Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 333–340
https://doi.org/10.54941/ahfe1007533

Automated Generation of Situational Judgment Tests for Civil Aviation Flight Attendants Using Large Language Models: Method and Preliminary Evaluation

Yaqian Liu[1,2], Qida Hao[3], Jian Cheng[3,4], Cuixia Ma[3,4], Bo Jia[5], Peiru Chen[5], Gang Jie[5], and Jingyu Zhang[1,2]

[1]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[2]Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[4]School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[5]Technology Application Research and Development Center Co., Ltd., China Eastern Airlines, Shanghai, 201700, China

ABSTRACT

In the field of civil aviation, the psychological competency characteristics of cabin crew members are directly related to service quality and flight safety. Although situational judgment tests (SJTs) have proven to be an effective assessment method, their development is costly and time-consuming. The breakthroughs in large language models (LLMs) offer new opportunities for the automated development of assessment tools. Using verbatim transcripts from critical incident interviews with frontline flight attendants as the primary data source, this study aims to construct and validate a retrieval-augmented generation (RAG)-driven workflow for automatically generating SJT items. An expert evaluation approach was employed to assess the quality of items generated by three large models (Model 1: qwen3-14b; Model 2: qwen3-32b; Model 3: deepseek-r1-32b). The results provide preliminary evidence for the feasibility of an automated development pathway for psychological assessment tools based on LLMs and RAG technology, which can significantly improve item development efficiency. However, this study represents an initial exploration, and further research as well as validation through large-scale empirical data are required to optimize and enhance model performance.

Keywords: Large language models (LLMs), Situational judgment test (SJT), Automated item generation (AIG), Retrieval-augmented generation (RAG), Flight attendant competency, Psychometrics

INTRODUCTION

Against the backdrop of rapid global development in the air transportation industry, increasingly complex operational environments, and growing

Received February 7, 2026; Revised April 1, 2026; Accepted April 16, 2026; Available online July 20, 2026

passenger service demands, cabin service work is characterized by high pressure, high uncertainty, and high responsibility. Flight attendants are required not only to handle passenger service situations but also to assume safety management responsibilities during emergencies. The psychological competency characteristics of cabin crew members are directly related to service quality and flight safety (Ji et al., 2019). However, in current flight attendant selection processes, airlines predominantly rely on traditional psychological assessment tools (such as self-report questionnaires like the Big Five Inventory). The results of such assessments are susceptible to social desirability bias and response strategies, which undermine their authenticity and predictive validity, making it difficult to efficiently and accurately evaluate the comprehensive capabilities of cabin crew in complex scenarios.

The Situational Judgment Test (SJT) is a personnel selection method that simulates realistic work situations and presents a series of possible responses. The test requires participants to rate, rank, or choose among these responses based on their judgment of what would be most effective in that situation (Qi & Dai, 2003). By simulating real task scenarios, SJTs serve as an effective tool for measuring competencies and have been shown to outperform traditional self-report questionnaires in predicting actual job performance (Xu et al., 2024). However, their development heavily relies on experts, making it costly and time-consuming (Zhang et al., 2024). As a result, the application of SJTs in civil aviation remains at an early stage and faces numerous challenges.

Recent breakthroughs in large language models (LLMs) present new opportunities for the automated development of assessment tools (Lee et al., 2025). First, LLMs can generate diverse, contextually relevant, and high-quality test items based on large-scale text corpora with minimal human intervention, significantly reducing both production costs and time. Second, LLMs can retrieve and process up-to-date information in real-time (e.g., industry trends, regulatory changes) and can be continuously trained or updated to regularly generate new situational items, ensuring the timeliness and variety of test content. Third, LLMs can learn to extract key information from responses and perform standardized scoring, reducing human subjectivity in the rating process, enhancing scoring consistency, and providing immediate feedback.

However, existing LLM-based approaches still face core challenges: (1) While efficient, methods driven by base large language models often lack domain-specific accuracy and are prone to generating "hallucinations," failing to ensure that the content aligns with real-world regulations and practices in civil aviation (Burke, 2025; Jiang & Feng, 2025). (2) There is a lack of technical mechanisms for effectively integrating professional domain knowledge into the AI generation process, resulting in generated test items that often fall short in situational authenticity, content validity, and measurement depth required for high-stakes personnel selection.

Retrieval-Augmented Generation (RAG) addresses these issues by retrieving relevant information from external knowledge bases in real-time during the generation process, providing the model with reliable information support and effectively reducing inaccuracies and ungrounded content (Lewis et al., 2020). By incorporating retrieved domain knowledge into the prompt

as contextual information, RAG enhances the quality of generated content. Combining LLMs with RAG allows for the integration of professional domain knowledge into the item generation process, mitigates model hallucinations, lowers training costs, and improves the professionalism, accuracy, interpretability, and traceability of generated materials (Wu et al., 2025).

Based on this, the present study aims to apply a large language model approach grounded in Retrieval-Augmented Generation (RAG) to construct and validate an automated method that integrates domain knowledge with LLM generation capabilities, enabling the efficient, batch production of situational judgment test items for civil aviation flight attendants.

METHODS

Knowledge Base Construction

First, the system acquires three types of core raw data:

(1) Interview Transcripts: These comprise verbatim transcripts from one-on-one in-depth interviews with 30 frontline flight attendants. Each transcript, ranging from 10,000 to 20,000 Chinese characters, is rich with critical incidents and behavioral descriptions from real-world work scenarios.
(2) Existing Item Bank: A set of Situational Judgment Test (SJT) items validated through either public or internal experiments. This bank includes item stems, response options, answer keys, and explanatory rationales.
(3) Relevant Literature and Industry Reports: This includes professional works and academic papers from fields such as psychology and organizational behavior, as well as aviation industry documents, providing theoretical support and professional context.

Subsequently, differentiated data processing and ingestion strategies were applied based on the characteristics of each data source. Finally, a hybrid knowledge base system was established, consisting of two major components:

(1) Structured Database (PostgreSQL): This centrally stores key incidents and competency features extracted from the interviews, along with standardized SJT items. The database supports precise SQL queries, correlation analysis, and statistical operations.
(2) Vector Knowledge Base: This stores semantic vector embeddings of academic literature and relevant documents. It supports fuzzy queries and associative knowledge discovery via natural language, providing a theoretical foundation for understanding queries and generating content.

Automated Item Generation

Building upon the domain knowledge base, a Retrieval-Augmented Generation (RAG)-driven workflow for the automated generation of SJT items was constructed. The core of this workflow lies in integrating unstructured

domain-specific experience with the generalized generative capabilities of a Large Language Model (LLM). Specifically, when the system receives an instruction to generate an item targeting a specific competency (e.g., emotion management), it first retrieves the precise definition of that competency and associated key incident snippets from the structured database. Simultaneously, it retrieves examples of style and format from the historical high-quality item bank. These retrieved domain-specific pieces of information serve as strongly constraining context and are input into the LLM alongside the generation prompt, guiding the model to generate standardized SJT items with situations that closely align with actual job requirements and well-designed response options. Furthermore, an automated deduplication mechanism based on semantic similarity is embedded within the workflow to ensure the novelty of generated items and the diversity of the final item bank. For this study, three different LLM architectures were employed independently to perform the automated generation task for flight attendant SJT items: Model 1: qwen3-14b; Model 2: qwen3-32b; Model 3: deepseek-r1-32b.

Item Quality Evaluation

The quality of the LLM-generated items was comprehensively assessed using an expert evaluation method. Four experts in the fields of psychometrics and industrial-organizational psychology independently evaluated a set of 60 test items (20 randomly selected from each model's output). The expert panel rated the items based on a predefined item evaluation scale. This scale, developed by integrating evaluation criteria from relevant literature (Gorgun & Bulut, 2025; Zhang et al., 2024), encompasses 10 evaluation indicators. Each indicator was rated on a 1–5 point Likert scale, with higher scores indicating better performance on that indicator as judged by the experts.

The specific evaluation indicators are as follows:

(1) Language Fluency: The fluency of the language expression in the item.
(2) Language Conciseness: The succinctness and brevity of the item's language.
(3) Language Compliance: The degree to which the language is standard and free from bias or inappropriate expression.
(4) Logicality: The rigor and coherence of the logical structure within the item text.
(5) Situation Clarity: The clarity and specificity of the situational description.
(6) Situation Plausibility: The reasonableness and realism of the depicted situation.
(7) Stem-Option Correspondence: The degree of match between the item stem and the provided response options.
(8) Scoring Rationality: The reasonableness of the scoring criteria assigned to each response option.
(9) Content Validity: The extent to which the item content aligns with the definition of the measured trait.
(10) Discriminatory Power: The item's ability to differentiate between individuals of varying ability levels.

Results

Overall Performance Across Evaluation Indicators

Based on the descriptive statistical results, the overall performance of the three item-generation models across all indicators was relatively favorable. The performance was particularly strong in the areas of Language Fluency, Logicality, Situation Clarity, and Stem-Option Correspondence. The overall performance of the test items on the ten evaluation indicators is presented in Table 1.

Table 1: Overall performance across evaluation indicators.

Indicator	Mean	SD
Language Fluency	4.12	0.60
Language Conciseness	3.82	0.87
Language Compliance	2.84	0.76
Logical Consistency	3.95	0.70
Scenario Clarity	3.91	0.87
Scenario Reasonableness	3.76	1.07
Alignment between Stem and Options	3.99	0.72
Grading Reasonableness	3.59	1.03
Content Validity	3.68	0.88
Discrimination	3.19	1.02

Table 2: Mean values of different models across evaluation indicators.

Indicator	qwen3-14b	qwen3-32b	deepseek-r1-32b
Language Fluency	3.85	4.23	4.27
Language Conciseness	3.50	3.83	4.13
Language Compliance	2.75	3.00	2.77
Logical Consistency	3.75	3.98	4.13
Scenario Clarity	3.50	4.08	4.15
Scenario Reasonableness	3.45	3.85	3.98
Alignment between Stem and Options	3.85	3.97	4.15
Grading Reasonableness	3.55	3.65	3.58
Content Validity	3.65	3.68	3.72
Discrimination	3.25	3.15	3.18
Number of times ranked first	1	2	7
Average score across indicators	3.51	3.74	3.81

Based on the data presented in Table 2, it can be concluded that Model 3 (deepseek-r1-32b) demonstrated optimal performance across seven indicators: Language Fluency, Language Conciseness, Logicality, Situation Clarity, Situation Plausibility, Stem-Option Correspondence, and Content

Validity. Its strengths are primarily reflected in language quality, logical coherence, situational design, and content validity. Model 2 (qwen3-32b) exhibited the best performance in controlling Language Compliance and held a slight advantage in Scoring Rationality. Model 1 (qwen3-14b) showed a marginal advantage in the Discriminatory Power indicator. However, its performance was relatively weaker across most other indicators.

Comparative Analysis of Difficulty Levels

Furthermore, we predefined the difficulty levels of the generated items. Under each testing dimension across different models, the item sets included questions at both easy and difficult levels. During the expert evaluation phase, each expert was asked to subjectively assess the overall difficulty of each item based on their own knowledge and experience. This difficulty assessment also employed a 1–5 point scale, where a higher score indicated that the expert perceived the item as more challenging for test-takers to answer correctly.

Table 3: Question difficulty ratings under different preset difficulty levels.

Model	Preset Difficulty Level	Mean	SD
qwen3-14b	Easy	2.40	0.88
qwen3-14b	Difficult	2.65	0.99
qwen3-32b	Easy	2.20	0.89
qwen3-32b	Difficult	2.50	0.83
deepseek-r1-32b	Easy	2.40	0.94
deepseek-r1-32b	Difficult	2.60	0.99
Overall	Easy	2.33	0.90
Overall	Difficult	2.58	0.93

Based on the results presented in Table 3, regardless of whether considering the three models collectively or examining them individually, the difficulty ratings assigned by experts to items pre-defined as "difficult" were consistently higher than those assigned to items pre-defined as "easy." This finding provides preliminary evidence that the difficulty level settings for the assessment items are consistent with expectations.

CONCLUSION

Regarding item generation efficiency, the average time to generate a complete SJT item (including the stem, four response options, and scoring rubric) was approximately 2 minutes. This is significantly lower than the time required for manual authoring, demonstrating the substantial potential of this method for rapidly constructing large-scale item banks. From the perspective of overall item quality evaluation, expert assessment results indicate that the current items have achieved a high standard in linguistic characteristics and logicality, suggesting a solid quality foundation in basic design elements.

However, the relatively lower scores on indicators such as Discriminatory Power and Scoring Rationality suggest that there is still room for improvement in the psychometric properties of the items. The average score for the difficulty indicator aligns with the intended design goals, reflecting the reasonableness of the item difficulty settings. The existing models were able to achieve the target of difficulty grading as intended, providing a reliable foundation for subsequent item development and quality control.

In terms of model comparison, the three models exhibited distinct performance differences. Model 3 (deepseek-r1-32b) demonstrated the most outstanding performance across several key indicators, including language quality, logicality, and situational design, highlighting its potential as a superior framework for item design. Therefore, we recommend prioritizing Model 3 (deepseek-r1-32b) as the primary framework for item design in practical applications, particularly in scenarios with high demands for language quality and situational realism.

Concurrently, relying solely on the expert evaluation method for item assessment has inherent limitations, especially concerning psychometric indicators such as content validity, difficulty, discriminatory power, reliability, and validity. Future work could involve refining the item generation model based on expert feedback and administering the items in large-scale testing to obtain sufficient data for statistical analysis. This would facilitate continuous improvement in the scientific rigor and effectiveness of the generated items.

In conclusion, this study provides preliminary evidence for the feasibility of an automated development pathway for psychological assessment tools centered on LLMs and RAG technology. This approach not only significantly enhances item development efficiency but also, through the integration of massive critical incidents, ensures a high degree of relevance between item content and actual job requirements. It holds promise for achieving personalized and dynamically updatable assessment tools. However, this research remains at an early stage. The performance of the generated items on more stringent psychometric indicators, such as predictive validity, requires further validation through large-scale empirical testing. Optimization and refinement of model details will necessitate continued research in the future.

ACKNOWLEDGMENT

The authors would like to acknowledge.

This study was supported by Natural Science Fundation of China (U2133209).

REFERENCES

Burke, C. M. (2025). AI-assisted exam variant generation: A human-in-the-loop framework for automatic item creation. *Education Sciences, 15*(8), 1029. https://doi.org/10.3390/educsci15081029

Gorgun, G., & Bulut, O. (2025). Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educational Measurement: Issues and Practice, 44*(1), 96~107. https://doi.org/10.1111/emip.12663

Ji, M., Liu, B., Li, H., Yang, S., & Li, Y. (2019). The effects of safety attitude and safety climate on flight attendants' proactive personality with regard to safety behaviors. *Journal of Air Transport Management, 78*, 80–86.

Jiang, Z., & Feng, S. (2025). UsmleGPT: An AI application for developing MCQs via multi-agent system. *Software Impacts, 23*, 100742.

Lee, P., Son, M., & Jia, Z. (2025). AI-powered Automatic Item Generation for Psychological Tests: A Conceptual Framework for an LLM-based Multi-Agent AIG System. *Journal of Business and Psychology,* 1–29.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Qi, S. Q., & Dai, H. Q. (2003). The property, function and the development of situational judgment tests. *Psychological Exploration, 23*(4), 42–46.

Wu, Y., Yu, T., Yu, Q., Lu, Y., Zhang, X., Huo, S., ... & Li, J. (2025). A Study on question and answer of a large language model for spleen and stomach diseases based on retrieval-augmented generation technology. *World Chinese medicine, 20*(19), 3516–3523.

Xu, J., Luo, F., Ma, Y., Hu, L., & Tian, X. (2024). Automated scoring of open-ended situational judgment tests. *Acta Psychologica Sinica, 56*(6), 831.

Zhang, Z., Tu, Z., Chen, Y., Yiyao, X., Feng, Y., & Zhang, W. (2024). Automated Item Generation for Personality Assessment: Development and Validation of Large-Language-Model-Derived HEXACO Situational Judgment Tests. Available at SSRN 5378520.