Open Access

Public Organizations Management

Vol. 14(3), (Series 55): 25-44/ 2026

https://doi.org/10.30473/ipom.2026.75081.5213

E-ISSN: 2538-600X P-ISSN: 2322-522X

ORIGINAL ARTICLE

Validation of the Artificial Intelligence–Driven Delphi Method in Management Research Based on the Turing Test

Hamed Dehghanan[1], Zahra Pouramini[*2]

1. Associate Professor, Department of Management and Accounting, Allameh Tabataba`i University, Tehran, Iran.
2. Ph.D. Student, Department of Management and Accounting, Allameh Tabataba`i University, Tehran, Iran.

***Correspondence**
Zahra Pouramini
E-mail: z.pooramini@ut.ac.ir

Receive Date: 06/July/2025
Revise Date: 15/Oct/2025
Accept Date: 19/Oct/2025

How to cite
Dehghanan, H., Pouramini, Z (2026). Validation of the Artificial Intelligence–Based Delphi Method in Management Research Based on the Turing Test. *Public Organizations Management*, 14(3), 25 -44.

EXTENDED A B S T R A C T

Introduction

In the past decade, the expansion of artificial intelligence (AI) and large language models (LLMs) has brought a fundamental transformation to qualitative research methods and expert-based decision-making processes. The Delphi method, traditionally grounded in the collective judgments of human experts, has entered a new phase of evolution with the advent of AI. The present study aimed to design and validate an AI-based Delphi method while leveraging the imitation game, or Turing test, as a tool to assess the credibility of its outputs. The central question guiding this research was:
"How can the Turing test be utilized as a novel approach to evaluate the validity and similarity of AI-based Delphi results compared to traditional human-based Delphi?"
This question is particularly significant because researchers engaging with modern language models face a fundamental challenge: whether AI-generated responses can be considered as reliable as human judgments in terms of logic, reasoning, and coherence. This study seeks to provide a well-grounded answer and establish a methodological framework for the combined use of Delphi and AI in applied research contexts.

Mothodology

This research is of an applied-developmental nature and was conducted with the objective of validating the AI-based Delphi method through the Turing test. The study population consisted of 30 purposively selected participants, including human resources experts, AI specialists, and independent evaluators, ensuring the data could be examined from human, technical, and impartial judgment perspectives.

In prior studies, the AI-based Delphi methodology was comprehensively designed and explained, and in a separate study, it was applied to prioritize the developmental needs of human resources managers in the context of coaching. In the current study, the method is only briefly introduced, and the data derived from previous research serve as input for the Turing test.

In the AI-based Delphi process, large language models (ChatGPT, Copilot, and Gemini) acted as virtual experts, responding to research questions and presenting their reasoning explicitly through structured prompts. Subsequently, the outputs of these models were validated using the Turing test. Evaluators engaged in brief dialogues without knowing the source of the responses (human or AI) and were required to determine the nature of the respondent. Responses that could not be confidently distinguished were considered "successful" in the Turing test.

Thus, this study represents a continuation of prior research, examining the similarity, coherence, and credibility of AI-based Delphi outputs in a real-world research setting using actual data and the Turing test.

Findings

Findings indicated that 56% of participants were female and 44% male, with a mean age of 36 years (range: 27–52) and approximately 72% holding a master's degree or higher. This diversity of expertise and experience provided a robust basis for analyzing AI model behavior in comparison with human responses. The data employed were secondary sources, comprising results from three rounds of traditional and AI-based Delphi in previous studies by the authors, focused on identifying developmental priorities for human resources managers. These datasets included competency ratings, comparisons between traditional and AI-based approaches, and reliability and validity matrices, with over 85% of the data confirmed as reliable and valid.

During the study, key questions were collected from both human and AI sources at each Delphi stage, and evaluators were tasked with distinguishing between human and machine responses, in line with the Turing test framework. Qualitative analysis revealed that AI-generated responses were often so similar to human answers that evaluators struggled to distinguish them, with an overall Turing test success rate of only 23%, indicating that over two-thirds of responses could not be correctly classified. This finding highlights the strong ability of the models to simulate human behavior.

Content analysis showed that evaluators primarily relied on cues such as grammatical and stylistic errors, vocabulary diversity, interactive responsiveness, display of emotion and empathy, variation in tone, and references to personal experience—features more prevalent in human responses. In contrast, AI responses were characterized by high linguistic precision, formal structure, uniform tone, accurate and error-free answers, consistent response speed, and absence of personal narratives or experiential examples. Behavioral and linguistic patterns further revealed that participants expected human responses to be more dynamic, personal, and occasionally non-linear, whereas machine responses were predominantly formal, precise, and structured.

Discussion and Conclusion

These findings not only confirm the capability of large language models to emulate human-like behavior and response style but also underscore the significance of linguistic, behavioral, and emotional cues in distinguishing between humans and AI. Furthermore, integrating human and AI Delphi data and analyzing evaluator judgments demonstrated that AI models could produce highly coherent responses, logically consistent themes, and alignment with management literature, establishing their effectiveness as tools for analyzing developmental needs of human resources managers.

Overall, the results indicate that the AI-based Delphi method, combined with the Turing test, can effectively evaluate the validity and similarity of machine-generated responses to human ones, enabling the production of reliable and credible data in real-world research settings.

As the first study to employ the Turing test for evaluating a research methodology, this study demonstrates that AI-based Delphi can generate data that are reliable and highly similar to human responses. The findings show that

large language models successfully emulate human behavior and response style, making it challenging for participants to distinguish human from machine responses. This evidence strengthens the credibility of the AI-based Delphi method, suggesting it can serve as an effective research tool without compromising the quality of human-like analysis.

Based on the findings, practical recommendations include optimizing the design of Turing test scenarios and criteria to more accurately assess the capabilities of language models, integrating the Delphi method with the Turing test for practical validation of outputs, fully documenting the process and managing potential biases, attending to ethical considerations and data confidentiality, and developing and standardizing prompt-engineering skills for effective interaction with language models. Additionally, potential limitations—such as the influence of participants' attitudes and experiences, ethical and confidentiality constraints, sensitivity of results to prompt design and model characteristics, limitations in sample diversity and size, and the inability to fully control the knowledge and biases embedded in language models—should be considered in the design and analysis of future studies.

These findings provide practical guidance for researchers employing AI-based Delphi methods, enabling the combined use of human and AI collective intelligence with higher accuracy and reliability.

K E Y W O R D S
Artificial Intelligence, Delphi Method, Turing Test, Personal Development.