Research on Generative AI-Enabled Development of Contextualized Science Test Questions for Junior High Schools Based on SOLO Taxonomy

Yujie Tang

Jinghengyi School of Education, Hangzhou Normal University, Hangzhou, Zhejiang,311121

Abstract: This study investigates the potential and limitations of applying generative artificial intelligence (AIGC) to the development of contextualized science test questions for junior high schools based on the SOLO taxonomy. By comparing one-step and step-by-step prompting strategies, SOLO-based test questions on the topic of "properties of carbon dioxide" were generated. The analysis focuses on AIGC's performance in context design, question chain construction, and progression of thinking levels. The findings indicate that step-by-step instructions more effectively guide AIGC to generate novel, structured, and logically coherent contexts and question chains. However, issues such as lack of scientific accuracy, overly explicit prompts, and weak logical connections between sub-questions remain. Therefore, AIGC can serve as an efficient auxiliary tool for test question development, providing creative ideas and initial drafts, but teachers are still required to conduct scientific validation, logical restructuring, and language optimization to achieve human-AIGC collaboration. Future efforts to further improve AIGC-generated question quality may include providing positive and negative examples, using step-by-step iterative instructions, and embedding domain-specific knowledge constraints.

Keywords: generative artificial intelligence, SOLO taxonomy, contextualized test questions

DOI:10.12417/3029-2328.26.04.022

1.Research Background

With the rapid development of generative artificial intelligence (AIGC) technology, its potential for application in the field of education has become increasingly prominent. AIGC can generate high-quality text, images, and even multimedia content through natural language processing and deep learning techniques, providing new tools and methods for education and teaching.[1]In the field of educational measurement and evaluation, AIGC has been preliminarily applied in areas such as automated scoring, test item generation, and personalized learning path design, demonstrating advantages of high efficiency, convenience, and scalability. Particularly in generating objective test items, AIGC has already been able to produce basic, memory-based questions to a satisfactory extent, thereby reducing teachers' burden in test development to a certain degree.

However, most large language models still exhibit limitations when required to generate complex test items that assess higher-order thinking, contextual integration, and the synthesis of disciplinary competencies. These limitations often include overly explicit question phrasing, low information complexity, and low cognitive demand, making it difficult to effectively assess students' core competency levels.[2]Under the "One Core, Four Layers, Four Wings" evaluation framework of the new Gaokao evaluation system, the suitability of AIGC-generated test items remains to be improved. Therefore, this study aims to address the gap in AIGC's capacity for developing subjective test items by constructing a question development framework that integrates the SOLO taxonomy with the Gaokao evaluation system, thereby providing a structured and standardized pathway for question development. On one hand, by optimizing prompt engineering and human-AIGC collaboration mechanisms, the quality and applicability of AIGC-generated test items can be further enhanced, thereby reducing teachers' question development workload and costs, and promoting the intelligent transformation of educational evaluation. On the other hand, given the current issue of test items becoming progressively easier and losing differentiation power due to test-taking strategies, AIGC can effectively enhance the novelty and challenge of test items by introducing unfamiliar contexts, interdisciplinary materials, and innovative question designs, thus better assessing students' authentic disciplinary competencies and thinking levels.

2.Literature Review

The promulgation of the compulsory education curriculum plans and subject-specific curriculum standards by the Ministry of Education marks the official transition of China's compulsory education toward a new era centered on core competencies. The transformation of educational concepts has consequently driven changes in teaching practices, including textbook compilation, classroom instruction, and assessment and evaluation.Competency orientation, integrated teaching-learning-assessment, and technology empowerment are the directions and requirements for the reform of China's basic education evaluation in the new era, resonating with the trends in international educational evaluation reform. Against this backdrop, traditional evaluation methods focusing on fragmented knowledge recall and simple skill application can no longer meet the requirements of competency-oriented assessment. To achieve scientific evaluation of students' core competency development levels, a paradigm shift in evaluation is imperative. The key to this shift lies in two pillars: first, in the presentation format of test items, authentic, rich, and well-structured contextual carriers must be used to concretize abstract competency requirements into observable cognitive tasks; second, in terms of cognitive objectives and evaluation criteria for test items, a theoretical framework capable of accurately depicting the depth and structural complexity of thinking is urgently needed to achieve effective differentiation and objective measurement of higher-order thinking abilities.

2.1 Contextualized Test Questions: A Necessary Path for Competency Evaluation

The fundamental transformation of China's Gaokao evaluation system is symbolically marked by the construction of the *China Gaokao Evaluation System*, which systematically addresses the issues of "why to assess, what to assess, and how to assess" through the "One Core, Four Layers, Four Wings" framework. The "One Core" refers to the core functions: fostering virtue and nurturing talents, selecting talents for higher education, and guiding teaching, which establish the fundamental purpose and direction of the Gaokao. The "Four Layers" refer to the assessment content: core values, disciplinary competencies, key abilities, and essential knowledge, specifying the scope of assessment and emphasizing the evaluation of students' qualities in comprehensively applying disciplinary thinking and methods to solve complex problems under value guidance. The "Four Wings" refer to the assessment requirements: foundational, comprehensive, applied, and innovative, requiring that test design balance foundation and innovation, theory and practice. Context serves as the core carrier of the Four Layers and Four Wings. The principle of "no context, no test question" has become a basic consensus in contemporary test development.

From the perspective of task context types, systematic classification research has formed a multi-level framework. Yang Fan et al., when explaining the reform pathway for the biology subject examination of the Gaokao, proposed a fundamental three-part classification: life practice contexts, learning exploration contexts, and scientific experiment and inquiry contexts.[3]This framework aligns closely with the Gaokao evaluation system and has been widely applied across disciplines. Other scholars have conducted more detailed classification studies. Cao Xiaoxiao, based on an in-depth analysis of Gaokao mathematics questions, proposed that contexts can be subdivided into more specific types, such as personal life, public common sense, production practice, social issues, scientific frontiers, mathematical culture, and historical materials, aiming to comprehensively assess students' knowledge, abilities, and competencies through contextual diversity.[4] In the PISA scientific literacy assessment, contexts are divided into three levels: personal, regional/national, and global, and further embedded into four major thematic areas: health, natural resources, environmental hazards, and scientific and technological frontiers.

2.2 SOLO Taxonomy: An Effective Tool for Externalizing Implicit Thinking

The SOLO taxonomy was proposed by Biggs and Collis (1982) to assess students' learning outcomes and to guide teaching processes and methods. This theory holds that students' learning outcomes can be categorized into five levels (Table 1). It focuses not on how much knowledge students have acquired, but on the complexity of their thinking structures. SOLO test items are measurement tools that apply the fundamental ideas of the SOLO taxonomy to evaluate students' open-ended responses accordingly.[5]

Table 1. Five levels of the SOLO taxonomy

Level	Description
Prestructural	The student has little to no understanding of the problem.
Unistructural	The student responds using one relevant piece of information.
Multistructural	The student responds using multiple relevant but independent pieces of information.
Relational	The student integrates multiple relevant pieces of information to address complex problems.
Extended abstract	The student can abstractly generalize from the problem and apply it to new situations.

Biggs initially used the SOLO taxonomy as a scoring tool for open-ended questions. However, he found that direct use of raw open-ended questions presented several drawbacks: heavy scoring workload, difficulty in standardizing criteria, and low scoring reliability. To address these issues, he innovatively developed the SOLO taxonomy from a "scoring tool" into a "framework for test question development." The specific operational steps are shown in Table 2.

Table 2. Process of developing SOLO test questions

Step	Specific Operation
Step 1: Decompose the problem	Decompose a complex open-ended problem into four sub-questions according to the four levels of thinking complexity (unistructural, multistructural, relational, extended abstract).
Step 2: Sequence the questions	Arrange these four sub-questions from easy to difficult, from concrete to abstract, forming a question chain. Each subsequent question builds upon the previous one.
Step 3: Align with levels	Each sub-question explicitly corresponds to one SOLO level, together forming a complete SOLO test question.

Based on the logical structure among the sub-questions, evaluators can clearly determine a student's thinking structure level from their response performance. Ideally, the relationship between response patterns and corresponding thinking levels is shown in Table 3.

Table 3. Relationship between student response patterns and corresponding SOLO levels

	Sub-Q1	Sub-Q2	Sub-Q3	Sub-Q4	SOLO Level
Pattern 1	Incorrect	Incorrect	Incorrect	Incorrect	Prestructural
Pattern 2	Correct	Incorrect	Incorrect	Incorrect	Unistructural
Pattern 3	Correct	Correct	Incorrect	Incorrect	Multistructural
Pattern 4	Correct	Correct	Correct	Incorrect	Relational
Pattern 5	Correct	Correct	Correct	Correct	Extended abstract

If a student answers the first sub-question correctly but fails the others, it indicates the unistructural level. If a student answers the first two correctly, it indicates the multistructural level, and so on. However, "non-allowed response patterns" may occur, where a student answers a relational sub-question correctly but fails a multistructural one. Possible reasons include issues with the test question design (e.g., the relational sub-question being less difficult than the multistructural one) or student-specific errors. For student-specific errors, the closest response pattern can be used for matching. For test question quality, the Guttman scale can be used for assessment.

3.Research Process and Findings

This study selected Deepseek as a representative AIGC tool and used two approaches to deliver instructions: one-step instruction and step-by-step instruction. For the step-by-step approach, the study first provided relevant literature on SOLO taxonomy, SOLO test question development, and content and academic requirements from the compulsory education science curriculum standards for the specific topic as attachments. The AIGC was then asked

to formulate a SOLO-based evaluation goal for the topic, specifying what should be achieved at each level (prestructural, unistructural, etc.). Second, literature related to the problem context was provided, and the AIGC was asked to design a problem context based on the methods and requirements of contextualized test question development, and to decompose the problem into four sub-questions corresponding to the SOLO levels. The one-step approach involved providing all the above literature and materials simultaneously and requesting the generation of a contextualized SOLO test question. The specific prompt compositions are shown in Table 4.

Table 4. Prompt content

Approach	Prompt	Components
One-step	You are a science education assessment expert. Based on SOLO taxonomy and compulsory education science core competencies, develop a contextualized SOLO test question on "properties of carbon dioxide." The four sub-questions should correspond to the four levels, with progressively increasing difficulty, and should be interlocking.	Role definition + based on (reference) + complete (action) + achieve (goal) + satisfy (requirements)
Step-by-step	Step 1: You are a science education assessment expert. Based on SOLO taxonomy, formulate a SOLO-based evaluation goal for "properties of carbon dioxide" (based on curriculum standards).	Role definition + based on (reference) + complete (action) + achieve (goal) + satisfy (requirements)
	Step 2: Learn the methods for developing contextualized test questions. 1. Design an authentic, interesting, and challenging contextual background for "properties of carbon dioxide." 2. Based on this background, decompose the core problem into four SOLO sub-questions (unistructural, multistructural, relational, extended abstract).	Based on (reference) + achieve (goal) + satisfy (requirements)

3.1 Analysis of Question Contexts

Under the one-step instruction approach, the generated problem context belonged to the scientific experiment type. In terms of authenticity, it met the criteria for a high-quality problem context. However, from the perspective of novelty, the experiments were largely textbook experiments with which students would be highly familiar. Moreover, students could answer the sub-questions without relying on the context. Therefore, this context was still a low-quality problem context overall.

Under the step-by-step instruction approach, the AIGC first generated a problem context titled "Intelligent Greenhouse Gas Monitoring and Conversion System." Compared to the one-step approach, this context exhibited higher novelty and lower overlap with students' learning materials. However, it suffered from insufficient scientific accuracy and forced connections between sub-questions, leading to answers being implied in the question stem. After human intervention, the revised context demonstrated enhanced novelty and inquiry potential. It could assess students' higher-order thinking by guiding them to analyze the reasons why two substances cause dough to "expand," summarize their similarities and differences, and infer their uses.

4.Research Findings and Implications

Through a comparative analysis of AIGC's performance under one-step and step-by-step instruction modes, this study verifies the potential and challenges of AIGC in developing contextualized SOLO science test questions for junior high schools.

4.1 AIGC-Enabled Development: Coexistence of Strengths and Limitations

This study confirms that AIGC demonstrates unique strengths in enabling the development of contextualized SOLO test questions, but its application also has clear boundaries. First, AIGC demonstrates high efficiency in providing question development materials, inspiring ideas, and rapidly generating initial drafts. Particularly under

step-by-step instructions, AIGC can understand the structured requirements of the SOLO taxonomy and generate problem contexts and sub-question frameworks with progressive thinking levels and a degree of novelty. It can provide teachers with high-quality starting points and blueprints, greatly reducing their cognitive load in designing contexts and question chains from scratch, demonstrating the significant potential of AIGC in the "assistant" role.

4.2 Future Optimization Directions: Toward Refined and Structured Prompt Engineering

Based on the results of this study, future efforts to use AIGC to assist question development can optimize prompt design in the following ways. First, provide structured examples and negative lists: The prompt should include not only the theoretical text of the SOLO taxonomy but also one or two positive and negative examples. Positive examples demonstrate how high-quality contexts and sub-questions interlock and how to phrase questions appropriately; negative examples explicitly indicate issues to avoid, such as providing answers in the stem or using overly guiding language. Contrastive learning can more effectively align AIGC's output with expectations.

Second, implement step-by-step iteration and further refinement: For example, after generating the context, add an instruction: "Please review the scientific accuracy of the above context and point out any statements inconsistent with scientific common sense or junior high school students' cognitive level." After generating the sub-questions, add: "Please check that the question chain proceeds from easy to difficult and from concrete to abstract, and that each subsequent question builds upon the previous one." This chain of iterative, self-reviewing instructions can effectively improve output quality. Finally, embed domain-specific knowledge constraints when necessary: Explicitly incorporate core requirements of subject teaching into the instructions. By binding instructions to domain-specific knowledge—such as key concepts, common errors, and teaching priorities—the content generated by AIGC can be made more pedagogically useful.

References:

[1] Zheng Gengbiao.A Preliminary Exploration of History Learning Assessment Design Based on Generative Artificial Intelligence[J].History Teaching(First Half Monthly),2024,(03):20-29.

[2] Jia Linzhi.Analysis of the Characteristics of"Authentic Contexts"and"Authentic Problems"in Core Competency Assessment:Taking Biology as an Example[J].Shanghai Educational Research,2020,(09):77-81.

[3] Pan Hong,Zhang Yalin.Contextualized Chemistry Test Questions in Secondary Schools:Significance,Elements, and Design[J].Teaching Reference of Middle School Chemistry,2022,(11):66-71.

[4] Cao Xiaoxiao.An Analysis of the Value,Quality,and Assessment Status of Contexts in Gaokao Mathematics Test Questions[J].Teaching and Administration,2025,(01):63-66.

[5] Wu Weining,Li Jia,Kong Huisi.Development and Quality Testing of SOLO Test Questions[J].Educational Measurement and Evaluation(Theory Edition),2009,(03):45-48.