BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100248

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/

# Expert consensus and reliability validation of the portfolio assessment guideline for Chinese practical writing: An empirical study based on fleiss' kappa

Ying Wang (王莹)[a,b,1,*], Ibnatul Jalilah Yusof[b]

[a] Department of Humanities and Arts Education, Hebei Finance University, Lianchi District, Baoding, Hebei Province, China
[b] Faculty of Educational Science and Technology, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

ARTICLE INFO

*Keywords:*
Assessment guideline
Chinese practical writing
Fleiss' kappa
Portfolio assessment
Reliability validation

ABSTRACT

*Purpose:* Portfolio assessment has been increasingly recognized as an effective approach to fostering comprehensive writing ability. However, its application in Chinese practical writing remains limited. The lack of standardized evaluation criteria has hindered its reliability and broader implementation. This study aimed to systematically develop and validate a Portfolio Assessment Guideline for Chinese practical writing, focusing on inter-rater reliability and coverage across four core dimensions: content, logical structure, language, and format.
*Methods:* Five higher education experts with extensive experience in practical writing instruction and research independently rated the guideline and its scoring rubrics for two key genres (summary and official notice), and inter-rater agreement was assessed using Fleiss' Kappa coefficients.
*Findings:* The Kappa values for six core modules ranged from 0.79 to 1.00, with an overall Kappa of 0.87 across 14 sub-dimensions, indicating "almost perfect" agreement. Genre-specific analysis showed high overall consistency for summary (κ=0.87) and official notice (κ=0.89), with the summary's "logical structure" dimension achieving "substantial agreement" (κ=0.68). Based on expert feedback, descriptive indicators were refined without altering the core framework.
*Value:* The findings provide robust evidence for the psychometric quality of the guideline, supporting its potential application in higher education and professional training for enhancing Chinese practical writing abilities.

## 1. Introduction

Portfolio assessment, as a continuous paradigm, integrates the assessment of both learning processes and competency development by systematically collecting learners' multiple outputs over a given period (e.g., drafts, revisions, reflective records). It has been widely applied in the field of language education, particularly in writing assessment [1,2]. Compared with traditional standardized testing, portfolio assessment offers the advantage of capturing learners' performance in authentic contexts, thereby balancing both product-oriented and process-oriented values [3,4]. This approach is particularly well aligned with the nature of practical writing. As functional texts, Chinese practical writings (e.g., plans, notices, requests) emphasize genre conventions, situational appropriateness, and accuracy of expression. The development of such ability relies on multiple rounds of writing practice and revision based on feedback, which calls for an evaluation tool that values both process and reflection [5].

In China's higher education system, practical writing courses serve as a core vehicle for cultivating students' professional communication ability, covering a wide range of genres such as governmental documents, administrative writings, and news reports [6]. However, existing research indicates two major limitations in current approaches to assessing Chinese practical writing: first, the lack of a specialized evaluation framework, as most assessments rely heavily on instructors' subjective judgment with blurred definitions of key dimensions such as "thematic content" and "formatting conventions"; and second, insufficient validation of assessment tools, particularly regarding inter-rater reliability, which undermines the objectivity and fairness of the results

Peer review under the responsibility of The International Open Benchmark Council.
* Corresponding author: Department of Humanities and Arts Education, He Finance University, Lianchi District, Baoding, Hebei Province, China.
*E-mail address:* ying20@graduate.utm.my (Y. Wang).
[1] This research was completed while the first author, Wang Ying, was pursuing her doctoral degree at Universiti Teknologi Malaysia.

https://doi.org/10.1016/j.tbench.2025.100248

Available online 2 December 2025
2772-4859/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

[5]. Inter-rater reliability, as a core indicator of the quality of any assessment tool, essentially ensures that different raters' judgments of the same performance are not unduly influenced by individual biases. It is therefore a prerequisite for the widespread application of assessment instruments [7].

Fleiss' Kappa, an extension of Cohen's Kappa, is specifically designed to test the consistency of categorical ratings among multiple raters ($\geq$3). By controlling for chance agreement, it has been established as a reliable metric for measuring inter-rater agreement in fields such as medical diagnosis and educational measurement [8–10]. However, its application in the field of Chinese writing assessment remains rare, particularly in reliability testing of portfolio-based assessments of practical writing. Existing studies mostly adopt Pearson correlation coefficients or Cohen's Kappa, which are insufficient to address the reliability requirements of multi-rater, multi-dimensional evaluations.

Against this backdrop, the present study systematically examines the inter-rater reliability of the Portfolio Assessment Guideline for Chinese Practical Writing developed in earlier research, using Fleiss' Kappa as the statistical method. Specifically, the study addresses the following research questions:

Do the six core modules of the guideline demonstrate reliable inter-rater consistency when evaluating four key abilities: thematic content, logical structure, linguistic expression, and formatting conventions?

For the two representative genres, summary and official notice, how consistent are raters' judgments with regard to the specific assessment indicators defined in the guideline?

Based on the consistency analysis, what directions for refinement and implications for wider application can be derived for the guideline?

## 2. Literature review

### 2.1. Portfolio assessment in writing education

The value of portfolio assessment in writing education has been widely recognized. Its core advantage lies in overcoming the limitations of traditional tests, which capture only "one-shot performances." By longitudinally tracking learners' textual production (e.g., drafts, revisions, reflective journals), portfolio assessment provides a dynamic picture of competency development [1,2]. In higher education contexts, portfolio assessment serves both formative and summative purposes: the former enhances learning improvement through continuous feedback, while the latter enables a holistic judgment of competency achievement based on multiple sources of evidence [11]. Empirical studies have shown that this approach effectively improves students' metacognitive awareness, writing autonomy, and learning motivation [12].

From an international perspective, portfolio assessment has yielded abundant findings in ESL/EFL writing. It not only helps address specific weaknesses of learners at different proficiency levels (e.g., sentence construction among lower-proficiency learners) [13], but also, through innovative forms such as e-portfolios, strengthens feedback interaction and peer review, thereby promoting the comprehensive development of writing skills ranging from basic to advanced [14,15]. Comparative studies further confirm its advantages over traditional summative assessments in enhancing learners' writing complexity, accuracy, fluency, and self-efficacy [16].

In the Chinese writing domain, exploration of portfolio assessment began relatively late [17]. Early studies primarily focused on basic education, such as Yu [18], who proposed an implementation procedure for "composition growth portfolios" in high school, and Wei [19], who designed a primary school e-portfolio on the Moodle platform. Although some studies (e.g., [20]) have attempted to apply portfolio assessment to teaching Chinese as a second language, confirming its positive effects on promoting learners' autonomy and reflective ability, the field still faces clear limitations: (i) a scarcity of empirical research, with most studies confined to theoretical frameworks or process design [21]; (ii) limited application in higher education, especially in practical genres such as functional writing; and (iii) insufficient scientific validation of assessment outcomes, as the lack of quantitative analysis makes it difficult to establish reliability and validity for broader adoption [22].

Overall, while existing research has predominantly focused on the impact of portfolio assessment on learning outcomes, little attention has been given to the reliability mechanisms (e.g., inter-rater consistency) and standardization pathways of portfolio-based assessment in Chinese practical writing. This research gap provides the central entry point for the present study.

### 2.2. Assessment of practical / functional writing

Practical or functional writing distinguishes itself from literary writing by its explicit communicative purposes, specific pragmatic contexts, and standardized textual structures [6]. The core of this type of writing lies in "situational appropriateness," requiring writers to select suitable genre structures, linguistic styles, and formatting conventions according to communicative needs [23]. In Chinese higher education, practical writing courses cover a variety of genres such as official notices, summaries, and reports. Their assessment must move beyond the "literary orientation" often found in general writing evaluation, instead focusing on functional indicators such as accuracy of information transmission, compliance with formatting standards, and pragmatic appropriateness [5].

Scholars have reached consensus regarding the core components of practical writing ability. Unlike literary writing, which emphasizes creativity, practical writing focuses on four major dimensions: thematic content (relevance and completeness of information), logical structure (conformity to genre frameworks and coherence), linguistic expression (appropriateness and accuracy of register), and formatting conventions (standardization of layout and presentation) [24,25]. This categorization not only echoes the general writing elements such as "form" and "mechanics" proposed by Diederich et al. [26], but also reflects the genre-specific characteristics of practical writing, aligning with the requirements for functional writing in China's college entrance examination marking criteria.

Nevertheless, current practices in assessing practical writing reveal significant shortcomings. Recent studies [27] have pointed out the issue of "academic transplantation," whereby existing assessment tools simply borrow dimensions from academic writing (e.g., originality, depth of argumentation), thereby neglecting the functional essence of practical writing. More critically, specialized assessment tools for Chinese practical writing remain scarce, and their reliability, particularly regarding inter-rater consistency in multi-rater settings, has not been rigorously validated from a psychometric perspective. This deficiency undermines the objectivity and comparability of evaluation results [5]. The situation highlights the pressing need to develop standardized assessment guidelines and to rigorously validate their reliability.

### 2.3. Inter-Rater reliability in writing assessment

Inter-rater reliability is a core indicator of the scientific rigor of any assessment tool. It essentially examines whether different raters can make consistent and stable judgments about the same evaluation object [28]. In writing assessment, the rating process is inevitably influenced by subjective factors such as raters' prior experience and their varied interpretations of standards [29]. Therefore, reliability validation becomes a prerequisite for both the development and application of assessment tools. Only when raters can achieve a high level of consensus based on the same criteria can assessment results be regarded as objective and valid.

The choice of consistency testing methods depends on the type of data and the number of raters involved [30]:

**Pearson correlation coefficient** is applicable to continuous data (e.g., scores on a 0–100 scale). It measures the linear association between two raters' scores, but it cannot distinguish between absolute agreement

and trend agreement, nor can it account for chance agreement.

**Cohen's Kappa** is designed for two raters evaluating unordered categorical data (e.g., "excellent / good / average"). By controlling for chance agreement, it provides more accurate results. However, it is limited to two-rater contexts [29].

**Intraclass Correlation Coefficient (ICC)** can be extended to multiple raters and is suitable for continuous or ordinal data. Yet, its applicability to categorical data is weaker, and its results can vary considerably depending on whether an absolute agreement or relative agreement model is selected.

In multi-rater contexts (more than two raters), the limitations of these methods become particularly evident: correlation coefficients cannot control for chance agreement, and Cohen's Kappa requires multiple pairwise comparisons to indirectly infer overall consistency. Both approaches are inadequate for complex evaluation scenarios, such as portfolio assessment, where multiple experts provide categorical ratings across multiple dimensions.

A critical research gap lies in the limited application of appropriate reliability analysis methods in the evaluation of Chinese practical writing. Specifically, there has been insufficient systematic examination of inter-rater agreement on detailed criteria such as "thematic content" and "formatting conventions." As a result, the scientific robustness of existing evaluation tools remains under-validated.

### 2.4. Application of fleiss' kappa in educational assessment

Fleiss' Kappa, an extension of Cohen's Kappa, is specifically designed to measure the consistency of categorical data involving multiple raters ($m \geq 2$), multiple objects ($n \geq 1$), and multiple categories ($k \geq 2$) [8]. Its core advantages are as follows:

By calculating the ratio of observed agreement to expected agreement (chance probability), it effectively eliminates the influence of chance agreement, thereby addressing the limitation of correlation coefficients in controlling for random consistency [31].

It is suitable for scenarios with unequal numbers of raters (e.g., some objects rated by three raters, others by five), making it more flexible than the intraclass correlation coefficient (ICC).

It demonstrates stronger adaptability to categorical data (e.g., "meets requirements / partially meets requirements / does not meet requirements"), which makes it particularly well suited to writing assessment contexts that involve multiple dimensions and multiple rating levels [32].

The interpretive standards proposed by Fleiss [33] have been widely adopted in academic research.Table 1

In the field of educational assessment, Fleiss' Kappa has been widely used for testing inter-rater reliability in language assessment, such as validating the effectiveness of writing scoring rubrics [32] and evaluating the impact of rater training in speaking assessments. Its value in measuring multi-rater agreement on categorical data has also been well established in fields such as medicine and psychology [10].

However, its application in Chinese writing assessment remains rare. In particular, within the complex context of portfolio assessment, which involves multiple text versions and multiple competency dimensions, no studies to date have systematically employed this method to examine the inter-rater reliability of assessment guidelines. This gap leaves the psychometric properties (e.g., reliability) of Chinese practical writing portfolio assessment tools without sufficient empirical evidence, thus constraining their standardization and broader implementation.

## 3. Research methodology

This section elaborates on the research design, participants, assessment instrument, data collection procedures, and analytical methods. The aim is to address the central question: "What is the inter-rater reliability of the Chinese Practical Writing Portfolio Assessment Guideline?" and to ensure the scientific rigor and replicability of the study.

### 3.1. Participants

Content validity is a fundamental criterion in the development of assessment instruments. Its essence lies in verifying, through expert judgment, the degree of alignment between the instrument's content and the intended evaluation objectives. Specifically, this includes the appropriateness of tasks, clarity of standards, alignment with curricular goals, and feasibility of implementation [34,35]. Although no consensus exists regarding the optimal number of experts, Lynn [36] suggested a range of 3–10, while Polit and Beck [37] considered 2–15 to be acceptable. The key requirement is that experts must possess sufficient professional expertise related to the assessment object [38].

In line with these principles, five experts were recruited as raters in this study. All were associate professors or above in the field of Chinese language education at domestic universities, and they met the following qualifications:

More than ten years of teaching experience in practical writing courses, with a solid understanding of the pedagogical characteristics of key genres such as summary and official notice;

Experience in leading or contributing to the compilation of practical writing textbooks, thus possessing systematic knowledge of assessment standards;

Familiarity with the development of writing assessment tools and a sound understanding of the basic logic of reliability testing.

All raters had previously received training in the use of analytic scoring rubrics. However, they had no prior exposure to the specific portfolio assessment guideline under investigation, in order to minimize potential bias and ensure objectivity in scoring.

### 3.2. Assessment instrument

The development of the Portfolio Assessment Guideline for Chinese Practical Writing was driven by the need to address the lack of standardized criteria in the assessment of Chinese practical writing portfolios. The guideline integrates Moya's systemic assessment framework [39], Lam's theory of self-regulated learning [40], and Delett's structured design principles [41], thereby encompassing three essential dimensions: assessment, teaching, and learning.

#### 3.2.1. Portfolio assessment guideline for Chinese practical writing

The development of the guideline followed a "goal-setting – content selection – standard formulation – implementation validation" sequence, with the following core steps:

**Goal setting**: The guideline emphasizes three main goals: (i) enhancing students' ability in four major areas of practical writing, namely content relevance and completeness, organizational structure, linguistic appropriateness and accuracy, and formatting conventions;; (ii) cultivating a self-regulated learning cycle of planning, monitoring, evaluating through reflective journals and multi-source feedback; (iii) providing teachers with multidimensional data to inform pedagogical

**Table 1**
Interpretation of Kappa values for Inter-rater agreement.

| Kappa Value Range | Level of Agreement | Interpretation |
|---|---|---|
| Kappa $< 0.40$ | Poor agreement (Unacceptable) | Indicates significant divergence in raters' understanding of the evaluation criteria. |
| $0.40 \leq$ Kappa $\leq 0.75$ | Intermediate to good agreement (Acceptable) | Suggests that the standards are generally clear but still have room for refinement. |
| Kappa $> 0.75$ | Excellent agreement (Highly reliable) | Indicates strong consensus among raters and highly reliable results. |

adjustments.

**Content structure**: Drawing on Moya's "process–product" evaluation and Lam's SRL cycle, the portfolio consists of three types of materials: writing samples (including annotated drafts, revised drafts with justification, and final drafts with format checklists), reflective journals (records of learning strategies and self-evaluations of goals), and feedback records (self-assessment checklists, peer review forms, and teacher diagnostic reports).

**Scoring system**: Based on the standards of the Chinese College Entrance Examination and the specific features of practical writing, the scoring rubric comprises four dimensions and eleven indicators, supported by a three-level rating scale (with descriptors for each level).

**Implementation management**: In accordance with Delett's principle of "task–standard alignment," the teaching activities were designed as a cyclical sequence of input, output, reflection, revision. The portfolio adopts a "genre–category" filing system, organized by weekly timeline, to ensure systematic collection and management of materials.

#### 3.2.2. Expert rating form

To examine the inter-rater reliability of the guideline, an Expert Rating Form was designed based on its core components. The form specifies six modules: assessment goals, portfolio content, evaluation standards, instructional procedures, management guidelines, and implementation recommendations, which were further operationalized into 14 rating dimensions (e.g., "clarity of indicator descriptions," "practicality of instructional procedures").

Each dimension employed a 3-point ordinal scale (1 = not applicable/inappropriate, 2 = partially applicable/problematic, 3 = applicable/appropriate). This design enabled the collection of expert judgments on the consistency of each dimension, in alignment with the analytical requirements of Fleiss' Kappa.

### 3.3. Research procedures

#### 3.3.1. Rater training

A two-hour standardized workshop was conducted to ensure consistency among raters. The training included: an overview of the guideline's development background and core objectives; operational definitions of the four scoring dimensions and eleven indicators; instructions on the use of the three-level analytic rubric.

#### 3.3.2. Independent rating and feedback collection

Within one week, raters independently evaluated the Portfolio Assessment Guideline using the Expert Rating Form. All ratings were recorded in electronic forms, and raters were prohibited from any communication during the process. Upon completion, both the rating forms and qualitative feedback from the experts were collected to support subsequent data analysis and guideline refinement.

### 3.4. Data analysis

Data analysis was conducted using Excel, following these steps:

**Data organization**: Expert ratings were converted into a "rater–dimension" matrix (rows = raters, columns = rating dimensions, cells = rating levels).

**Reliability calculation**: Fleiss' Kappa values were computed for (a) the six core modules, (b) the fourteen rating dimensions as a whole, and (c) the four primary writing dimensions (content, structure, language, format) across two key genres (summary and official notice).

**Result interpretation**: Consistency levels were evaluated according to Fleiss' criteria (<0.40 = poor, 0.40–0.75 = fair to good, >0.75 = excellent) [33]. Dimensions with low agreement were further analyzed in light of the experts' qualitative feedback to identify directions for refinement.

## 4. Results

### 4.1. Overall consistency of the assessment guideline

Fleiss' Kappa analysis revealed that the inter-rater consistency coefficients across the six core modules of the guideline ranged from 0.79 to 1.00, with an overall Kappa value of 0.87 across the 14 dimensions (Table 2). According to Fleiss' standard (≥0.75 = excellent), this indicates excellent consistency [33]. Notably, the "Implementation Objectives" and "Instructional Procedures" modules achieved perfect agreement ($\kappa = 1.00$), suggesting a strong consensus among raters regarding the evaluation purposes and operational procedures of the guideline. The "Assessment Criteria" module ($\kappa = 0.84$) demonstrated slightly higher consistency compared to modules such as "Portfolio Content" and "Teacher Toolkit" ($\kappa = 0.79$), indicating that raters' understanding was more stable for the core evaluation dimensions.

### 4.2. Consistency of scoring rubrics for specific genres

To evaluate the effectiveness of the guideline across different genres of practical writing, two types of scoring rubrics: summary and official notice, were examined. Both genres demonstrated high inter-rater consistency (see Table 3). The overall Fleiss' Kappa for official notices ($\kappa = 0.89$) was slightly higher than that for summaries ($\kappa = 0.87$).

At the dimension level, the "Structure" dimension in summaries showed the lowest agreement ($\kappa = 0.68$, good), whereas all other dimensions achieved excellent agreement ($\kappa \geq 0.79$). For official notices, both the "Language" and "Format" dimensions reached perfect agreement ($\kappa = 1.00$), while "Content" and "Structure" achieved excellent agreement ($\kappa = 0.79$).

**Table 2**
Fleiss' Kappa values for each module of the assessment guideline.

| Module | Scoring Dimensions (Items) | No. of Items | κ | Interpretation |
|---|---|---|---|---|
| Implementation Objectives | 1. Clarity of objective statements<br>2. Comprehensiveness of ability cultivation | 2 | 1.00 | Excellent |
| Portfolio Content | 3. Relevance of material types to ability assessment<br>4. Rationality of genre requirements | 2 | 0.79 | Excellent |
| Assessment Criteria | 5. Genre-specific adaptability of rubrics<br>6. Operability of indicator descriptions<br>7. Differentiation of three-level scoring system<br>8. Appropriateness of four-dimension weighting | 4 | 0.84 | Excellent |
| Teacher Toolkit | 9. Scientific rigor of dynamic tracking sheets<br>10. Articulation between rubrics and tracking sheets | 2 | 0.79 | Excellent |
| Instructional Procedures | 11. Guidance effectiveness for ability development<br>12. Validity of multi-source feedback | 2 | 1.00 | Excellent |
| Management Norms | 13. Feasibility of timeline<br>14. Scientific categorization of portfolios | 2 | 0.79 | Excellent |
| Overall | | 14 | 0.87 | Excellent |

**Table 3**
Fleiss' Kappa Values for Specific Genre Scoring Rubrics.

| Genre | Module | Scoring Dimensions (Items) | No. of Items | κ | Interpretation |
|---|---|---|---|---|---|
| Summary | | | | | |
| | Content | 1. Task Coverage<br>2. Problem Objectivity<br>3. Experience Abstraction Level<br>4. Data Support<br>5. Causal Attribution Logic<br>6. Improvement Feasibility | 6 | 0.93 | Excellent |
| | Structure | 7. Framework Coherence<br>8. Hierarchy Clarity | 2 | 0.68 | Good |
| | Language | 9. Expression Objectivity<br>10. Graphic Assistance | 2 | 0.79 | Excellent |
| | Format | 11. Element Completeness | 1 | 1.00 | Excellent |
| | Overall | | 11 | 0.87 | Excellent |
| Official Notice | | | | | |
| | Content | 1. Core Elements<br>2. Information Completeness<br>3. Information Conciseness | 3 | 0.79 | Excellent |
| | Structure | 4. Framework Coherence<br>5. Hierarchy Clarity | 2 | 0.79 | Excellent |
| | Language | 6.Terminology Standardization<br>7. Expression Precision | 2 | 1.00 | Excellent |
| | Format | 8. Letterhead<br>9. Body<br>10. Closing Elements | 3 | 1.00 | Excellent |
| | Overall | | 10 | 0.89 | Excellent |

### 4.3. Evaluation of guideline applicability

Experts reached perfect consensus across the four core indicators of guideline applicability (κ = 1.00), confirming that the guideline: (i) demonstrates a high degree of alignment between assessment dimensions and key abilities; (ii) provides scoring criteria with strong operability; (iii) is suitable for both undergraduate education and professional training contexts; and (iv) holds potential for large-scale implementation.

### 4.4. Analysis of expert feedback

While the core framework of the guideline received strong recognition, expert feedback indicated that certain dimensions required further refinement.

#### 4.4.1. Core issues requiring optimization and revision measures

Within the assessment criteria module, the item "operability of indicator descriptions" was rated "2″ by two experts. For instance, in the "summary" genre, the original standard for "degree of experience extraction" merely referred to "transferable methodology" without specifying "transfer contexts." After revision, the indicator was refined into a three-level description: "transferable principles (including applicable contexts) → superficial experience (context not specified) → non-transferable," supplemented with concrete examples such as "small- and medium-scale campus events."

The original weight of "degree of experience extraction" in the "summary" genre was 5 points, which all five experts considered too low; for the "official notice" genre, the original weight of "degree of information conciseness" was 5 points, with three experts recommending an increase. Following revision, the weight of "degree of experience extraction" was adjusted to 10 points (while maintaining the total weight of the content module), and "degree of information conciseness" was adjusted to 8 points.

The original standard for "clarity of hierarchy" in the "summary" genre mentioned only "graded headings," without covering intra-paragraph logic; in the "official notice" genre, the original description of "loose structure" lacked quantitative benchmarks. After revision, "clarity of hierarchy" was expanded into a composite standard of "graded headings + paragraph coherence + intra-paragraph logic," with explicit penalty criteria such as "absence of a purpose paragraph" and "disordered sequence of key elements."

#### 4.4.2. Points of divergence and resolution

One expert recommended adding genres such as "plans" and "reports" to the portfolio. The research did not adopt this suggestion in order to avoid excessive burden on teachers and students, but plans to address it later by adopting a "core genres + optional genres" model.

One expert argued that "degree of chart/visual aid use" in the "summary" genre is non-essential for purely textual summaries and suggested reducing its weight. In response, the research renamed this dimension "clarity of information presentation," retained its weight, but broadened the scope of evaluation to include textual expression.

## 5. Discussion

### 5.1. Reliability characteristics of the assessment guideline

This study found that the Chinese Practical Writing Portfolio Assessment Guideline demonstrated high inter-rater reliability (overall κ = 0.87), indicating that the dimensional design and descriptors of the guideline are highly clear and operationalizable. This finding is consistent with the conclusions of Lloyd et al. [10], who, in their study of short-answer tasks in large-scale statistics courses, enhanced inter-rater consistency (Fleiss' κ = 0.68) by refining the behavioral descriptors of their rating scale (e.g., a three-tiered operational definition of "step completeness"). Their results confirmed the widely accepted consensus that explicitly defined descriptors are the core prerequisite for reliability. A further comparison suggests that the higher reliability observed in this study (0.87 > 0.68) may be attributed to the guideline's concretized definitions of writing-specific dimensions such as format conventions and logical structure (e.g., "the header must include the issuing number and the authorizing officer"), which reduced the room for raters' subjective interpretation [42].

Regarding differences across modules, the "Implementation Objectives" and "Instructional Procedures" modules achieved perfect agreement (κ = 1.00), not only supporting Rezaei and Lovorn's [29] argument that "clarity of rating criteria determines reliability," but also echoing the core finding of Klenowski et al. [43] that the reliability of portfolio assessment primarily depends on the explicitness of its purposes and processes. These modules focus on the how-to aspects (e.g., "submission of the revised draft in Week 6″), which constitute low-ambiguity tasks requiring minimal subjective inference from raters. By contrast, the κ values for the "Portfolio Content" and "Operational Norms" modules (κ = 0.79) were slightly lower, likely because these modules involve more context-dependent judgments (e.g., "representativeness of portfolio samples"). Such contextual dependency, as highlighted by Eckes [42], is a core source of rater divergence, which indicates that even after training, raters' interpretations of "contextual appropriateness" may still be shaped by their individual experiences.

### 5.2. Genre specificity and rating consistency

The rating consistency of official notices was extremely high (overall $\kappa = 0.89$, with both the language and format dimensions achieving $\kappa = 1.00$). This reflects the fact that the high degree of conventionalization within this genre reduces raters' judgmental difficulty. Such a result aligns with the effect of genre-specific evaluation criteria: research has indicated that compared with general dimensions, genre-specific rubrics (e.g., explicit textual structure and formatting) are more conducive to rater consensus [44].

In contrast, the structural dimension of summary writing yielded a relatively lower κ value (0.68), suggesting that flexible conventions (e.g., logical coherence) are more susceptible to raters' subjective interpretations. This finding echoes Weigle's classic assertion that rating reliability tends to decline in more open-ended genres lacking a unified paradigm [45]. Therefore, future descriptors for such genres should incorporate prototypical exemplars to provide raters with a shared frame of reference.

### 5.3. Implications for practical writing assessment

First, genre-adaptiveness of scoring rubrics is crucial. The findings of this study corroborate the claim of Lloyd et al. [10]: raters achieved a Fleiss' Kappa of 0.68 when employing detailed rubrics in short-answer tasks, with even higher consistency observed in aligned genres (e.g., texts with explicit formats). The high consistency in official notices illustrates that criteria grounded in genre features (e.g., "authoritativeness" and "format completeness") can effectively enhance reliability.

Second, rater training should be genre-specific. For instance, in the structural dimension of summary writing, training should include "prototypical paradigm comparisons" to unify raters' evaluative standards through case-based learning. This aligns with Rezaei & Lovorn's [29] conclusion that training combined with operationalized descriptors is a key pathway to improving reliability.

Finally, assessment tools need to allow for dynamic refinement. Lloyd et al. [10] demonstrated that the continuous revision and supplementation of scoring exemplars can further reduce rating discrepancies. The expert feedback in this study, which emphasized the need to "add positive and negative cases and refine ambiguous descriptors," is highly consistent with this perspective.

### 5.4. Limitations and future research

The limitations of this study include: (i) a relatively small sample of raters ($n = 5$), which may restrict generalizability; (ii) the examination of only two types of practical writing, excluding other common genres such as reports and contracts; and (iii) the absence of further validation of structural validity and discriminant capacity. Future research could: (i) incorporate a larger and more diverse group of raters to test consistency; (ii) expand the genre coverage to build a rubric repository for practical writing; and (iii) adopt mixed methods (e.g., cognitive interviews) to uncover the decision-making mechanisms underlying rating discrepancies.

## 6. Conclusion

Through Fleiss' Kappa analysis, this study confirms that the Portfolio Assessment Guideline for Chinese Practical Writing demonstrates excellent inter-rater reliability, with consistency across its six core modules and two genre-specific scoring criteria reaching or approaching the level of "excellent." This finding provides empirical support for the psychometric quality of the guideline and offers a practical framework for the standardization of Chinese practical writing assessment.

The results further indicate that the degree of rigidity in genre conventions significantly influences rating consistency, highlighting the need for a balanced approach between a "general framework" and "genre-specific indicators" in practical writing assessment. Revisions to the guideline based on expert feedback have enhanced its applicability, making it not only suitable for higher education but also a valuable reference for writing evaluation in professional training contexts.

Future research should validate the stability of the guideline with larger samples and a wider range of genres, and integrate evidence of validity to construct a more comprehensive assessment system. Ultimately, such efforts will advance the evaluation of Chinese practical writing from "experience-based judgment" toward "evidence-based assessment."

## Funding

This paper has no funding support.

## CRediT authorship contribution statement

**Ying Wang:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Ibnatul Jalilah Yusof:** Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] R. Lam, Portfolio Assessment For the Teaching and Learning of Writing, Springer, Singapore, 2018.
[2] L. Hamp-Lyons, W. Con, Assessing the Portfolio Principles For practice, Theory and Research, Hampton Press, 2000.
[3] J.S. Barrot, Effects of Facebook-based e-portfolio on ESL learners ' writing performance, Lang. Cult. Curric. 34 (1) (2021) 95–111.
[4] P. Zhang, G. Tur, A systematic review of e-portfolio use during the pandemic: inspiration for post-COVID-19 practices, Open. Prax. 16 (3) (2024) 429–444.
[5] H.T. Huang, Master 's thesis, Nanning Normal University, 2023, https://doi.org/10.27037/d.cnki.ggxsc.2023.000911.
[6] Y.Y. Wang, Practical Writing Skills and Standards, People 's Posts and Telecommunications Press, 2022.
[7] Räz, T. (2023). Inter-rater reliability is individual fairness. arXiv preprint arXiv: 2308.05458. https://doi.org/10.48550/arXiv.2308.05458.
[8] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull 76 (5) (1971) 378.
[9] A. Zapf, S. Castell, L. Morawietz, A. Karch, Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC. Med. Res. Methodol 16 (93) (2016) 1–10, https://doi.org/10.1186/s12874-016-0200-9.
[10] S. Lloyd, M. Beckman, D. Pearl, R. Passonneau, Z. Li, Z. Wang, Foundations for NLP-assisted formative assessment feedback for short-answer tasks in large-enrollment classes, arXiv preprint arXiv:2205.02829, https://arxiv.org/abs/2205.02829, 2022.
[11] V. Klenowski, Developing Portfolios For Learning and assessment: Processes and Principles, Routledge, 2002.
[12] T. Burner, The potential formative benefits of portfolio assessment in second and foreign language writing contexts: a review of the literature, Stud. Educ. Eval. 43 (2014) 139–149.
[13] I. Listiana, F.N. Yusuf, S.M. Isman, Portfolio assessment: benefits for students At different writing proficiency level, J. Pendidik. Bhs. Dan. Sastra 20 (2) (2021) 243–256.
[14] W. Ngui, V. Pang, W. Hiew, K.W. Lee, Exploring the impact of e- portfolio on ESL students'writing skills through the lenses of Malaysian undergraduates, Comput.-Assist. Lang. Learn. Electron. J. 21 (3) (2020) 105–121.
[15] N. Pourdana, K. Tavassoli, Differential impacts of e-portfolio assessment on language learners ' engagement modes and genre-based writing improvement, Lang. Test. Asia 12 (1) (2022) 7.
[16] B.O.S. Al-Hawamdeh, N. Hussen, N.S.G. Abdelrasheed, Portfolio vs. summative assessment: impacts on EFL learners ' writing complexity, accuracy, and fluency (CAF); self-efficacy; learning anxiety; and autonomy, Lang. Test. Asia 13 (1) (2023) 12.
[17] J. Yang, Master 's thesis, Lanzhou University, 2020, https://doi.org/10.27204/d.cnki.glzhu.2020.003088.
[18] Y. Yu, Establishing a "composition growth portfolio": an attempt to effectively improve middle school students ' writing levels, Chin. Lang. Teach. Middle. Sch. (12) (2007) 39–41. CNKI:SUN:ZYJX.0.2007-12-017.

[19] Y.M. Wei, Master 's thesis, Northeast Normal University, 2012.
[20] Y. Liu, Master's thesis, Shandong Normal University, 2018.
[21] J.F. Mo, The application of"portfolio evaluation. Chinese Language Teaching, Henan Education (Basic Education Edition), 2021, pp. 65–66.
[22] X.H. Wang, Master 's thesis, Central China Normal University, 2019.
[23] W. Grabe, R. Kaplan, Theory and Practice of Writing, 1996. London and New York.
[24] S.Q. Lu, Master 's thesis, Shaanxi Normal University, 2015.
[25] X.X. Zuo, Master 's thesis, Luoyang Normal University, 2023.
[26] P.B. Diederich, J.W. French, S.T. Carlton, Factors in judgments of writing ability, ETS. Res. Bull. Ser. 1961 (2) (1961) i–93.
[27] E.A. Shabani, J. Panahi, Examining consistency among different rubrics for assessing writing, Lang. Test. Asia 10 (1) (2020) 12.
[28] G.T.L. Brown, H.L. Andrade, F. Chen, Accuracy in student self-assessment: directions and cautions for research, Assess. Educ.:. Princ. Policy. Pract. 22 (4) (2014) 444–457, https://doi.org/10.1080/0969594X.2014.996523.
[29] A.R. Rezaei, M. Lovorn, Reliability and validity of rubrics for assessment through writing, Assess. Writ. 15 (1) (2010) 18–39.
[30] T.F. McNamara, Language Testing, Oxford University Press, 2000.
[31] N. Gisev, J.S. Bell, T.F. Chen, Interrater agreement and interrater reliability: key concepts, approaches, and applications, Res. Soc. Adm. Pharm. 9 (3) (2013) 330–338.
[32] J.Y. Lee, Doctoral dissertation, Pennsylvania State University, 2022.
[33] J.L. Fleiss, Statistical Methods For Rates Andproportions, 2ndEd, John Wiley, New York, 1981, pp. 38–46.
[34] E. Almanasreh, R. Moles, T.F. Chen, Evaluation of methods used for estimating content validity, Res. Soc. Adm. Pharm. 15 (2) (2019) 214–221.
[35] S.J. Osterlind, What is Constructing Test items? Springer, Netherlands, 1998, pp. 1–16.
[36] M.R. Lynn, Determination and quantification of content validity, Nurs. Res 35 (6) (1986) 382–386.
[37] D.F. Polit, C.T. Beck, The content validity index: are you sure you know what's being reported? Critique and recommendations, Res. Nurs. Health 29 (5) (2006) 489–497.
[38] C. Welch, Item and prompt development in performance testing, Handb. Test. Dev. (2006) 303–327.
[39] S.S. Moya, J.M. O'malley, A portfolio assessment model for ESL, J. Educ. Issues. Lang. Minor. Stud. 13 (1) (1994) 13–36.
[40] R. Lam, Promoting self-regulated learning through portfolio assessment: testimony and recommendations, Assess. Eval. High. Educ. 39 (6) (2014) 699–714.
[41] J.S. Delett, S. Barnhardt, J.A. Kevorkian, A framework for portfolio assessment in the foreign language classroom, Foreign. Lang. Ann. 34 (6) (2001) 559–568.
[42] T. Eckes, Operational rater types in writing assessment: linking rater cognition to rater behavior, Lang. Assess. Q. 9 (3) (2012) 270–292.
[43] V. Klenowski, S. Askew, E. Carnell, Portfolios for learning, assessment and professional development in higher education, Assess. Eval. High. Educ. 31 (3) (2006) 267–286.
[44] Z.A. Philippakos, C.A. MacArthur, The use of genre-specific evaluation criteria for revision, Lang. Lit. Spectr. 26 (2016) 41–52.
[45] S.C. Weigle, Assessing Writing, Cambridge University Press, 2002.