

Evaluating Score Dependability in ChatGPT-Supported AP Chinese Speaking Tasks

Dan Song
University of Iowa
dan-song@uiowa.edu

Won-Chan Lee
University of Iowa
won-chan-lee@uiowa.edu

Abstract

This study applied generalizability theory to examine score variation and dependability in AP Chinese speaking tasks completed with and without ChatGPT support. Although ChatGPT-supported tasks were associated with higher scores, the NoGPT condition consistently exhibited higher dependability coefficients. Increasing the numbers of tasks and raters further improved score dependability.

1 Introduction

The increasing use of generative artificial intelligence (GAI) tools such as ChatGPT has created new opportunities and challenges for second-language learning and assessment (Song & Tang, 2025). Large language models (LLMs) are capable of generating human-like responses and following complex instructions, leading to their growing use in educational settings (Ouyang et al., 2022; Bulut et al., 2024). In language learning contexts, ChatGPT may support learners by generating ideas, providing vocabulary, and organizing responses. Although such support may enhance writing and speaking performance, it also raises important questions regarding score interpretation and score dependability (Jiao et al., 2026; Song et al., 2026).

Reliability is a fundamental concern in language assessment because test scores should reflect learners' language ability rather than irrelevant sources of variation (Bachman, 1990). In performance-based assessments, score variation may arise from task characteristics, rater severity, and interactions among these facets (Weigle, 2002). Generalizability (G) theory provides a useful framework for examining multiple sources

of measurement error and evaluating score dependability under alternative measurement conditions (Brennan, 2001b; Cronbach, 1972).

Despite the growing interest in AI-assisted language learning and assessment, relatively little is known about the dependability of scores obtained from speaking tasks completed with AI support. The following section reviews several previous research on ChatGPT in language assessment and the application of G theory in performance assessment.

2 Literature Review

2.1 ChatGPT in Language Assessment

The emergence of GAI tools such as ChatGPT has attracted increasing attention in language education and assessment. LLMs can generate human-like responses and provide learners with support for brainstorming, vocabulary selection, and response organization (Ouyang et al., 2022). Previous studies have suggested that AI-assisted learning may improve learners' performance in second-language writing tasks (Song & Tang, 2025). In addition to supporting language production, GAI tools may influence how assessment scores are interpreted. Learners can use ChatGPT to generate ideas, improve vocabulary, and organize responses, potentially enhancing task performance.

Although such support may facilitate language learning, it may also blur the distinction between learners' independent abilities and externally assisted performance. Additionally, researchers have also highlighted the importance of examining the validity and dependability of scores obtained under AI-supported conditions (Bulut et al., 2024). Understanding how AI support affects score interpretation has become an increasingly important issue in language assessment.

In addition, several recent research has examined the role of LLMs in assessment contexts, including automated scoring and AI-assisted learning. For example, Jiao et al. (2026) investigated rater effects across multiple LLMs and found differences in scoring behaviors among AI models. Similarly, Song et al. (2026) applied G theory to automated essay scoring and demonstrated that score reliability varied across scoring configurations. Although research on AI in language assessment continues to expand, relatively little attention has been given to the dependability of scores obtained from speaking tasks completed with AI support.

2.2 Generalizability Theory in Performance Assessment

G theory extends classical test theory by decomposing observed score variation into multiple sources of measurement error (Brennan, 2001b; Cronbach, 1972). Unlike classical test theory, which provides a single estimate of reliability, G theory enables researchers to examine the contributions of different facets, such as tasks and raters, to score variation. The framework consists of a G-study, which estimates variance components, and a D-study, which evaluates score dependability under alternative measurement conditions.

Compared with classical test theory, G theory offers greater flexibility for investigating multiple sources of measurement error in performance assessments (Brennan, 2001b; Shavelson & Webb, 1991). In writing and speaking assessments, score variation may arise not only from differences among learners but also from task difficulty, rater severity, and their interactions. By estimating variance components associated with these facets, G theory provides evidence regarding the consistency and interpretability of assessment scores. In addition, D-studies allow researchers to evaluate how changes in measurement conditions, such as the number of raters or tasks, influence score dependability. These features make G theory a suitable method for investigating AI-supported assessment contexts.

Because performance assessments involve multiple sources of measurement error, G theory has been widely applied in language assessment research. Previous studies have shown that task characteristics, rater severity, and interactions among these facets may influence score

dependability (Bachman, 1990; Weigle, 2002). In writing and speaking assessments, G theory provides a useful framework for determining whether observed score differences primarily reflect meaningful differences among learners or unintended sources of variation. However, little is known about how AI-supported speaking tasks influence score variation and dependability. Therefore, the present study applies G theory to investigate score dependability across AP Chinese speaking tasks completed with and without ChatGPT support.

3 Research Questions

Building on previous research and the identified gap in the literature, the present study addressed the following two research questions (RQs):

RQ1. To what extent do participants, tasks, raters, and their interactions contribute to variation in Overall scores obtained from AP Chinese speaking tasks completed with and without ChatGPT support?

RQ2. How does score dependability change under different combinations of tasks per condition and raters, and what differences are observed between the NoGPT and GPT conditions?

4 Method

4.1 Participants and Data

The dataset consisted of 50 participants who completed four AP Chinese speaking tasks. For each task, participants were given four minutes to prepare and two minutes to present their responses. Tasks 1 and 2 were completed without ChatGPT support, whereas Tasks 3 and 4 were completed with ChatGPT support.

The four speaking tasks were selected from the Cultural Presentation free-response section of recent AP Chinese Language and Culture examinations. Participants were asked to present information related to Chinese cultural topics. In the ChatGPT condition, participants were permitted to use ChatGPT during the preparation period to generate ideas and organize their responses before delivering their presentations.

Each response was independently scored by four trained human raters using an Overall score

ranging from 0 to 6. Because each participant completed four tasks and each task was rated by four raters, the complete dataset contained 800 score observations. The use of multiple raters enabled the examination of rater-related variance and score dependability using G theory.

4.2 Preliminary Descriptive Analysis

Before conducting the G-study, participant-level descriptive statistics were calculated to compare Overall scores across conditions. For each participant, mean scores were computed separately for tasks completed with and without ChatGPT support by averaging across the two tasks and four raters within each condition. A repeated-measures analysis of variance (ANOVA) was subsequently conducted to examine whether score differences between conditions were statistically significant.

4.3 G-Theory Design

A univariate $p \times t \times r$ design was implemented in this study. Participants served as the object of measurement, whereas tasks and raters were treated as measurement facets. The G-study estimated variance components associated with participants, tasks, raters, and their interactions. Following the G-study, D-studies were conducted to examine score dependability under alternative measurement conditions by varying the number of tasks per condition and the number of raters. Generalizability coefficients (E_p^2) and phi coefficients (Φ) were estimated for both the NoGPT and GPT conditions.

5 Results

5.1 Descriptive Statistics

Participant-level descriptive statistics for Overall scores are presented in Table 1. For each participant, mean scores were calculated by averaging across two tasks and four raters within each condition. Participants obtained higher Overall scores on tasks completed with ChatGPT support ($M = 3.82$, $SD = 1.06$) than on tasks completed without ChatGPT support ($M = 2.96$, $SD = 1.24$).

Condition	N	Mean	SD	Min	Max
ChatGPT	50	3.82	1.06	1.62	6.00
No ChatGPT	50	2.96	1.24	0.63	6.00

Table 1: Overall Scores by Condition

To further examine score differences between conditions, a repeated-measures ANOVA was conducted. The result revealed a statistically significant effect of condition, $F(1, 49) = 40.64$, $p < .001$. Participants obtained significantly higher Overall scores on tasks completed with ChatGPT support than on tasks completed without support.

In addition, participants in the GPT condition demonstrated slightly lower score variability ($SD = 1.06$) than those in the NoGPT condition ($SD = 1.24$). This pattern suggests that ChatGPT-supported responses were associated with somewhat more consistent performance across participants.

5.2 G-Study Results

The G-study estimated variance components associated with participants, tasks, raters, and their interactions. Table 2 presents the estimated variance components.

Effect	NoGPT	GPT
Participant (p)	1.377	0.952
Task (t)	0.094	0.014
Rater (r)	0.053	0.042
Participant $\times$ Task (pt)	0.142	0.185
Participant $\times$ Rater (pr)	0.101	0.203
Task $\times$ Rater (tr)	-0.001	-0.004
Participant $\times$ Task $\times$ Rater (ptr + residual)	0.387	0.331

Table 2: Variance Components for Overall Scores

Based on the result, participant variance accounted for the largest proportion of score variation under both conditions, with variance component estimates of 1.378 for the NoGPT condition and 0.952 for the GPT condition. These results suggest that Overall scores primarily reflected meaningful differences among students. In contrast, task and rater effects were comparatively small. Task variance was estimated at 0.095 under the NoGPT condition and 0.014 under the GPT condition, whereas rater variance was 0.054 and 0.042, respectively. These findings indicate that tasks and raters contributed relatively little systematic variance to Overall scores.

However, interaction effects and residual variance represented additional sources of measurement error. For example, participant-by-task interaction variance was estimated at 0.142 for the NoGPT condition and 0.185 for the GPT condition, which suggested that participants performed somewhat differently across speaking

tasks. Similarly, participant-by-rater interaction variance increased from 0.101 under the NoGPT condition to 0.203 under the GPT condition, which indicated that some participants were affected differently by individual raters. Residual variance was also substantial, with estimates of 0.387 and 0.331 for the NoGPT and GPT conditions, respectively. These findings highlight the importance of considering interaction effects when evaluating the quality of speaking assessment scores.

Overall, the results provide evidence that the majority of score variation was attributable to differences among learners rather than unintended sources of measurement error.

5.3 D-Study Results

To evaluate score dependability under alternative measurement conditions, D-studies were conducted by varying the numbers of tasks and raters. Table 3 presents the estimated Ep^2 and Φ across different task and rater configurations. The values in the first column represent the number of tasks included within each condition. Thus, a value of 1 indicates that one task was used within the NoGPT condition and one task was used within the GPT condition, whereas a value of 2 indicates that two tasks were used within each condition.

Number of Tasks	Number of Raters	NoGPT Ep^2	NoGPT Φ	GPT Ep^2	GPT Φ
1	1	0.781	0.731	0.678	0.662
1	2	0.781	0.731	0.678	0.662
1	3	0.819	0.768	0.724	0.709
1	4	0.839	0.787	0.750	0.736
2	1	0.790	0.747	0.674	0.651
2	2	0.863	0.825	0.775	0.758
2	3	0.891	0.855	0.816	0.801
2	4	0.905	0.870	0.838	0.825

Table 3: D-Study Results Across Tasks and Raters

As shown in Table 3, score dependability generally improved as the numbers of tasks and raters increased. Across all configurations, the NoGPT condition consistently exhibited higher coefficients than the GPT condition. The highest coefficients were observed when two tasks and four raters were included, with $Ep^2 = 0.905$ and $\Phi = 0.870$ for the NoGPT condition, and $Ep^2 = 0.838$ and $\Phi = 0.825$ for the GPT condition. Although score dependability improved with additional tasks

and raters, the improvements became smaller as more raters were added.

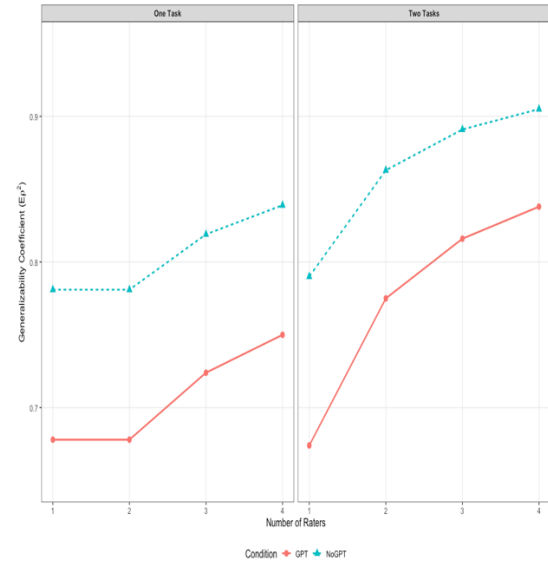


Figure 1. Ep^2 by Tasks and Raters

In addition, Figure 1 further illustrates the changes in Ep^2 across task and rater configurations. Under both task conditions, score dependability generally improved as the number of raters increased. Across all configurations, the NoGPT condition consistently exhibited higher coefficients than the GPT condition. Nevertheless, both conditions displayed similar patterns of growth, which suggested that the relative differences in score dependability remained generally stable across task and rater configurations.

6 Discussion

Participant variance represented the largest source of score variation, whereas task and rater effects were relatively small. This suggests that Overall scores primarily reflected meaningful differences among participants. The relatively small task and rater effects also indicate that the Overall scoring rubric functioned consistently across raters and speaking tasks. These findings suggest that sources of measurement error other than participant differences contributed relatively little to Overall scores.

The D-study further showed that score dependability generally improved as the numbers of tasks and raters increased, although the amount of improvement became smaller as more raters were added. From a practical perspective, these findings suggest that increasing the numbers of tasks and raters may enhance score dependability,

but the benefits become smaller as additional raters are introduced. Therefore, decisions regarding measurement design should balance score precision with practical considerations such as scoring efficiency, time, and resource availability.

The results also demonstrate the value of G theory for evaluating score quality in AI-supported assessment contexts. Unlike classical test theory, G theory makes it possible to examine multiple sources of measurement error simultaneously and to evaluate score dependability under alternative measurement conditions. As generative AI tools become increasingly integrated into language learning environments, understanding how AI support influences score quality and score interpretation will become increasingly important. Overall, these findings support the usefulness of G theory for investigating score quality in AI-supported language assessment contexts.

At the same time, although participants obtained higher scores on tasks completed with ChatGPT support, these differences should be interpreted cautiously. Because the ChatGPT and non-ChatGPT conditions were administered in a fixed order, the effects of ChatGPT support cannot be separated from potential task-order or practice effects.

7 Limitations and Future Research

Several limitations should be considered. Firstly, the study was based on a sample of 50 participants and four speaking tasks. Secondly, task order was confounded with ChatGPT condition, which prevented causal interpretations of the observed score differences. Furthermore, the analysis focused solely on overall scores. Future research may consider employing larger samples, counterbalanced designs, and analytic scoring domains to provide a more comprehensive understanding of the effects of ChatGPT support on speaking performance.

Future research may also examine analytic score domains, including task completion, delivery, and language use, to determine whether AI support influences different aspects of speaking performance. In addition, incorporating a larger number of tasks and raters would provide further evidence regarding score dependability. Qualitative analyses of how learners interact with ChatGPT during task preparation may also contribute to a deeper understanding of AI-supported language assessment.

8 Conclusion

Using a univariate $p \times t \times r$ design, this study applied G theory to examine score dependability across AP Chinese speaking tasks completed with and without ChatGPT support. Participant variance accounted for the largest proportion of score variation, and score dependability improved as the number of tasks and raters increased. Although participants obtained higher scores on ChatGPT-supported tasks, the observed differences should be interpreted cautiously given the study design and associated limitations. Overall, the findings provide preliminary evidence regarding score interpretation in AI-supported speaking assessments.

As generative AI tools become increasingly integrated into language learning environments, the present study highlights the value of G theory as a framework for investigating multiple sources of measurement error and evaluating score quality in AI-supported assessment contexts.

References

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Brennan, R. L. (2001b). *Generalizability theory*. New York: Springer-Verlag.
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., ... & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *arXiv preprint arXiv:2406.18900*.
- Cronbach, L. J. (1972). *The dependability of behavioral measurements. Theory of generalizability for scores and profiles*. 1-33.
- Jiao, H., Song, D., & Lee, W. C. (2026). Evaluating Rater Effects of Large Language Models in Automated Essay Scoring: GPT, Claude, Gemini, and DeepSeek. *Educational Measurement: Issues and Practice*, 45(2), e70018.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35, 27730-27744.
- Shavelson, R. J., & Webb, N. M. (2012). Generalizability theory. In *Handbook of complementary methods in education research* (pp. 309-322). Routledge.
- Song, D., Lee, W. C., & Jiao, H. (2026). Exploring LLM autoscoring reliability in large-scale writing assessments using generalizability theory.

International Journal of Testing, 1-26.
<https://doi.org/10.1080/15305058.2026.2684584>

Song, D., & Tang, A. F. (2025). Exploring AI-assistance in L2 Chinese writing with standardized assessment tasks. *Language Learning & Technology*, 29(2), pp. 161–189.
<https://hdl.handle.net/10125/73615>

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.