Scientific Research Publishing

Open Journal of Applied Sciences, 2026, 16(8), 2842-2871
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Citation Competence in High- and Low-Rated Master of Education Theses and an Instructional Model

Xiuling Shi[1], Lifang Wei[2]

[1]School of English Studies, Zhejiang Yuexiu University, Shaoxing, China
[2]School of Foreign Languages, Shaoxing University, Shaoxing, China
Email: 764499355@qq.com

How to cite this paper: Shi, X.L. and Wei, L.F. (2026) Citation Competence in High- and Low-Rated Master of Education Theses and an Instructional Model. *Open Journal of Applied Sciences*, **16**, 2842-2871.
https://doi.org/10.4236/ojapps.2026.168159

Received: August 3, 2026
Accepted: August 24, 2026
Published: August 27, 2026

Abstract

Citation competence is a core component of English academic writing literacy. Its essence lies not merely in conforming to referencing conventions, but in drawing on others' research to construct arguments, express stance, and participate in scholarly dialogue. Using Petrić's (2007) framework of the rhetorical functions of citations as its analytical lens [1], this study compares the citation practices of 10 high-rated and 10 low-rated Master of Education (M.Ed.) theses in subject teaching (English) from a Chinese university, focusing on their Introduction and Literature Review modules and triangulating the results with the comment themes in 42 thesis review forms, to test whether Petrić's findings replicate among Chinese M.Ed. students. The results show a tendency pattern consistent in direction with Petrić's findings but moderate in magnitude: the gap lies not so much in the quantity of citation—the two groups' citation densities do not differ significantly—as in its quality. In thesis-level tests, statement of use is the only between-group difference to reach significance (aggregate share high 5.3% vs low 2.5%, exact $p = 0.011$; $q = 0.023$ within the two pre-designated confirmatory indicators, but $q = 0.090$ under full-family correction), while evaluation, exemplification, and most other higher-order functions show same-direction tendencies (links and the very sparse application category run in the opposite direction; the former, coding-sensitive, is not used as a basis for argument); the comment themes converge with the quantitative pattern, though they derive from the same review process that defined the groups. On this basis, the study interprets citation competence as a continuum extending from knowledge telling to knowledge transformation, targets pedagogical intervention at evaluation and statement of use, and constructs a three-layer "diagnosis-progression-feedback" model for citation competence instruction, thereby providing an empirical basis for culti-

vating L2 citation competence and a framework proposal open to classroom testing.

Keywords

Citation Competence, Rhetorical Functions of Citations, Knowledge Transformation, Master of Education (M.Ed.) Theses, L2 Academic Writing

1. Introduction

Citation is one of the fundamental features that distinguish academic writing from other genres. By citing prior literature, writers embed their own research within an existing network of knowledge, signaling their affiliation with the academic community while establishing the legitimacy and originality of their own claims. As Hyland (2000) points out, citation is not a value-neutral transfer of information but a core mechanism in the social construction of knowledge; what writers say certainly matters, but how they invoke others to support what they say bears equally on the shaping of their academic identity. Citation competence therefore goes far beyond correct referencing: it is a rhetorical competence—the integrated ability to select appropriate citation functions, to manage the relationship between source material and one's own views, and to establish links between sources effectively while expressing an evaluative stance [2] [3].

For Chinese writers of English as a second language, citation is precisely where competence gaps are especially readily exposed. Previous research has repeatedly shown that L2 writers tend to favor paraphrase and the stacking of information in their citations while remaining conspicuously weak in higher-order functions such as evaluation, comparison, and establishing links between sources [4] [5]. This tendency to privilege telling over transformation makes citation competence a highly sensitive indicator of the developmental level of L2 academic writing.

This study focuses on M.Ed. theses in subject teaching (English) and adopts a design that compares a high-rated group with a low-rated group for three reasons. First, M.Ed. students constitute a large yet long-overlooked population of L2 writers in academic writing research; their theses are mostly empirical studies of English teaching in primary and secondary schools, and their theses afford a direct window on the citation practices of this population. Second, thesis review forms (hereinafter "review forms") provide natural high/low anchors and a basis of expert judgment: reviewers' scores and comments on dimensions such as "topic selection and literature review" and "compliance with thesis conventions" (including the conventions of figures, tables, and referencing) offer an external criterion, independent of the researcher, for stratifying the theses. Third, a comparative design goes beyond the frequency description of a single group and directly reveals developmental differences in citation competence, thereby addressing a key question: can the pattern that Petrić (2007) identified in Central European English-

medium master's theses—the high-rated group using more non-attribution functions while the low-rated group relied predominantly on attribution—be replicated in the context of Chinese M.Ed. theses [1]?

Accordingly, this study poses three research questions: 1) How do high-rated and low-rated theses differ in the distribution of rhetorical functions of citations in the introduction and literature review sections? 2) How do citation-related evaluative themes in the reviewers' comments (e.g., the summary of the review, the research gap, the theoretical basis, citation conventions, and literature support in the discussion) diverge between the two groups? 3) How do these quantitative and qualitative differences map onto the competence continuum from knowledge telling to knowledge transformation? Answering these questions not only tests and extends the explanatory power of Petrić's framework in the new context of Chinese M.Ed. students, but also enables EAP citation instruction to locate the competence bottleneck; the study is thus of both theoretical and practical significance.

2. Literature Review

2.1. The Evolution of Citation Research from Scientometrics to Rhetorical Functions

Systematic research on citation practices originated in scientometrics and the sociology of science. In their analysis of physics papers, Moravcsik and Murugesan (1975) were among the first to classify citations according to their nature and function (e.g., conceptual versus operational, organic versus perfunctory, confirmative versus negational), revealing that citations are not value-neutral markers of information but carry the author's evaluative stance [6]. This functional orientation laid the foundation for subsequent EAP research.

As citation studies entered linguistics and genre research, Swales (1990), within his framework of genre analysis, proposed the influential distinction between integral and non-integral citations: in the former, the cited author's name functions as a grammatical constituent of the sentence (e.g., as its subject), foregrounding the author; in the latter, the reference is placed outside the syntactic structure, typically in parentheses, foregrounding the propositional content. This syntactic dimension has become a basic tool for research on citation form and, together with the dimension of rhetorical function, constitutes the surface and deep facets of citation competence. Thompson and Ye (1991) demonstrated that reporting verbs themselves carry the citing writer's evaluative stance toward the reported material, opening up the evaluative dimension of citation language [7]. Hyland (1999, 2000) further situated citation within the construction of disciplinary discourse, using cross-disciplinary corpora to reveal marked disciplinary differences in citation density, reporting verbs, and integral/non-integral patterns, and emphasizing that citation is a social-interactional resource for negotiating the legitimacy of knowledge [2] [8]. Subsequently, Thompson and Tribble (2001) used corpus methods to develop Swales's binary framework into an operational tool for EAP teaching [9], and Thompson (2005) examined the distribution of the "focus"

and "stance" of intertextual reference in doctoral theses [10]. Tracking journal corpora in four disciplines from 1965 to 2015, Hyland and Jiang (2019) further document a sustained rise in citation density, underscoring the growing weight of citation in knowledge construction [11].

Meanwhile, another line of research has addressed textual borrowing and citation ethics among L2 writers. Pennycook (1996) offered a cultural reflection on the concept of "plagiarism," while Pecorari (2003, 2006) showed through empirical analysis that the pervasive "patchwriting" in L2 graduate theses is often not deliberate plagiarism but a symptom of insufficient citation competence, rooted in a double deficit of linguistic resources and rhetorical awareness [12]-[14].

2.2. The Petrić (2007) Framework of Nine Rhetorical Functions of Citations

Within this line of inquiry, Petrić's (2007) study offers the greatest methodological value for the present research. Focusing on L2 writers, she analyzed 16 master's theses in gender studies from an English-medium university in Central Europe (8 high-rated theses graded A and 8 low-rated theses graded mostly B) and systematically identified nine rhetorical functions of citations: attribution, exemplification, further reference, statement of use, application, evaluation, establishing links between sources, comparison of one's own findings with other sources, and other. Petrić's central finding is that attribution was overwhelmingly dominant in both groups, indicating that novice academic writers' citation remains primarily a matter of knowledge display. However, the high-rated group showed markedly higher proportions of higher-order functions such as evaluation, statement of use, application, and establishing links between sources, as well as a higher share of citations carrying multiple functions; these differences were especially pronounced in the introduction and literature review chapters (the specific figures are juxtaposed with the present study's results in Chapter 5). Drawing on Bereiter and Scardamalia's (1987) distinction between knowledge telling and knowledge transformation, Petrić accordingly argued that the gap between high- and low-rated theses is, in essence, the gap between passively listing sources and actively constructing arguments through citation [15]. She also identified two qualitative problems in the low-rated group: expressive deficiencies caused by limited language proficiency, and inappropriate use of rhetorical functions (e.g., praising an author's status as a "prominent scholar" [p. 250] rather than evaluating the research itself).

This functional perspective has been repeatedly corroborated in recent research. Comparing non-native expert and novice scientific writers, Mansourizadeh and Ahmad (2011) found that novices likewise favored attribution and rarely used citations to support their own argumentation [16]; Lin and Lau's (2021) examination of evaluative language in master's thesis defenses and Sun and Jiang's (2025) analysis of how doctoral writers deploy appraisal resources to acknowledge research limitations both indicate that evaluation, as a higher-order function, is a key marker of stratification in academic writing competence [17] [18]. The frame-

work has likewise been carried toward the writer's own perspective: Petrić and Harwood (2013) elicited a successful L2 writer's self-reported citation functions, linking the textual functions of citations to task requirements and task representation [19].

In terms of research design, the present study can be viewed as a conceptual replication of Petrić (2007) in the Chinese M.Ed. context: the corpus likewise consists of master's theses by L2 writers, the grouping is likewise based on review ratings, and the category set is kept identical, which makes the between-group patterns comparable across contexts. At the same time, operationalizing the framework's categories is not without difficulty—as the reliability analysis in Section 3 and the discussion in Section 5 will show, the links category is highly sensitive to merging criteria, which itself constitutes first-hand evidence on the framework's conditions of application.

2.3. Positioning and Gaps of Citation Research on L2 Academic Writing in China

Over the past decade or so, research on citation in China has gradually gained momentum, proceeding mainly through corpus-based frequency description. Xu (2012) provided one of the earliest systematic characterizations of citation features in empirical English academic discourse [20]; Xu (2016) subsequently drew on both cross-sectional and longitudinal evidence to propose a three-dimensional structure of citation competence in L2 academic writing—breadth of content coverage, interpersonal interaction, and discourse integration—offering a localized competence framework for the field [5]. Wang *et al.* (2017) examined the citation features of English graduate students' academic writing [4]; Mu (2019) compared the forms and rhetorical functions of citations in English and Chinese research articles from the perspective of intertextuality [21]; and H. Xu (2011) focused on authorial stance markers in the academic discourse of advanced Chinese learners of English [22]. More recently, Wei *et al.* (2023) attempted to assess the originality of research articles on the basis of citation intent, extending citation research into the domain of research evaluation [23]. These studies echo international competence research framed by Cumming *et al.*'s (2018) work on writing from sources, jointly sketching a multidimensional picture of citation competence [24].

A survey of the existing literature, however, reveals two gaps. First, research in China has mostly consisted of corpus-based frequency descriptions of a single group, seldom employing high-/low-rated comparisons to reveal the developmental gradient of citation competence. Second, research subjects have been concentrated on academic master's and doctoral theses and journal articles [25]—international examination of citation practices in master's theses has indeed continued [26] [27]—with little attention paid to M.Ed. students, a large and distinctive population of L2 writers. The present study is situated precisely at the intersection of these two gaps: using Petrić's (2007) functional framework as its lens and incorporating the expert judgment of the reviewers' comments, it conducts a stratified

comparison of citation competence in Chinese M.Ed. theses, seeking to supply citation competence research in China with the two missing dimensions of the developmental gradient and the M.Ed. population.

3. Research Design

3.1. Research Questions

This study centers on one core question: whether, in the Introduction and Literature Review modules of Chinese Master of Education (M.Ed.) theses in subject teaching (English), the theses receiving higher and lower review grades differ systematically in citation competence, where such differences concentrate, and how they should be explained. The specific research questions correspond to the three questions raised in the Introduction: 1) What differences exist between the two groups of theses in the distribution of the rhetorical functions of citation in the Introduction and Literature Review modules? At the operational level this is refined into two observation points: first, citation density and the formal distribution of integral versus non-integral citations; second, the distribution of attribution versus non-attribution functions (in particular evaluation, statement of use, and links) as measured by Petrić's (2007) nine-category functional framework. 2) How do the citation-related evaluative themes in the thesis review comments (summary of the review, research gap, theoretical basis, citation conventions, literature support in the discussion, etc.) diverge between the two groups, and can they be triangulated with the citation-function counts? 3) How do the above quantitative and qualitative differences map onto the competence gradient from knowledge telling to knowledge transformation? Questions (1) and (2) are answered directly by the results in Section 4, while Question (3) is addressed in the discussion in Section 5 by synthesizing the three lines of evidence—the quantitative results, the reviewers' comments, and the authentic examples.

3.2. Corpus and Grouping

The corpus comprises two components of 20 M.Ed. (subject teaching, English) theses: the Introduction (Chapter 1) and the Literature Review module (from Chapter 2 up to the methodology chapter, including the theoretical-basis chapter). All theses are empirical studies of English teaching in primary and secondary schools, written in English by the learners; the clean full texts run to 91 - 145 pages and 24,626 - 47,034 words. After segmentation and removal of headers and page numbers, the Introductions of the high-rated group total 15,396 words and their Reviews 69,741 words, against 16,256 and 63,475 words respectively for the low-rated group—approximately 165,000 words across the two modules (85,137 + 79,731 = 164,868 words).

Grouping was based on the 100-point scores and conclusion grades recorded in the professional-degree thesis review forms of the university from which the sample was drawn; the high-rated group is numbered H01-H10 and the low-rated group L01-L10, with 10 theses in each group. All these were submitted for external

review at this university in 2025; two low-rated theses had three reviewers each, yielding 22 review forms for that group, while each high-rated thesis had two reviewers, yielding 20. The high- and low-rated groups were demarcated by the research team by partitioning the theses according to the total scores and conclusion grades they had received on review: every thesis in the high-rated group carries a grade combination of AA or AB (three AA and seven AB), whereas the low-rated group is dominated by B grades (its single A comes from a three-reviewer combination that was reckoned as a BB-level combination for grouping purposes, and the group also contains two C forms). Because the partition drew on both scores and grade composition rather than a single score threshold, the two groups' student-level score ranges partly overlap (see below); the present study carried out its coding and analysis on this pre-established grouping. The archive of review forms available to the project covered 22 students (11 per group as filed); one student per group was not retained, and the coded corpus comprises the 10 theses per group for which both the complete review forms and the clean English full texts used for coding are on file. The student not retained on the high side carried an A + B combination (mean total 83.0), within the same AA/AB rule as the retained theses, and the one on the low side a C + B + C combination (mean total 69.3), the lowest-rated in the archive—an exclusion that attenuates rather than inflates the between-group contrast; both exclusions were fixed before citation coding began. The quantitative basis for the grouping is shown in Table 1 (means for each dimension are computed from the 42 review forms):

Table 1. Mean review scores of the high- and low-rated groups by scoring dimension.

Scoring dimension (full marks)	High-rated group mean	Low-rated group mean
Topic selection and literature review (20)	17.05 (85.3%)	15.95 (79.8%)
Originality and value of the thesis (35)	29.45 (84.1%)	27.14 (77.5%)
Research ability and foundational knowledge (30)	25.00 (83.3%)	23.45 (78.2%)
Compliance with thesis conventions (15)	12.50 (83.3%)	11.50 (76.7%)
Total (100)	84.00	78.05

The means for each dimension and the mean total score were computed separately from the raw scores and then rounded to two decimal places; the direct sum of the dimension means may therefore deviate from the total by ±0.01 owing to rounding.

The distribution of conclusion grades (counted by review form) is 13A/7B for the high-rated group and 1A/19B/2C for the low-rated group. In terms of student-level mean total scores, the two groups' ranges are 78 - 88.5 and 74 - 82.5 respectively—partially overlapping, which further indicates that the grouping is "relatively high/low" rather than an extreme contrast. To address this overlap directly, a sensitivity analysis regrouped the 20 theses purely by student-level mean total score (top 10 vs bottom 10). This alternative partition moves two theses across groups and leaves a single tie at the boundary (two students at 81.5) that can be

resolved in either direction; under both resolutions statement of use remains the only indicator to reach or approach thesis-level significance (exact $p = 0.002$ and $p = 0.054$ respectively) and no other indicator becomes significant, while the directions of statement of use, evaluation, and the integral/attribution pattern are unchanged—only the two sparsest tendency indicators (multi-function and exemplification) flip direction under one resolution, consistent with their tendency-only status in this paper. The two groups show a stable gap on the two dimensions most directly related to citation—"topic selection and literature review" and "compliance with thesis conventions"—which constitutes a valid basis for the high-low contrast. This grouping ensures that the distinction between the high- and low-rated groups is not the researcher's subjective judgment but a quantitative conclusion derived from independent peer review.

3.3. Analytical Framework and Coding Protocol

Citation functions were coded with the nine-category rhetorical-function framework established in Petrić's (2007) comparative study of high- and low-rated master's theses: attribution, exemplification, further reference, statement of use, application, evaluation, establishing links between sources (hereafter links), comparison of one's own findings with other sources (hereafter comparison_own), and other. Given that the author's own findings have not yet been presented in the Introduction and Literature Review modules, comparison_own was operationalized in this study as citations that explicitly connect others' findings with the present study. The coding unit is each citation instance in the text; an instance simultaneously serving two or more functions was also marked as a multi-function citation. The formal dimension (integral vs. non-integral) was recorded in parallel; cases such as multiple sources listed within a single set of parentheses and repeated citations across sentences were merged and counted according to pre-established rules, and the original English sentences were preserved thesis by thesis as illustrative examples and for verification. In the text and tables, comparison_own is a coding-variable abbreviation corresponding to Petrić's original term comparison of one's own findings with other sources; all other function names follow the original terms.

3.4. Triangulation with 42 Thesis Review Forms

To avoid thin interpretation resting solely on citation counts, this study incorporated the qualitative comments in 42 thesis review forms (20 for the high-rated group and 22 for the low-rated group) for triangulation. Each review form contains scores on four dimensions together with a text of "general comments plus revision suggestions"; the score for "topic selection and literature review" and comment prompts such as "literature review ability" and "conventions in the use of figures, tables and citations" point to the same construct as the citation-function coding. The reviewers' comments and the citation statistics from the full theses form a complementary chain of evidence: the former are judgments by peers external to the researchers on the quality of citation and the review, and the latter a quantifiable

record of citation behavior. The comments were coded by theme (summary of the review, research gap, theoretical basis, citation conventions, literature support in the discussion, etc.) and cross-checked against the citation-function counts, so that the quantitative conclusions gain corroboration from expert judgment. It should be noted that the comments share the same source as the scores on which the grouping was based, so their overall association with group membership is in part an endogenous result of how the groups were defined; the force of the triangulation lies in the fact that the comment themes point specifically to the citation/review dimensions and converge, on specific dimensions, with the citation-function counts completed separately and independently. Comment themes were classified by low-inference matching of explicit wording (for instance, "no discernible research gap" required an explicitly corresponding statement in the comments); this classification was completed by the research team without independent double coding, so the theme frequencies should likewise be read directionally.

3.5. Coding Reliability

All citation-function coding was completed by the research team with the aid of text-analysis tools, and the coding results are to be interpreted under the constraint of the reliability check reported below. An independent second coder recoded the Literature Review modules of H07 and L05, and the results were compared with the primary coding (the category-level counts of the reliability check are available from the author upon request). The total numbers of citations identified by the two coders are highly close (H07: 71 vs 70, a total-count agreement of 98.6%; L05: 125 vs 125, 100%—a ratio of totals, not an item-by-item matching rate); the distributions on the formal dimension are highly consistent (H07: integral/non-integral 52/19 vs 51/19, the one-instance difference arising in citation identification; L05: 80/45 vs 80/45, identical). Among the functional categories, the differences for evaluation (16 vs 18), statement of use (6 vs 4), exemplification, and application are all within 2 instances; the disagreement concentrates between links and attribution (L05: links 26 vs 14, attribution 84 vs 97)—items the primary coder counted as links under the "consensus statement plus multi-source list" rule were mostly coded as attribution by the second coder, so the absolute proportions of the two categories fluctuate jointly with the merging standard; in addition, the marking of multi-function citations shows relatively larger divergence (H07: 10 vs 6), for which reason the between-group difference in multi-function share is interpreted as a tendency only. Accordingly, the main conclusions of this study rest on the attribution-dominant pattern, on which the two coders agreed in direction, and on the consistently coded evaluation and statement of use categories, which carry explicit linguistic markers; the absolute values of attribution and links are interpreted with caution, and links is not used as a basis for argument. It should also be noted that the re-coding of two theses cannot be equated with the reliability of full-scale, independent blind coding by multiple coders, and the absolute values of the coding should accordingly be treated with caution. Two further con-

straints are stated explicitly: because the citation-by-citation pairing of the two coders' decisions was not archived, item-level agreement coefficients for each functional category (e.g., per-category κ) cannot be computed retrospectively—the figures above are category-level count comparisons, not item-level agreement rates; and no formal procedure blinded the second coder to the theses' group membership. Both constraints reinforce the decision to rest no conclusion on the absolute values of any single category. The two coders concurred on the directional between-group conclusions (attribution dominance and impoverished evaluative functions in the low-rated theses).

3.6. Statistical Methods

Between-group comparisons take the thesis as the unit of analysis (10 per group), using exact Mann-Whitney U tests (full enumeration of the permutation distribution, two-sided) with the rank-biserial correlation r reported as the effect size. On the theoretical grounds of Petrić's (2007) findings and the "interpersonal interaction" dimension of Xu (2016), evaluation and statement of use were designated as confirmatory indicators and the remaining indicators as exploratory, with Benjamini-Hochberg false-discovery-rate correction for multiple comparisons. Item-level two-proportion tests and Cohen's h (high 1,437 vs low 1,452 instances) serve as descriptive reference only—citation instances cluster within theses, and treating them as independent observations would underestimate standard errors; Fisher's exact test was used for sparse counts. All tests were computed with the project's own program on the raw coded data (program and coded data are available from the author upon request). The designation of the confirmatory indicators rests on the literature-based priors above rather than on formal preregistration before data collection; the corresponding results under full-family correction are reported alongside the thesis-level tests in Section 4 for readers to weigh.

4. Results

4.1. Between-Group Quantitative Comparison of Citation Functions

Table 2 presents the between-group comparison with the Introduction and Literature Review modules combined (functional percentages are each function's share of the group's total citation instances; owing to multi-function citations, the functional percentages do not sum to 100%):

Table 2. Citation behavior of the high-rated and low-rated groups (Introduction + Literature Review combined).

Indicator	High-rated group (H01 - H10)	Low-rated group (L01 - L10)
Module word count	85,137	79,731
Citation instances	1437	1452

Continued

Citation density (per 1000 words)	16.88	18.21
Integral citations (%)	58.5%	70.9%
attribution	890 (61.9%)	954 (65.7%)
Non-attribution	547 (38.1%)	498 (34.3%)
evaluation	114 (7.9%)	84 (5.8%)
statement of use	76 (5.3%)	36 (2.5%)
links	274 (19.1%)	312 (21.5%)
exemplification	119 (8.3%)	77 (5.3%)
application	17 (1.2%)	20 (1.4%)
comparison_own	2	0
Multi-function citations (%)	3.8%	2.1%

The counts of comparison_own are too small (2 and 0 instances) for percentages to be reported.

The between-group differences form a pattern that runs counter to the intuition that "higher-rated means more citing", yet corresponds closely to the distinction between knowledge telling and knowledge transformation invoked by Petrić (2007). First, the low-rated group shows a higher citation density (low 18.21 vs high 16.88 per 1,000 words) and a higher proportion of integral citations (low 70.9% vs high 58.5%), indicating heavier reliance on list-like, item-by-item reporting in the "author (year) + reporting verb" pattern. Second, the low-rated group has a higher share of attribution (low 65.7% vs high 61.9%; the same coding standard was applied to both groups, so the direction is comparable, though the absolute difference is affected by the merging convention for links and should be interpreted with caution), that is, more of its citations remain at the knowledge-telling level of "who said what", whereas the non-attribution functions are relatively more active in the high-rated group (high 38.1% vs low 34.3%)—corresponding precisely to that distinction between knowledge telling and knowledge transformation. Third, the two functions that most directly embody authorial stance—evaluation (high 7.9% vs low 5.8%) and statement of use (high 5.3% vs low 2.5%)—are both higher in the high-rated group, the latter by more than double and the only indicator to reach thesis-level significance (see Table 3); exemplification (high 8.3% vs low 5.3%) is likewise more frequent in the high-rated group, and comparison_own appears only in the high-rated group (2 vs 0 instances, Fisher exact $p = 0.247$, illustrative rather than statistical evidence). Fourth, the high-rated group's share of multi-function citations is about 1.8 times that of the low-rated group (high 3.8% vs low 2.1%; item-level $\chi^2 = 7.18$, $p = 0.007$ but thesis-level $p = 0.486$, hence read as a tendency), indicating that its individual citations carry more complex rhetorical tasks.

Table 3 reports the statistical tests of the between-group differences. Exact Mann-Whitney tests with the thesis as the unit of analysis show that statement of use is the only difference to reach thesis-level significance (p = 0.011, rank-biserial r = 0.66, a large effect; still significant after BH correction within the two pre-designated confirmatory indicators, q = 0.023, though attenuated to q = 0.090 under full-family correction across all eight indicators—see the note to Table 3), and that its advantage is concentrated in the Introduction module (9/143 vs 1/121, the nine instances coming from six different theses; Fisher exact p = 0.023, item-level basis). The remaining indicators show same-direction tendencies that do not reach thesis-level significance; the two groups' citation densities do not differ significantly (p = 0.579)—the high-rated group did not simply "cite more", and this very absence of a difference supports the judgment that the gap does not lie in quantity; links is non-significant even at the item level (χ^2 = 2.62, p = 0.106), consonant with the cautious stance taken in the reliability analysis.

Table 3. Statistical tests of between-group differences (thesis-level tests primary; item-level effect sizes supplementary).

Indicator	Thesis-level medians (high/low)	U	Exact p	r	Item-level Cohen's h
statement of use %	5.17/1.75	17	0.011	0.66	+0.148
evaluation %	6.48/4.43	39	0.436	0.22	+0.085
Multi-function %	2.10/1.31	40.5	0.486	0.19	+0.101
Integral %	62.30/84.51	30	0.143	0.40	−0.259
attribution %	66.84/69.49	45	0.739	0.10	−0.078
exemplification %	4.70/4.51	42	0.579	0.16	+0.119
links %	14.08/17.87	47	0.838	0.06	−0.060
Citation density (per 1000 words)	17.29/17.60	42	0.579	0.16	-

Thesis-level metrics are each thesis's count of the function as a percentage of that thesis's total citation instances (density in citations per 1,000 words), with medians computed by group; exact p values come from the full enumeration of the C(20,10) = 184,756 possible group assignments (two-sided); item-level Cohen's h is based on the pooled group proportions (1,437 vs 1,452 instances) and, because citations cluster within theses, serves as an effect-size reference only. The confirmatory indicators are evaluation and statement of use (corrected with m = 2); under full-family BH correction (m = 8), the q value for statement of use is .090. r is the rank-biserial correlation (absolute value); the direction of each difference follows the medians column. The sparse application category (high 1.2% vs low 1.4%) and the near-zero comparison_own are not entered into the thesis-level tests, being reported only in Table 2 and via Fisher's exact test.

The aggregate percentages in Table 2 are weighted by citation instances, giving longer theses more weight (H06 alone contributes 20.8% of the high-rated group's instances); on the unweighted "one thesis, one vote" basis the direction of every indicator is unchanged (evaluation 7.94% vs 6.30%, statement of use 5.49% vs 2.43%, integral 63.0% vs 71.6%, density 16.99 vs 18.54 per 1,000 words, the high-rated group first in each pair), and a sensitivity check recomputed with H06 excluded likewise alters the direction of none of the indicators.

Viewed by module, the differences are unevenly distributed between the Introduction and the Literature Review (Table 4):

Table 4. Between-group comparison of citation functions by module (Introduction vs Literature Review).

Indicator	H-Introduction	L-Introduction	H-Review	L-Review
Word count	15,396	16,256	69,741	63,475
Citation instances	143	121	1294	1331
Density (per 1,000 words)	9.29	7.44	18.55	20.97
Integral %	37.8%	65.3%	60.8%	71.4%
attribution %	71.3%	71.1%	60.9%	65.2%
evaluation %	9.8%	9.9%	7.7%	5.4%
statement of use %	6.3%	0.8%	5.2%	2.6%
links %	7.7%	8.3%	20.3%	22.7%
exemplification %	2.1%	5.0%	9.0%	5.3%
application %	2.8%	5.0%	1.0%	1.1%
Multi-function %	0%	0%	4.3%	2.3%

The most striking feature in the Introduction is statement of use: 6.3% in the high-rated group against a mere 0.8% in the low-rated group, showing that high-rated writers explicitly declare at the outset the theories or frameworks they intend to adopt, whereas some low-rated writers substitute the authority of policy documents for research positioning (e.g., 6 of the 9 citations in L08's Introduction are to the English curriculum standards documents). Notably, citation density in the Introduction is in fact higher in the high-rated group (9.29 per 1,000 words) than in the low-rated group (7.44); the corpus-wide density reversal derives mainly from the Literature Review chapter (high 18.55 vs low 20.97), suggesting that the low-rated group's "citing more" is concentrated in list-like paraphrase within the review. The formal contrast is equally conspicuous in the Introduction: the integral proportion is only 37.8% in the high-rated group but reaches 65.3% in the low-rated group, a gap of 27.5 percentage points—the largest between-group difference among all percentage indicators; its item-level effect size for the two modules combined is also the largest of all indicators ($h = -.26$), yet the thesis-level test does not reach significance ($p = 0.143$, Table 3) and the reliability check covered only the review modules, so it is best read as a strong tendency rather than a settled verdict—high-rated writers more often foreground propositional content with non-integral parenthetical citations at the opening, whereas low-rated writers habitually begin sentence after sentence with the "author (year)" pattern. The Literature Review chapter, in turn, is the principal arena for the gaps in evaluation

(7.7% vs 5.4%) and statement of use (5.2% vs 2.6%), whereas links (20.3% vs 22.7%), in view of the reliability finding that links counts are unstable, is not treated as major evidence. It should also be noted that in the Introduction module exemplification and application run opposite to the overall direction (higher in the low-rated group), but the counts are sparse (high 3 and 4 vs low 6 and 6 instances) and cannot support interpretation; further reference and other are 0 across the corpus, which is why the tables list only the seven categories that occur.

A thesis-by-thesis robustness check lends further support to these conclusions: in the high-rated group, the mean per-thesis attribution share is 63.5% (SD 12.7, range 38.9 - 80.6) and the mean number of evaluation instances per thesis is 11.4 (range 4 - 33); in the low-rated group, the mean per-thesis attribution share is 65.4% (SD 15.7, range 36.9 - 87.4) and the mean evaluation count is only 8.4 (range 3 - 22). Although the between-group gap in mean attribution share is only about 2 percentage points, the within-group dispersion should not be overlooked—the standard deviation is larger in the low-rated group (15.7 vs 12.7), indicating extreme cases within it (in one thesis the attribution share reaches 87.4%, amounting almost to pure stacking of paraphrase), while the between-group gap in evaluative functions is on the whole preserved at the per-thesis level (though the two groups' per-thesis ranges also overlap). Both groups score 10/10 for the presence of an explicit research gap or a summary of the review, indicating that the difference lies not in presence versus absence but in the quality and density of evaluation and integration: quite a few low-rated theses possess the formal skeleton of a summary of the review and a research gap yet lack critical weighing of the cited literature, so that their "reviews" function more as compilations of sources than as dialogue with them. In its functional distribution this pattern points in the same direction as Petrić's (2007) findings on L2 master's theses in a European context: high-rated writers do not cite more; rather, they are better at making each citation serve stance expression and research positioning, thereby elevating citation from a vehicle of knowledge telling into a means of knowledge transformation.

4.2. Between-Group Distribution of Comment Themes

When the citation-related comments in the 42 review forms are classified by theme, their distribution across the two groups accords with the quantitative results on the whole (except that for citation conventions the two groups' frequencies are close while differing in nature—see below), displaying a clear "gradient" difference (Table 5):

Table 5. Between-group distribution of citation/review-related comment themes (by review form).

Comment theme	High-rated group (/20 forms)	Low-rated group (/22 forms)
Praise for a "comprehensive/systematic review with apt commentary"	17/20	10/22
"Review lacks a summary/no discernible research gap"	1/20	5/22

Continued

"Theoretical basis weak/missing/detached from the study"	2/20	7/22
"Reference labelling/citation format not standard"	6/20	7/22
"Discussion not supported by literature/no comparison with prior studies"	1/20	5/22
"Literature outdated/work of the last three years missing"	1/20	4/22

A clear gradient is visible: the low-rated group is repeatedly criticized for deficiencies in the "basic elements"—only about half of its review forms (10/22) receive positive comments of the "comprehensive/systematic review" type (17/20 for the high-rated group), while "review lacks a summary, no discernible research gap" (low 5/22 vs high 1/20), "theoretical basis weak or detached from the study" (low 7/22 vs high 2/20), and "discussion not supported by literature, no comparison with prior studies" (low 5/22 vs high 1/20) cluster in the low-rated group. Such comments do not contradict the earlier finding that both groups score 10/10 for an explicit summary or research gap: textual verification shows that in the criticized theses the summaries and research gaps are largely nominal—present in form yet devoid of critical weighing of the cited literature; that the reviewers "cannot discern" them demonstrates precisely that an element failing to perform its rhetorical function appears, in the reader's eyes, as absent. The criticisms directed at the high-rated group operate at a higher level, mostly finer-grained demands such as "insufficient critical analysis", "closer articulation with the local context", and "effect sizes of prior studies should be reported", pointing to the deepening of existing knowledge transformation rather than to the presence or absence of elements. Citation-convention problems occur at similar rates in the two groups (low 7/22 vs high 6/20) but differ in nature: in the high-rated group they mostly concern details such as volume/issue numbers and page numbers, whereas in the low-rated group they include substantive defects such as mismatches between in-text citations and the reference list, unattributed figures and tables, and chronological errors in citation. The comment themes and the functional distribution in Table 2 converge—the quantitative shortfall in the evaluation function manifests itself precisely as the experts' qualitative verdicts of "summarizing without evaluating" and "no discernible gap". This convergence carries two qualifications stated in Section 3.4: the comment themes were classified by the research team through low-inference matching of explicit wording, without independent double coding, and the 42 forms from which they derive are the same documents whose scores and grades defined the two groups; the triangulation is therefore convergent evidence from the same assessment process rather than corroboration from a fully independent source.

4.3. Corroboration from Authentic Examples

The differences above are directly attested in the original citation sentences. (Ci-

tations appearing inside or alongside the quoted student sentences below are part of the corpus data under analysis, not sources cited by this paper, and are accordingly not listed in the References.) Evaluative citations in the high-rated group typically deliver substantive criticism of specific studies rather than generic endorsement. For instance, in reviewing textbook research, the Literature Review chapter of H04 writes:

"Sun (2022) evaluated the reading sections' critical thinking cultivation of the New Standard English compulsory textbooks... However, this research process neglected the selection of selective compulsory and optional textbooks."

Reporting first and then pointing out the limitation is a typical realization of the evaluation function. The high-rated group also uniquely exhibits comparison_own, bringing others' findings into dialogue with the present study, as in the Literature Review chapter of H04:

"Zhou's (2018) findings complement the present study by emphasizing the key factors in shaping the cultivation of critical thinking."

This sentence corresponds exactly to the quantitative conclusion that comparison_own appears only in the high-rated group (2 vs 0)—high-rated writers do not merely report; they draw prior findings into a dialogic frame with their own research.

By contrast, "evaluation" in the low-rated group is mostly polite commendation that never engages with the merits and flaws of the work. Consider the treatment of Guilloteaux and Dörnyei (2008) in the Literature Review chapter of L06:

"Guilloteaux and Dörnyei (2008) made a notable contribution to this by categorizing motivational strategies into two distinct types: teaching motivation strategies and learning motivation strategies."

The tone is affirmative, yet it stops at restating their classification without judging its value or limitations. Citations of the "general statement plus multi-source parenthetical list" type are likewise common in the low-rated group, functionally leaning towards the stacking of attribution, as in the Literature Review chapter of L08:

"The ultimate goal of learning activities is to cultivate English subject core competencies, seeking to achieve students' comprehensive development of language ability, cultural awareness, thinking capacity and learning ability (Gao, 2018; MOE, 2018; Wang, 2018; Zhang *et al.*, 2019)."

Although the juxtaposition of multiple sources appears to signal "linking", it in essence gathers and lists viewpoints, mirroring the reviewers' criticisms of "summarizing without evaluating" and "no discernible gap". Thus the three strands of evidence—the quantitative functional distribution, the thematic criticisms in peer review, and the original citation sentences—converge to support the core judgment of this study: the difference between the two groups lies not so much in the quantity of citations as in whether citation can be elevated from knowledge telling to knowledge transformation that carries evaluation and research positioning.

5. Discussion

5.1. A Qualitative Difference from Knowledge Telling to Knowledge Transformation

The most intriguing finding of this study runs precisely counter to the naive intuition that higher-rated theses cite more, and more densely. When the introduction and literature review modules are pooled, the citation density of the low-rated group (18.21 per 1,000 words) is in fact higher than that of the high-rated group (16.88 per 1,000 words)—the emphasis here is on direction rather than magnitude; at the thesis level the two densities do not differ significantly (Chapter 4, **Table 3**)—and its proportion of integral citations is also higher (70.9% vs 58.5%, corresponding to 29.1% vs 41.5% for non-integral citations). This means that the low-rated group did not simply cite less; rather, these writers transported large quantities of literature into their reviews item by item, in a list-like "author (year) + reporting verb" fashion. In Bereiter and Scardamalia's (1987) terms, this is typical knowledge telling: the writing task is reduced to "saying what one knows," and, judging from the textual surface, the more one cites, the closer the citations come to serving as proof that "I have done the reading". Such paraphrase-dominated listing is akin in nature to Pecorari's (2003, 2006) account of "patchwriting"—usually not deliberate plagiarism but a symptom of insufficient citation competence [13] [14].

The high-rated group, by contrast, shows signs of a transition toward knowledge transformation: fewer citations, deployed with greater purpose. First, its share of attribution is slightly lower (high 61.9% vs low 65.7%; the direction matches Petrić's (2007) finding that the low-rated group shows the higher attribution share, though the magnitude here is far more moderate—nearly 13 percentage points in her study, low 91.54% vs high 78.77%, against only about 4 percentage points here), meaning relatively fewer citations remain at the level of "who said what". This attribution-dominant overall pattern is also consistent with Mansourizadeh and Ahmad's (2011) observation of non-native novice writers. Second, the two functions that most directly carry authorial stance—evaluation (high 7.9% vs low 5.8%; note that the low-rated group here still devotes 5.8% of its citations to evaluation, far from the near-zero 0.63% of Petrić's low-rated group, whose high-rated counterpart stood at 5.51%) and statement of use (high 5.3% vs low 2.5%, more than double [16]; in the introductions the gap is more pronounced—6.3% vs 0.8% (9/143 vs 1/121, Fisher exact $p = 0.023$), matching the direction of Petrić's 9.42 vs 2.90 for introductions, somewhat smaller in percentage points yet steeper as a ratio)—are both higher in the high-rated group, statement of use being the only indicator to reach thesis-level significance (exact $p = 0.011$; see **Table 3** in Chapter 4), while evaluation shows a same-direction tendency that does not reach significance ($p = 0.436$); and comparison of one's own findings with other sources (comparison_own), which brings others' findings into dialogue with the present study, occurs only in the high-rated group (2 instances vs 0—small in absolute terms, and thus

illustrative rather than statistical evidence). Third, the high-rated group's share of multi-function citations is roughly 1.8 times that of the low-rated group (high 3.8% vs low 2.1%; in Petrić's study, 9.34% vs 3.79%, the same direction; item-level $p = 0.007$ but thesis-level $p = 0.486$, hence a tendency), with individual citations shouldering more complex rhetorical work. In other words, beyond attribution, high-rated writers more often let a citation double as evaluation and positioning, and less often stop at mere display. This combination of "lower density, higher-order functions" can be regarded as a behavioral signature of citation competence moving from telling to transformation; it also suggests that among the three dimensions of citation competence discussed by Xu (2016)—breadth of content coverage, interpersonal interaction, and discourse integration—the "interpersonal interaction" (evaluative stance) dimension may discriminate most sharply between the groups (as far as the functional indicators observed in this study are concerned).

Taken together, the present study reproduces the direction of Petrić's (2007) between-group pattern in functional distribution—an attribution-dominant low-rated group and relatively more active higher-order functions in the high-rated group—but with gaps that are on the whole more moderate, of which only statement of use reaches thesis-level statistical significance (Chapter 4, Table 3). Strictly, what this design licenses is a claim of association rather than of replicated competence differences: most functional contrasts are non-significant at the thesis level, and the reviewers' comments used for triangulation originate in the same assessment process that defined the groups (Section 3.4), so the observed contrasts are best read as associations between citation-function profiles and review-based ratings. The causes of this "same direction, moderate magnitude" pattern are inseparable from the boundaries within which the present conclusions hold, and the two are addressed together in the next subsection.

5.2. Tendency-Level Conclusions and Their Limits

The primary cause of the moderate magnitudes is that neither group is an extreme group: the mean total review scores of the high- and low-rated groups are approximately 84 versus 78, and the reviewers' overall verdicts are mostly A and B (13A/7B in the high-rated group; 1A/19B/2C in the low-rated group). Both groups consist of "relatively strong" Master of Education (M.Ed.) theses, rather than spanning the two extremes of quality as in Petrić; a dramatic between-group gap should not be expected in the first place. Accordingly, most functional differences amount to only a few percentage points and are non-significant at the thesis level (Chapter 4, Table 3); in terms of the per-thesis mean share of attribution, the high- and low-rated groups stand at 63.5% and 65.4% respectively, a gap of less than 2 percentage points (to be distinguished from the overall shares of 61.9% vs 65.7% reported above with the two modules combined). Thesis-by-thesis robustness checks further show that within-group dispersion exceeds the between-group gap: the standard deviation of the per-thesis attribution share reaches 15.7 in the

low-rated group (12.7 in the high-rated group); some low-rated theses have attribution shares as high as 87.4%, verging on pure paraphrase, while others fall as low as 36.9%, indistinguishable from the high-rated group—"group" is therefore a tendency, not an iron law.

A second boundary concerns corpus coverage: this study sampled only the introduction (Chapter 1) and the literature review module—both literature-dense sections in which paraphrase is the norm—and excluded the analysis chapters, where the application function diverged most sharply in Petrić's corpus (reaching 15.2 vs 3.61 between the high- and low-rated groups), thereby compressing the space in which higher-order functions could diverge and possibly underestimating the true gap between the groups.

Third, for establishing links between sources (links), this study points in the opposite direction from Petrić: the low-rated group's share (21.5%) is actually higher than the high-rated group's (19.1%), contradicting Petrić's finding that the high-rated group produced more links (5.99 vs 3.16) and sitting uneasily with theoretical expectations; moreover, judging what counts as a "link" depends considerably on coders' subjective delimitation, so this study does not build its argument on links. The very small application category (high 1.2% vs low 1.4%) also runs in the opposite direction, though with a gap of only 0.2 percentage points on sparse counts it carries no interpretive weight. The direction of citation density likewise runs counter to Petrić's (6.85 in her high-rated and 6.2 in her low-rated theses per 1,000 words); her density, however, was computed over whole theses, whereas ours covers only the Introduction and Literature Review modules, so the two figures are not directly comparable—and the difference is in any case non-significant at the thesis level here. The most defensible evidence lies in evaluation and statement of use—two categories with explicit linguistic markers and relatively consistent coding—and in the directional judgment, convergently supported by the reviewers' comments, that the low-rated group is attribution-dominant and evaluation-poor.

One methodological difference unrelated to effect size should also be noted: this study triangulated the counts with the 42 review forms, so that interpretation need not rest on frequencies alone—themes such as "the review lacks a summary and no gap is discernible," "the discussion is not compared with previous studies," and "weak theoretical basis" cluster in the low-rated group (for the frequencies see Table 5 in Chapter 4), mirroring the quantitative shortfall in the evaluation function; such convergent evidence from the reviewers' comments—external to the researchers though not independent of the grouping process (Section 3.4)—was not available in Petrić's study. The present study gains ecological validity with a design of "moderate contrast plus multi-source corroboration," at the cost of effect sizes less striking than Petrić's—hence the conclusions should be stated as "tendencies" rather than "settled verdicts".

Accordingly, the theoretical significance of this study can be distilled into a single sentence: citation competence is best understood as a continuous gradient

from knowledge telling to knowledge transformation, rather than a matter of presence versus absence. It should be noted that Bereiter and Scardamalia's (1987) original distinction is a typology of two writing models; rereading it as a continuous gradient is an extension made by the present study on the basis of the evidence that within-group dispersion exceeds the between-group gap, not the original sense of the framework. Notably, both groups score 10/10 in providing an explicit research gap or a summary of the review, indicating that the low-rated group already possesses the skeleton of a literature review; what it lacks is the "last mile" of making citations carry evaluation and positioning. This resonates with the competence denoted by the "interpersonal interaction" dimension of Xu's (2016) three-dimensional framework, and it pinpoints the target for pedagogical intervention: there is no need to train "how to cite" from scratch; rather, effort should concentrate on remedying the higher-order functions of evaluation, statement of use, and comparison of one's own findings with other sources—which is precisely where the pedagogical model below takes aim.

6. A Diagnosis-Progression-Feedback Model of Citation Competence Instruction for M.Ed. Students

Grounded in the empirical findings above, this section constructs a citation competence instruction model that takes Petrić's nine citation functions as its analytical yardstick and the remediation of higher-order functions as its target. What the model specifies is not "which classroom activities to run," but on what evidence to teach, in what sequence to teach, and by what criterion to judge whether the teaching has worked; the selection and arrangement of concrete classroom activities are left to users, to be organized according to the model's operational guidelines. The model proposed in this section is a pedagogical framework proposal based on the empirical findings above; it has not yet been tested through classroom teaching experiments—though classroom evaluation of instruction in writing from sources has proved feasible in EAP settings [28]—and the empirical testing of its effectiveness is taken up in Section 7.3.

6.1. Rationale and Overview of the Model

The model rests on three empirical conclusions from Section 4 and 5. First, the gap between the two groups lies mainly not in citation quantity but in the two stance-bearing higher-order functions of evaluation and statement of use—this fixes the model's targets. Second, the low-rated group already possesses the skeleton of a literature review (both groups score 10/10 in providing an explicit research gap or review summary); what is missing is the "last mile" of making citations carry evaluation and positioning—so instruction need not start from "how to cite," but should focus on functional upgrading. Third, within-group dispersion exceeds the between-group gap—so instructional decisions must start from individual diagnosis rather than treating "the group" as one undifferentiated block.

Accordingly, the model adopts a "three layers, one loop" architecture: a diagnostic

input layer at the front; a three-stage progression layer in the middle (functional awareness—linguistic resources—integrated application); and an evaluation-feedback layer at the end, with rubric-based re-assessment looping back to diagnosis. The target output is observable movement of learners' citation behavior along the "knowledge telling → knowledge transformation" gradient. The model is designed for M.Ed. thesis writers in subject teaching (English)—the same population from which this study's corpus was drawn. The model is visualized in Figure 1.

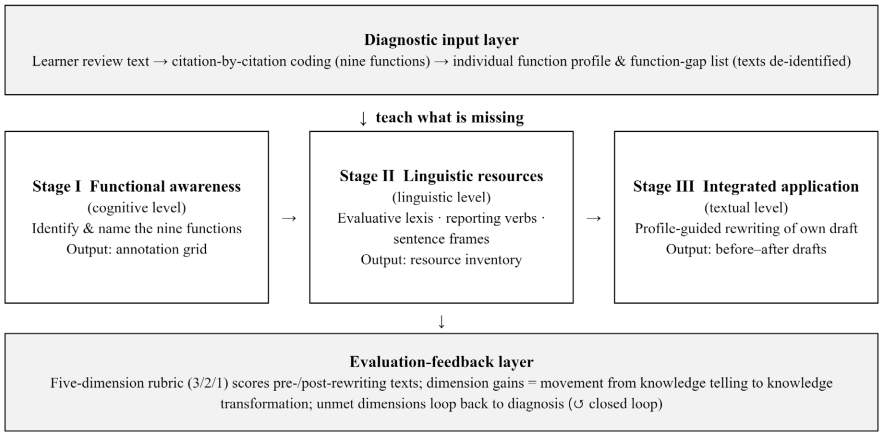

Figure 1. The "diagnosis-progression-feedback" model for citation competence instruction.

6.2. Components and Operating Mechanism

The diagnostic input layer codes the learner's review text citation by citation, using the nine functions as the coding yardstick, to generate an individual "citation function profile" with the same indicator set as Section 4: citation density, integral/non-integral shares, attribution share, the shares of evaluation, statement of use, exemplification and links, and the share of multi-function citations (all learner texts are de-identified, with author information removed). Gap determination requires a reference baseline: the per-thesis means of this study's high-rated group may serve as the initial reference (e.g., evaluation 7.94%, statement of use 5.49%)—an indicator clearly below the reference, or an attribution or integral share clearly above the corresponding per-thesis reference (63.5% and 63.0% respectively), marks a candidate target; these reference values come from the present corpus and are a starting point only, to be updated as local data accumulate. Diagnosis is the basis of every instructional decision: teach what is missing.

In the three-stage progression layer, the three stages operate at the cognitive, linguistic, and textual levels respectively, advancing step by step, with the target functions running through all three; the operational guidelines of each stage—goal, core operation, input-output, and progression criterion—are given in Table 6.

Table 6. Operational guidelines of the three-stage progression layer.

Stage (level)	Operational goal	Core operation	Input → Output	Progression criterion (suggested)
I Functional awareness (cognitive)	Identify and name the rhetorical function of every citation (aligned with Petrić's nine categories and Cumming *et al.*'s (2018) taxonomy)	Citation-by-citation annotation and classification of authentic texts	De-identified texts + function definitions → personal annotation grid	Annotation stably converges with teacher checking
II Linguistic resources (linguistic)	Master the linguistic means for realizing evaluation, statement of use, and comparison with one's own study	Stance ordering and frame-based composition with evaluative adjectives/adverbs, stance-differentiated reporting verbs, contrastive connectives, and "they say—I say" frames [29]	Annotation grid + resource list → personal inventory of linguistic resources	Can produce stance-differentiated paraphrases of the same source
III Integrated application (textual)	Upgrade purely attributive citations into multi-function citations carrying evaluation or positioning, and declare the adopted framework explicitly	Profile-guided targeted rewriting of the learner's own draft	Own draft + profile and targets → before-after pair of drafts	Re-assessment meets the mastery threshold (see evaluation-feedback layer)

Progression criteria are suggested thresholds, to be calibrated by instructors; the linguistic means listed are resource types rather than specific activities, whose selection is left to users under the model's operational guidelines.

The evaluation-feedback layer assesses the text before and after rewriting with the five-dimension rubric (**Table 7**). The teacher's rating serves as the rating of record, while self- and peer assessment provide reference points for calibration. The change in a dimension's score between the two versions can then be read as an observable measure of how far the learner has moved from knowledge telling toward knowledge transformation. As a reference line for mastery, the model suggests that the target dimension reach at least 2 (Adequate) and improve by at least 1 point over the pre-rewriting text—a standard that is advisory in nature and awaits classroom validation. Dimensions that fall short return to the diagnostic layer for a new round of target-setting and intervention; should a dimension fall short in two consecutive rounds, the corresponding linguistic means have evidently not yet been mastered, and the learner is best served by revisiting the linguistic-resources stage before attempting further rewriting.

As for the operating mechanism, the model as a whole is driven by function gaps: the diagnostic layer determines what to teach, the progression layer prescribes the order in which to teach it, and the feedback layer judges whether another round is needed; the three layers are linked by two kinds of data—the func-

tion profile and the rubric scores—so that every instructional decision remains traceable to evidence. At the same time, each learner's profiles, targets, drafts, and score changes are filed round by round in a personal writing portfolio, serving formative assessment and accumulating data for future research. On the question of when the cycle ends, the model suggests a natural exit once all five dimensions reach at least 2; if this has not been achieved after two full cycles (including remedial work on linguistic resources), the learner is transferred to regular writing instruction rather than looping indefinitely.

6.3. Operating Procedure and Assessment Rubric

The model operates through five successive phases—diagnosis, target-setting, intervention, re-assessment, and iteration—which together form one complete instructional cycle. Instruction begins with diagnosis: the learner's review text is collected and coded citation by citation with the nine functions, yielding the citation function profile. On the basis of the profile, two or three priority function targets are then fixed for each learner—in light of this study's findings, evaluation and statement of use will normally come first. Intervention follows, proceeding through Stages I to III in order, with the target functions running through the operations of all three stages. Once the rewriting is complete, the new text is reassessed with the same rubric and the change in each dimension is computed. Iteration closes the cycle: dimensions that meet the standard are consolidated into writing habits, while those that fall short return to target-setting as the focus of the next round. A hypothetical case may serve to illustrate the model in operation: suppose a learner's profile shows attribution at 85%, evaluation at a mere 3%, and statement of use and multi-function citations both at zero—far removed from the reference values; statement of use and evaluation are accordingly set as the targets; the intervention unfolds across the three stages with these two functions in constant focus; at re-assessment, the "research positioning" dimension (corresponding to the statement-of-use target) rises from 1 to 2 and is consolidated, while the "evaluative stance" dimension (corresponding to the evaluation target) remains at 1 and therefore returns to target-setting as the sole target of the following round.

As the model's measurement component, the rubric characterizes citation competence along the five dimensions of functional diversity, evaluative stance, research positioning, integration of sources, and citation conventions; the level descriptors are given in Table 7.

The five dimensions correspond directly to the shortfalls identified in this study: functional diversity and evaluative stance target the deficit in the evaluation function; research positioning targets the deficit in statement of use and the "no discernible gap" problem; integration of sources addresses "list-like enumeration" (this dimension rests on the comment evidence—"discussion not compared with prior studies", low 5/22 vs high 1/20—and on the observed instances of L08-style multi-source parenthetical lists, rather than on the links counts, on which this study does not build its argument); and citation conventions respond to the label-

ling problems present in both groups yet different in nature: 7/22 low-rated and 6/20 high-rated review forms were flagged for citation-format problems—in the high-rated group mostly minor slips in details such as volume/issue and page numbers, in the low-rated group substantive defects such as mismatches between in-text citations and the reference list and unattributed figures and tables; the rubric's "Adequate (2)" and "Needs improvement (1)" anchors correspond respectively to these two kinds of problems. By taking the difference between pre- and post-rewriting scores on each dimension, teachers can quantify how far a learner has moved from knowledge telling toward knowledge transformation, rendering this movement along the gradient observable, feedback-ready, and assessable.

Table 7. A five-dimension rubric for citation competence (the model's measurement component, used for self-, peer, and teacher assessment).

Dimension	Excellent (3)	Adequate (2)	Needs improvement (1)
Functional diversity	Attributive and non-attributive citations balanced; evaluation, statement of use, and comparison of one's own findings with other sources appear in multiple places	Predominantly attribution, with occasional evaluation	Almost entirely attribution; paraphrases merely listed
Evaluative stance	Paraphrase generally followed by substantive judgment (merits and flaws, limitations, open questions)	Evaluation present but mostly polite praise	Retelling without evaluating
Research positioning	Adopted theory/framework explicitly declared; citations serve the study's research gap	Summary of the review present but only loosely tied to the study	No summary; no discernible gap
Integration of sources	Multi-source comparison brings out consensus and divergence, forming a dialogue	Multiple sources juxtaposed but merely listed	Single sources transported one by one
Citation conventions	In-text citations consistent with the reference list; markings complete	Occasional omissions of volume/issue or page numbers	Citations inconsistent with the reference list; source markings missing

Each dimension of the rubric is scored on three equidistant levels—Excellent (3), Adequate (2), and Needs improvement (1)—for a maximum total of 15 points across the five dimensions.

7. Conclusions and Limitations

7.1. Major Findings

Using Petrić's (2007) nine-category framework of the rhetorical functions of citations as a lens, this study conducted a stratified comparison of the Introduction and Literature Review modules of 10 high-rated and 10 low-rated Master of Education (M.Ed.) theses in subject teaching (English) from a Chinese university, triangulated with the qualitative comments in the 42 review forms. The central finding contradicts the naive intuition that "higher-rated means more citations": the gap between the two groups lies not so much in the quantity of citation as in its quality. Overall, this study reproduces Petrić's (2007) between-group pattern in

direction, but with more moderate magnitudes (links and the very sparse application category run in the opposite direction; the former is not used as a basis for argument). Because most contrasts are non-significant and the review comments share their provenance with the grouping, these findings are stated as associations between citation-function profiles and review-based ratings rather than as demonstrated differences in underlying competence.

First, with regard to the "quantity" at issue in research question (1), the differences run exactly counter to intuition. The low-rated group not only showed a higher citation density (with the two modules combined, low 18.21 vs high 16.88 citations per 1,000 words; in the Introduction the high-rated group's density was in fact higher, the reversal being concentrated in the Literature Review chapter; at the thesis level the two groups' densities do not differ significantly (Chapter 4, Table 3)—the high-rated group did not simply cite more), but also a higher proportion of integral citations (low 70.9% vs high 58.5%) and a higher share of the attribution function (low 65.7% vs high 61.9%). This indicates that the low-rated group was more inclined toward list-like cataloguing in the "author (year) + reporting verb" pattern, transporting sources one by one into the review and remaining at the level of "who said what"—precisely what Bereiter and Scardamalia (1987) termed "knowledge telling".

Second, with regard to quality, the high-rated group cited less but with greater purpose, showing signs of a transition toward "knowledge transformation". The two higher-order functions that most directly carry the writer's stance—evaluation (high 7.9% vs low 5.8%) and statement of use (high 5.3% vs low 2.5%, more than double that of the low-rated group)—were both higher in the high-rated group, statement of use being the only indicator to reach thesis-level significance (exact $p = 0.011$; $q = 0.023$ among the two pre-designated confirmatory indicators, $q = 0.090$ under full-family BH correction), with evaluation a same-direction tendency; comparison of one's own findings with other sources (comparison_own), which brings prior findings into dialogue with the present study, was found only in the high-rated group (though with an absolute count of merely 2 instances, this constitutes illustrative rather than statistical evidence); the proportions of non-integral citations and multi-function citations were likewise higher (multi-function high 3.8% vs low 2.1%, roughly 1.8 times that of the low-rated group, non-significant at the thesis level and thus a tendency). The advantage of the high-rated group lay not in citing more, but in deploying citation as a rhetorical resource for expressing evaluative stance and anchoring the positioning of the study. This suggests that, among the three dimensions discussed by Xu (2016), the "interpersonal interaction" (evaluative stance) dimension may be the most discriminating in the stratification of L2 citation competence (as far as the functional indicators observed in this study are concerned).

Third, turning to research question (2), the thematic patterns in the comments from the 42 thesis review forms align on the whole with the quantitative picture above, providing qualitative triangulation. The low-rated group was repeatedly

criticized for inadequately realized "basic elements" (mostly perfunctory summaries and research gaps proposed without weighing the literature, rather than missing sections)—"no summary of the review / no discernible research gap" (low 5/22 vs high 1/20), "weak theoretical basis or disconnection from the study" (low 7/22 vs high 2/20), and "discussion not supported by the literature and lacking comparison with prior studies" (low 5/22 vs high 1/20). The criticisms received by the high-rated group operated at a higher level, mostly pointing to the deepening of knowledge transformation, such as "insufficient critical analysis" and "weak articulation with the local context". The quantitative deficit in the evaluation function manifests itself precisely as the reviewers' qualitative judgments of "summarizing without evaluating" and "no discernible gap".

Accordingly, and in answer to research question (3), the practical implications of this study are clear and convergent: citation competence is best understood as a continuous gradient from "knowledge telling" to "knowledge transformation", rather than an all-or-nothing attribute. Both groups featured an explicit research gap or a summary of the review in 10/10 theses, indicating that the low-rated group already possessed the skeleton of a review; what it mainly lacked was the final step of making citations genuinely carry evaluation and research positioning. Pedagogical intervention therefore need not train "how to cite" from scratch (though the convention defects actually observed in the low-rated group, such as mismatches between in-text citations and the reference list, still require accompanying remediation), but should lock its target precisely onto strengthening the two higher-order functions of evaluation and statement of use—which is exactly where the pedagogical model of this paper takes aim.

7.2. Limitations

This study adopted a design of "moderate contrast plus multi-source corroboration"; its conclusions should be read as indicating tendencies rather than as definitive, and the following three limitations must be stated candidly (the limits of the coding procedure and of the links category have already been set out in Sections 3.5 and 5.2 respectively, and are not repeated here).

First, neither group is an extreme group. As noted in Section 5.2, the two groups' mean overall review scores were approximately 84 and 78 respectively, with review conclusions dominated by A and B grades; both are "relatively strong" M.Ed. theses rather than an extreme contrast, so no dramatic between-group gap should be expected in the first place. Apart from statement of use, most functional differences amount to only a few percentage points and are non-significant at the thesis level (Chapter 4, Table 3); within-group dispersion exceeds the between-group gap, and "group" therefore represents a tendency rather than an iron law.

Second, the evidence base is restricted to a single institution and a single program. The corpus was drawn exclusively from M.Ed. students in subject teaching (English), all of whom conducted empirical studies on English teaching in pri-

mary and secondary schools; the disciplinary discourse and training environment are relatively homogeneous, so generalization to other institutions, other disciplines, or indeed to academic master's and doctoral theses and journal articles must be made with particular caution.

Third, the sample is small and its module coverage incomplete. The sample comprises only 20 theses, and only two parts were extracted—the Introduction (Chapter 1) and the Literature Review module (including the theoretical-basis chapter). Both are inherently literature-dense stretches of discourse in which paraphrase is the norm, whereas the analysis chapters, where the application function diverged most sharply in Petrić's corpus (15.2 vs 3.61 between her high- and low-rated groups), were not included. This both compresses the space in which higher-order functions could differentiate, potentially underestimating the true gap between the two groups, and means that this study has not yet examined how citation competence performs in the results/discussion/analysis sections; the findings offer only a picture of the "Introduction + Literature Review" facet rather than the full landscape of citation behavior in these theses.

7.3. Future Research

In view of the above limitations, future research may deepen the inquiry in four directions. First, expanding the sample across institutions: include more institutions, more disciplines, and even academic master's and doctoral theses and journal articles, and deliberately recruit extreme groups—one dominated by A grades and one by C/D grades, genuinely spanning the two ends of quality—so as to test the robustness and boundary conditions of the "quality over quantity" conclusion embodied in the "knowledge telling-knowledge transformation" gradient. Second, extending to the analysis and discussion chapters: broaden the coding scope to the methodology, results, and discussion chapters, examining higher-order functions such as application and comparison_own, which differentiate more readily outside the literature review—the discussion section's citation behavior has been examined in its own right [27]—thereby filling the gap left by this study's coverage of only the first two modules. Third, refining coding and reliability: establish more explicit merging rules for sensitive categories such as links and expand the proportion of double coding, so that their absolute values can support argumentation. Fourth, moving from description to intervention: implement the pedagogical model of this paper (a three-layer "diagnosis-progression-feedback" closed loop whose progression backbone is the three-stage sequence of functional awareness—linguistic resources—integrated application, targeting evaluation and statement of use) in real classrooms, taking pre- and post-revision differences in rubric scores across dimensions as indicators, to empirically test whether this targeted instruction in strengthening evaluation and statement of use can effectively move learners from "knowledge telling" toward "knowledge transformation", making this movement along the citation-competence gradient genuinely observable, feedback-ready, and assessable.

Funding

This paper was supported by the 2024 Research Project of Zhejiang Yuexiu University (N2024004).

Author Contributions

Conceptualization, S-X.L.; methodology, S-X.L.; software, S-X.L. and W-L.F.; validation, S-X.L.; formal analysis, S-X.L.; investigation, S-X.L.; resources, S-X.L. and W-L.F.; data curation, S-X.L. and W-L.F.; writing—original draft preparation, S-X.L.; writing—review and editing, S-X.L.; visualization, S-X.L.; supervision, S-X.L. and W-L.F.; project admin-istration, S-X.L.; funding acquisition, S-X.L. All authors have read and agreed to the pub-lished version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Petrić, B. (2007) Rhetorical Functions of Citations in High- and Low-Rated Master's Theses. *Journal of English for Academic Purposes*, **6**, 238-253. https://doi.org/10.1016/j.jeap.2007.09.002

[2] Hyland, K. (2000) Disciplinary Discourses: Social Interactions in Academic Writing. Longman.

[3] Swales, J.M. (1990) Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

[4] Wang, J., Yang, K. and Sun, T. (2017) A Study of Citation Features in English-Major Graduate Students' Academic Writing. *Foreign Language World*, No. 5, 56-64. (In Chinese)

[5] Xu, F. (2016) Citation Competence in L2 Academic Writing and Its Developmental Features: Cross-Sectional and Longitudinal Evidence. *Journal of Foreign Languages*, No. 3, 73-82. (In Chinese)

[6] Moravcsik, M.J. and Murugesan, P. (1975) Some Results on the Function and Quality of Citations. *Social Studies of Science*, **5**, 86-92. https://doi.org/10.1177/030631277500500106

[7] Thompson, G. and Yiyun, Y. (1991) Evaluation in the Reporting Verbs Used in Academic Papers. *Applied Linguistics*, **12**, 365-382. https://doi.org/10.1093/applin/12.4.365

[8] Hyland, K. (1999) Academic Attribution: Citation and the Construction of Disciplinary Knowledge. *Applied Linguistics*, **20**, 341-367. https://doi.org/10.1093/applin/20.3.341

[9] Thompson, P. and Tribble, C. (2001) Looking at Citations: Using Corpora in English for Academic Purposes. *Language Learning & Technology*, **5**, 91-105. https://doi.org/10.64152/10125/44568

[10] Thompson, P. (2005) Points of Focus and Position: Intertextual Reference in PhD Theses. *Journal of English for Academic Purposes*, **4**, 307-323. https://doi.org/10.1016/j.jeap.2005.07.006

[11] Hyland, K. and Jiang, F. (2019) Points of Reference: Changing Patterns of Academic

Citation. *Applied Linguistics*, **40**, 64-85. https://doi.org/10.1093/applin/amx012

[12] Pennycook, A. (1996) Borrowing Others' Words: Text, Ownership, Memory, and Plagiarism. *TESOL Quarterly*, **30**, 201-230. https://doi.org/10.2307/3588141

[13] Pecorari, D. (2003) Good and Original: Plagiarism and Patchwriting in Academic Second-Language Writing. *Journal of Second Language Writing*, **12**, 317-345. https://doi.org/10.1016/j.jslw.2003.08.004

[14] Pecorari, D. (2006) Visible and Occluded Citation Features in Postgraduate Second-Language Writing. *English for Specific Purposes*, **25**, 4-29. https://doi.org/10.1016/j.esp.2005.04.004

[15] Bereiter, C. and Scardamalia, M. (1987) The Psychology of Written Composition. Lawrence Erlbaum Associates.

[16] Mansourizadeh, K. and Ahmad, U.K. (2011) Citation Practices among Non-Native Expert and Novice Scientific Writers. *Journal of English for Academic Purposes*, **10**, 152-161. https://doi.org/10.1016/j.jeap.2011.03.004

[17] Lin, C.-Y. and Lau, K. (2021) *Journal of English for Academic Purposes*, **53**, Article ID: 101035. https://doi.org/10.1016/j.jeap.2021.101035

[18] Sun, S. and Jiang, F. (2025) Utilization of Appraisal Resources for Acknowledging Limitations within Doctoral Theses across Disciplines. *Journal of English for Academic Purposes*, **75**, Article ID: 101511. https://doi.org/10.1016/j.jeap.2025.101511

[19] Petrić, B. and Harwood, N. (2013) Task Requirements, Task Representation, and Self-Reported Citation Functions: An Exploratory Study of a Successful L2 Student's Writing. *Journal of English for Academic Purposes*, **12**, 110-124. https://doi.org/10.1016/j.jeap.2013.01.002

[20] Xu, F. (2012) Citation Features in Empirical English Academic Discourse. *Journal of Foreign Languages*, **35**, 60-68.

[21] Mu, C. (2019) A Contrastive Study of Citation Forms and Rhetorical Functions in English and Chinese Research Articles from the Perspective of Intertextuality. *Journal of Qinghai Normal University* (*Philosophy and Social Sciences Edition*), No. 6, 116-124. (In Chinese)

[22] Xu, H. (2011) Authorial Stance Markers in the Academic Discourse of Advanced Chinese EFL Learners: A Corpus-Based Contrastive Study. *Foreign Language Education*, No. 6, 44-48.

[23] Wei, X., Jiang, Z., Chang, X. and Shen, L. (2023) Evaluating the Originality of Research Papers Based on Citation Intent. *Information Studies*: *Theory & Application*, No. 9, 24-30, 46. (In Chinese)

[24] Cumming, A., Yang, L., Qiu, C., Zhang, L., Ji, X., Wang, J., *et al.* (2018) Students' Practices and Abilities for Writing from Sources in English at Universities in China. *Journal of Second Language Writing*, **39**, 1-15. https://doi.org/10.1016/j.jslw.2017.11.001

[25] Li, Q. and Zhang, X. (2021) An Analysis of Citations in Chinese English-Major Master's Theses and Doctoral Dissertations. *Journal of English for Academic Purposes*, **51**, Article ID: 100982. https://doi.org/10.1016/j.jeap.2021.100982

[26] Ahn, C.-Y. and Oh, S.-Y. (2024) Citation Practices in Applied Linguistics: A Comparative Study of Korean Master's Theses and Research Articles. *Journal of English for Academic Purposes*, **69**, Article ID: 101369. https://doi.org/10.1016/j.jeap.2024.101369

[27] Samraj, B. (2013) Form and Function of Citations in Discussion Sections of Master's

Theses and Research Articles. *Journal of English for Academic Purposes*, **12**, 299-310. https://doi.org/10.1016/j.jeap.2013.09.001

[28] Wette, R. (2010) Evaluating Student Learning in a University-Level EAP Unit on Writing Using Sources. *Journal of Second Language Writing*, **19**, 158-177. https://doi.org/10.1016/j.jslw.2010.06.002

[29] Graff, G. and Birkenstein, C. (2010) "They Say/I Say": The Moves That Matter in Academic Writing. 2nd Edition, W. W. Norton.