on each section (relationships, hierarchy, examples, cross-links), but the time cost of reviewing is enormous.

Comparative Analysis of AI and Human Teachers

Relationship Score

Human teachers averaged 10.44 and AI averaged 10.89, a relatively small difference.

The correlation coefficient of 0.90 indicates a highly consistent trend in the relationship score ratings between AI and human teachers.

t examined ($t = -1.19$, $p = 0.25$), indicating that there was no statistically significant difference between the relationship scores of the human teachers and the AI.

Class Score

Human teachers had the same mean (18.33) and the same standard deviation as AI.

The correlation coefficient of 0.96 indicates that the trend of AI and human teachers' rank score ratings is very consistent.

The t-test ($t = 0.00$, $p = 1.00$) shows that they are identical with no significant difference.

Example Score

The mean value of both human teachers and AI is 0.22, and the correlation coefficient is 1.00, which means that the AI and human teachers' ratings are identical.

The t-test could not be calculated, probably because of the distribution of the data, but there is no significant difference between the two in terms of mean and correlation.

Crosslink Fraction

Human teachers averaged 5.00, AI averaged 3.89, and AI was slightly lower.

The correlation coefficient of 0.89 indicates that the trend in cross-linking scores between AI and human teachers remains fairly consistent.

t test (t = 1.46, p = 0.16), indicating no statistically significant difference between the two. (Table 2)

Table 2
Comparison of Human and AI Scores Across Different Score Types

Score Type	Human Mean	AI Mean	Correlation Coefficient	t-test (t, p)	Summary
Relationship Score	10.44	10.89	0.90	t = -1.19 p = 0.25	No significant difference
Class Score	18.33	18.33	0.96	t = 0.00 p = 1.00	Identical ratings with consistent variability
Example Score	0.22	0.22	1.00	Not calculated	Identical ratings
Crosslink Fraction	5.00	3.89	0.89	t = 1.46 p = 0.16	No significant difference

Statistics of Students' Concept Map Performance

Overall, students from TYK outperformed Nangang students in terms of total conceptual map scores, especially in the relationship and example categories. This suggests that TYK students tended to provide more examples and achieved higher overall coherence in their maps. However, Nangang students displayed more cross-linking among concepts, indicating stronger connections between different parts of their conceptual maps. (Table 3)

Table 3
Statistics of Students' Concept Map Performance

Score Category	Nangang (Mean ± SD)	Taoziqiao (Mean ± SD)	Comparison/Note
Relationship Score	9.68 ± 3.84	14.47 ± 5.82	Taoziqiao students scored significantly higher.
Hierarchy Score	16.58 ± 8.39	18.91 ± 5.67	Taoziqiao students scored slightly higher.
Example Score	0.66 ± 0.78	1.13 ± 2.04	Taoziqiao students provided more examples; higher SD indicates larger individual differences.
Cross-Link Score	3.95 ± 5.95	0.36 ± 2.70	Nangang students showed more frequent cross-linking among concepts.
Total Score	28.58 ± 14.31	33.91 ± 10.12	Taoziqiao students had a higher overall score for conceptual maps.

Discussion

Potential Influencing Factors

1. Instructional Strategies**:** Variations in teaching methods between Nan-Gang and TYK may influence how students construct concept maps. For instance, the pedagogical approach in Nan-Gang may emphasize making inter-conceptual connections, while TYK may prioritize the development of hierarchical structures.
2. Student Learning Backgrounds**:** Differences in students' prior knowledge of subject content may also impact concept map performance. TYK students may demonstrate stronger understanding of hierarchical relationships, resulting in higher scores in related categories. Conversely, Nan-Gang students may be more adept at identifying cross-conceptual links, thereby excelling in cross-linking assessments.
3. Scoring Rubrics and Complexity**:** Discrepancies in scoring standards or students' varied interpretations of how to construct concept maps may contribute to significant differences in performance. Further refinement of rubric clarity is necessary to address such inconsistencies.

Conclusion

1. Frequent updates by AI software providers present a key challenge for maintaining consistent scoring when using AI for assessment and feedback.
2. Even in the absence of such updates, temporal variations in AI scoring suggest that model stability must be ensured—an issue not exclusive to AI, as human raters also exhibit variation over time.
3. Evaluation results revealed a limited understanding of the concept of "hierarchy" among both AI systems and students. To address this, students should receive clearer definitions, illustrative examples, and additional practice before generating concept maps. The hierarchical dimension is particularly suitable for analyzing structured scientific texts.
4. AI systems sometimes misclassify concept relationships as examples, leading to slight scoring inconsistencies.
5. The current assessment approach measures quantity but lacks the ability to weight scores based on concept importance. For instance, in the Beigang River article, livestock wastewater is heavily emphasized. Students accurately representing this key issue should be awarded higher scores, but current scoring methods cannot account for such qualitative distinctions.
6. Complex conceptual maps can present interpretive challenges for teachers, often requiring subjective inference. In such cases, AI-assisted scoring can help identify the underlying logic and support scoring consistency.
7. While AI effectively identifies handwritten concepts and even recognizes spelling errors, its performance decreases significantly when faced with illegible handwriting. Misinterpretation of handwritten terms remains a challenge—particularly in languages like Chinese—highlighting the current limitations of AI compared to human evaluators.

Suggestions and Future Improvements

1. Enhancing the AI Scoring System

   AI scoring should incorporate weighted adjustments based on human grading patterns.
   For example:

- Relationships: 40%
- Hierarchy: 30%
- Examples: 20%
- Cross-links: 10%

A hybrid scoring model could be developed, allowing AI to learn from historical teacher grading data and better predict scores rather than relying solely on additive calculations.

2. Aligning AI with Human Scoring Standards

Regarding standardized scoring criteria, the limited influence of example scores and cross-links may stem from ambiguities in current guidelines. These should be refined to enhance scoring reliability. Additionally, prompt optimization is essential to align AI-generated feedback more closely with human ratings. The inclusion of "cross-linking" in scoring should also be revisited, as current AI systems often fail to recognize this element—perhaps due to its underrepresentation in human instruction.

Final Reflections

- Human evaluators demonstrate greater flexibility, particularly in assessing relationships and hierarchical structures.
- In contrast, current AI models apply equal weighting across all subcategories, resulting in overly mechanical assessments.
- Optimizing AI scoring to reflect human grading logic will enhance consistency and trust in automated systems.
- Human raters can likewise benefit from more standardized rubrics to ensure fairness and coherence in evaluations.

These findings provide valuable insights for improving AI-based scoring models and enhancing teacher assessment practices. Moreover, engaging students in outdoor exploration and direct interaction with natural environments fosters greater sensitivity toward sustainability issues. When students care deeply about the environment, teachers can more effectively guide them through issue-based instruction. Combined with AI-assisted feedback, this approach supports deeper learning and a more comprehensive educational experience.

References

Adlaon, R. (2012). *Assessing effectiveness of concept map as instructional tool in high school biology*. Louisiana State University Libraries. https://doi.org/10.31390/gradschool_theses.2425

Anohina, A., & Grundspenkis, J. (2007). A Concept Map Based Intelligent System for Adaptive Knowledge Assessment. *Frontiers in Artificial Intelligence and Applications.* 263–276. https://ebooks.iospress.nl/publication/3461

Anohina-Naumeca, A., Strautmane, M., & Grundspenkis, J. (2010). Development of the scoring mechanism for the concept map based intelligent knowledge assessment system. *Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies*. 376–381. https://doi.org/10.1145/1839379.1839446

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart & Winston.

Becker, C., Lauterbach, G., Spengler, S., Dettweiler, U., & Mess, F. (2017). Effects of Regular Classes in Outdoor Education Settings: A systematic review on students' learning, social and health dimensions. *International Journal of Environmental Research and Public Health*, *14*(5), 485. https://doi.org/10.3390/ijerph14050485

Bleckmann, T., & Friege, G. (2023). Concept maps for formative assessment: Creation and implementation of an automatic and intelligent evaluation method. *Knowledge Management & E-Learning*, *15*(3), 433–447. https://doi.org/10.34105/j.kmel.2023.15.025

Dettweiler, U., Ünlü, A., Lauterbach, G., Becker, C., & Gschrey, B. (2015). Investigating the motivational behavior of pupils during outdoor science teaching within self-determination theory. *Frontiers in Psychology, 6*, 125. https://doi.org/10.3389/fpsyg.2015.00125

Ho, V. W., Harris, P. G., Kumar, R. K., & Velan, G. M. (2018). Knowledge maps: A tool for online assessment with automated feedback. *Medical Education Online*, *23*(1), 1457394. https://doi.org/10.1080/10872981.2018.1457394

Hubal, R., Bobbitt, L., Garfinkle, S., Harris, S. C., Powell, B. D., Oxley, M. S., Anksorus, H. N., & Chen, K. Y. (2020). Testing of a Program to Automatically Analyze Students' Concept Maps. *Pharmacy* (Basel, Switzerland), *8*(4), 209. https://doi.org/10.3390/pharmacy8040209

Ivanova, M., Ivanova, T., & Terzieva, V. (2024). Automating Assessment within Intelligent Education. *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, Varna, Bulgaria, 2024, 1-6. https://doi.org/10.1109/IS61756.2024.10705174

Liefländer, A. K., Froehlich, G., Bogner, F. X., & Schultz, P. W. (2013). Promoting connectedness with nature through environmental education. *Environmental Education Research*, *19*(3), 370–384. https://doi.org/10.1080/13504622.2012.697545

Lubberts, P. T. (2009). Concept Maps in the Science Classroom. *Mathematical and Computing Sciences Masters*. Paper 1. https://fisherpub.sjfc.edu/cgi/viewcontent.cgi?article=1000&context=mathcs_etd_masters

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press.

Ogunsakin, I. (2024). Unlocking The Potential of Artificial Intelligence: A New Paradigm for Assessment in 21st Century Education. *International Journal of Research in STEM Education*, *6*(2), 37–49. https://doi.org/10.33830/ijrse.v6i2.1698

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Pramjeeth, S., & Ramgovind, P. (2024). Reconceptualizing assessment design in the age of AI: Is it genuine or faux? *New Directions for Teaching and Learning*, 1–12. https://doi.org/10.1002/tl.20625

Ruether, S. (2018). Barriers to Teachers' Use of Environmentally-Based Education in Outdoor Classrooms. *Walden Dissertations and Doctoral Studies*, 5776. https://scholarworks.waldenu.edu/dissertations/5776

Zhao, C. (2024). AI-assisted assessment in higher education: A systematic review. *Journal of Educational Technology and Innovation*, *6*(4). https://doi.org/10.61414/jeti.v6i4.209

Contact email: jacky745@tykjh.ntpc.edu.tw