

Reliability Study of Chinese Writing Tests Based on Generalizability Theory

Xuan Li*

Department of Humanities, Dalian University, Dalian 116000, Liaoning, China

**Author to whom correspondence should be addressed.*

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Chinese writing tests have strong subjectivity, and the scoring results are easily affected by various factors such as differences in raters, types of test tasks, and testing environment conditions. Traditional reliability calculation methods cannot accurately identify where the errors come from. Generalization theory can distinguish between measurement objectives and measurement related aspects, estimate the variance of multiple error components simultaneously, and provide guidance for optimizing evaluation plans through generalization research (G research) and decision research (D research). This article comprehensively summarizes the main content and theoretical advantages of generalization theory, and breaks down the error structure of Chinese writing tests from three dimensions: differences among candidates themselves, measurement errors, and interaction effects. After that, specific strategies are proposed to improve reliability, such as building a stable scoring system, optimizing question design, standardizing scoring processes, and establishing quality monitoring mechanisms. The research results indicate that the generalization theory provides accurate diagnosis and targeted optimization methods for Chinese writing tests, and has a positive effect on promoting the scientific and fair development of international Chinese education assessment.

Keywords: Conceptual theory; Chinese writing test; Reliability; Scoring error; International Chinese education

Online publication: April 26, 2026

1. Introduction

Writing ability is an important component of language proficiency, and writing tests are subjective language skill tests that not only reflect learners' language output ability, but also their overall level, testing their comprehensive language proficiency. Compared to other language skills, such as listening, speaking, and reading, research on writing ability testing is relatively weak and is a difficult point in language proficiency testing ^[1]. Writing tests are also a key form of assessment in international Chinese education, which can directly assess learners' proficiency in Chinese character usage, grammar, logic, and vocabulary reserves. It can also reflect students' cognitive and cultural understanding, help teachers grasp learning situations, adjust teaching plans, provide an objective basis for Chinese proficiency assessment and talent selection, and promote the improvement of the quality of teaching Chinese as a foreign language. However, in actual teaching, the grading of Chinese writing is easily influenced by multiple factors such as personal preferences of the grader, differences in understanding of the grading criteria, degree of question openness, and response scenarios, which can lead to insufficient consistency in grading and unstable evaluation results. For a long time, most related studies have used methods

such as the reliability coefficient and correlation coefficient for evaluation, which can only obtain the overall reliability level and cannot answer key questions such as “where the error comes from” and “where to make improvements”. Generalization Theory breaks the limitations of classical measurement theory by treating measurement errors as systematic factors that can be decomposed, analyzed, and controlled. Distinguishing between measurement targets and measurement profiles, it helps researchers accurately identify the sources of errors and guides the optimization of evaluation schemes. Introducing the theory of generalization into the research of Chinese writing tests in international Chinese language teaching not only helps to improve the scientificity and fairness of the assessment, but also promotes the development of international Chinese education evaluation towards a more accurate and standardized direction. Therefore, this article does not conduct empirical data analysis, but only systematically elaborates on the application of generalization theory in Chinese writing tests from a theoretical perspective.

2. The core connotation and theoretical advantages of generalization theory

The core idea of the generalization theory proposed by Cronbach et al. is that any measurement result is not absolute, but comes from a sampling of a larger “measurement domain”.

Ao Yongqian (2008) summarized the development origin and core structure of Generalized Theory (GT), proposing that it is an extension and improvement of the classical Test Theory (CTT). The article provides a detailed explanation of the birth background, core concepts (such as global score, generalization coefficient, reliability index), research design (such as single-sided crossover design, multi-faceted random design), and its advantages compared to CTT - the ability to simultaneously estimate the variance components of multiple error sources and improve the measurement process based on this result ^[2].

Ge Du (2004) introduced the first monograph in China that systematically studied the theory of generalization, “Generalization Theory and Its Applications in Evaluation” (Yang Zhiming, Zhang Lei, 2003). This book compared the shortcomings of classical test theory, fully explained generalization research, decision research, and multi-faceted error decomposition methods, and applied these methods to practical scenarios such as the college entrance examination and teaching evaluation. In Chinese writing tests, these methods can be directly used to analyze the reliability of raters, determine the most suitable number of raters, and control measurement errors caused by task types and scoring criteria, thereby providing systematic methodological support for improving the scientificity and fairness of writing tests ^[3]. This lays a solid theoretical foundation for this article: the generalization theory is very suitable for analyzing complex measurement scenarios such as Chinese writing tests with multiple potential errors in raters, task types, grading criteria, etc. The “generalized decision-making” approach mentioned in this theory can help subsequent research determine the appropriate number of raters and task sample size, and improve testing efficiency while ensuring reliability. Ao Yongqian’s (2008) theoretical summary laid the methodological foundation for using generalization theory to decompose the error variance of Chinese writing scores and estimate reliability coefficients in subsequent empirical research ^[2]. This study divides total variation into two parts: the measurement target and the measurement side. The measurement objective is generally the actual ability of learners; The measurement aspect refers to external conditions that may affect the score, such as raters, test questions, question types, time, environment, etc.

Yang Hongbo et al. (2020) used the theory of multivariate generalization to analyze the “Chinese Language Subject Knowledge and Teaching Ability” test paper in the primary and secondary school teacher qualification examination. They found that there were differences in the measurement reliability of different ability modules, among which the “Teaching Design” module had relatively higher reliability. Increasing the number of questions in this module can effectively improve the reliability of the entire test paper. This study provides methodological references for the modular design and reliability optimization of Chinese writing tests ^[4]. Zhu Shi (2022) used the theory of multiple generalizations to analyze the sources of errors in the grading of college English reading courses using a self-developed classroom teaching quality evaluation form, and determined the appropriate number of assessors. Although the subject of this study is English reading,

its approach of using mGENOVA software to analyze measurement errors and improve evaluation tools can directly provide methodological references for the study of rater reliability in Chinese writing tests ^[5]. Xu Changhuo and Tao Jingrong (2019) used the theory of generalization and concluded through the G study that the writing ability of candidates themselves is the main source of score differences, but the errors caused by the rater factor are greater than those caused by the task type. Through D research (decision research), it was further found that the reliability of argumentative essay testing was the highest, while the reliability of narrative essay testing was the lowest. When the number of evaluators is 3, the generalization coefficients of all four task types can reach the ideal standard. This study provides a direct methodological reference for adjusting the number of raters and optimizing task design in Chinese writing tests ^[6].

Overall, compared with traditional methods of measuring reliability, generalization theory is more suitable for subjective tests such as Chinese writing. Firstly, it can simultaneously handle multiple sources of error and will not only provide a vague and general reliability value. Secondly, it can analyze the interactive effects between different factors, such as situations where “raters give varying degrees of grading to candidates of different levels” and “there are interactions between question types and raters”. Generalization theory can directly provide guidance for practical testing work, clearly telling test designers whether to increase the number of raters, optimize question settings, or strengthen training in the scoring process. The ability to accurately locate problems and make targeted adjustments is difficult to achieve with traditional reliability methods.

3. Analysis of error sources in Chinese writing tests

The study of generalization theory can be divided into two stages. The first stage is the G study, which is a conceptual study. The main task is to identify all sources that can cause variation in the results, and then calculate the proportion of each source. These sources of variation include measurement targets, various measurement aspects, and their interactions, in order to clarify the causes of measurement errors. The second stage is D research (decision research), which adjusts measurement conditions based on G research, such as adding evaluators, changing the number of questions, changing task types, etc., to find the most reliable and cost-effective evaluation plan ^[7]. Based on the analysis framework of generalization theory, the errors in Chinese writing tests can be systematically sorted out into three levels: measurement objectives, measurement aspects, and interaction effects.

3.1. Measurement target level: The variation brought by the candidate themselves

This article sets the measurement target as the true writing ability of candidates, and the differences in this part are reasonable and have practical significance. The performance of candidates in the test will be affected by factors such as their on-site state, emotions, adaptability to the questions, fatigue level, and answering strategies, which can lead to unstable performance on site ^[8]. This type of error belongs to individual internal differences, and although it cannot be eliminated, its impact can be reduced by using a stable examination process and clear task instructions.

3.2. Measurement at the lateral level: errors caused by external factors

The measurement aspect is the main source of errors in Chinese writing, and this study includes raters, test tasks, and testing environments.

Firstly, there is a rater error. The subjective judgment of evaluators is one of the most prominent sources of error in writing tests. There are fundamental differences among raters in terms of strictness, language preferences, and evaluation habits; Inconsistent understanding of scoring criteria; Easy to experience convergence effect, halo effect, and contrast effect; Rating fatigue, emotions, and order can also affect rating stability. These will directly lead to score distortion.

Secondly, there is an error between the test questions and the tasks. There are differences in openness, topic familiarity, prompt information, and genre requirements among different writing tasks. Open-minded argumentative essays have a high degree of flexibility in grading after reading; The scoring of picture-based writing and situational writing with

strong closure is relatively stable. Vague instructions and unclear tasks in the questions will further widen the scoring gap.

Thirdly, there are other side errors. The testing time limit, answering environment, whether the grading process is standardized, and whether it is anonymous can indirectly affect the consistency of scores.

3.3. Interaction effect error: The most easily overlooked key error

The interaction effect is a key focus of conceptual theory and an important source of scoring errors in Chinese writing. For example, raters give stricter scores to candidates with higher levels and more lenient scores to candidates with lower levels. When the same grader scores different genres of essays, the standards used are not uniform, and the difficulty of different exam questions for candidates with different native language backgrounds also varies. These interactive effects cannot be recognized by traditional reliability methods, but they will significantly affect the fairness of the evaluation.

4. Optimization strategy for the Chinese writing test based on generalization theory

4.1. Build a stable and controllable scoring system

Firstly, implement a dual rater or multi-rater system and implement anonymous blind evaluation to reduce personal bias. Secondly, unify the scoring criteria and provide sample essays of different levels to help evaluators understand the specific manifestations of “good, medium, and poor”. Thirdly, conduct pre-scoring calibration training and unify the criteria for severity. Fourthly, establish a scoring and sampling mechanism to promptly correct deviations.

4.2. Optimize test design to reduce scoring flexibility

It is best to ensure that the task is clear, the instructions are clear, and the situation is specific. The proportion of closed writing questions can be appropriately increased, such as picture-based writing, element-specific writing, and application writing. Reduce questions that are too open and cannot be scored uniformly. Propositions should be tailored to learners’ daily learning and life, reducing additional scoring errors caused by cultural backgrounds and unfamiliarity with the topic.

4.3. Standardize the scoring process and control systematic deviations

Implement graded scoring and multi-dimensional sub-item scoring, scoring separately from dimensions such as content, language, structure, and writing, to reduce the bias caused by overall impression scores. Control the duration and workload of grading to reduce fatigue in grading. Strictly implement the dual evaluation and triple evaluation system, and arbitration must be conducted if the score difference exceeds the threshold.

4.4. Establish a normalized quality monitoring mechanism

Schools and examination institutions can regularly apply the concept of generalization theory to review grading errors, analyze rater consistency, question differentiation effectiveness, interaction bias, etc. Continuously improve the scoring criteria, proposition quality, and grading process to achieve dynamic improvement in evaluation quality.

5. Conclusion

The Chinese writing test itself has strong subjectivity and complex sources of error. Traditional measurement methods cannot meet the requirements of scientificity and fairness. Generalization theory will be analyzed from three levels: measurement objectives, measurement aspects, and interaction effects, which can accurately identify the sources of errors and provide clear directions for optimizing the reliability of writing assessment. In the international Chinese education context, optimizing the reliability of Chinese writing tests based on generalization theory can help improve grading consistency, reduce subjective bias, and make evaluation results more fair. Future research can continue to combine

multivariate generalization theory with empirical research to further analyze the comprehensive impact of multiple dimensions, such as questions, raters, and task types, promoting the development of Chinese writing assessment towards a more standardized, accurate, and professional direction.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Han YY, 2020, Empirical Study on the Validity of Writing Test Methods for Teaching Chinese as a Foreign Language, thesis, Shandong Normal University.
- [2] Ao YQ, 2008, A Review of Generalization Theory Research. *Journal of Wanxi University*, (2): 49–52.
- [3] Ge D, 2004, A New Quantitative Research Model Beyond Classical Testing Theory: A Review of “Generalization Theory and Application of Evaluation”. *Educational Development Research*, (Z1): 148–149.
- [4] Yang HB, Luo CY, Yang Z, 2020, Analysis of the Quality of the “Chinese Language Subject Knowledge and Teaching Ability” Paper in the Qualification Examination for Primary and Secondary School Teachers: A Perspective Based on Classical Measurement Theory and Multivariate Generalization Theory. *Examination Research*, 16(4): 46–52.
- [5] Zhu S, 2022, The Application of Multivariate Generalization Theory in the Evaluation and Design of English Reading Courses in Chinese Universities. *Modern English*, (21): 1–4.
- [6] Xu CH, Tao JR, 2019, The Impact of Task Type and Scorer Factor on Writing Scoring in Chinese Writing Test. *The Journal of Chinese Language and Literature*, 2019: 115.
- [7] Zheng XQ, Liu YZ, Li GM, 2025, Validity Testing and Evaluation Scheme Optimization of Document Basket Test Concept Based on Multivariate Generalization Theory. *Journal of Jiangxi Normal University (Natural Science Edition)*, 49(6): 623–634.
- [8] Kong FY, 2022, A Study on the Scoring of Chinese Writing for Preparatory Studies in China: Based on Generalization Theory and Multi-dimensional Rasch Model. *Examination Research*, (4): 41–51.

Publisher's note

Whioce Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.