Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.104113
Published by: Zeus Press

Development and Validation of a Scale for Assessing the Effectiveness of Generative AI-Enhanced English Writing Learning Among Chinese University Students

Jingyu Wang*

Department: School of Foreign Languages, Shandong Normal University, Jinan 250000, China

**Corresponding author: Jingyu Wang.*

Abstract

This study, grounded in the Technology Acceptance Model (TAM) and student engagement theory, aims to develop a psychometric instrument for assessing the effectiveness of generative AI-enhanced English writing learning among Chinese university students. Drawing upon established scales from domestic and international literature, the initial scale was adapted to the specific context of generative AI-assisted English writing, comprising three dimensions—AI Use Attitude, Learning Engagement, and Learning Outcomes—with 42 items. Following expert review, the scale was administered to university students across 28 provinces in China, yielding 282 valid responses. Data were analyzed using SPSS Statistics 27, employing item analysis, exploratory factor analysis (EFA), and reliability and validity testing. Item analysis, based on critical ratio (CR), corrected item-total correlation (CITC), communality, and factor loading, resulted in the deletion of six items (Q8, Q15, Q27, Q35, Q39, Q42), retaining 36 items in the final scale. Exploratory factor analysis supported a three-dimensional structure: "AI Use Attitude—Learning Engagement—Learning Outcomes." Reliability analysis demonstrated Cronbach's α coefficients exceeding 0.87 for all dimensions and the overall scale, with a split-half reliability of 0.947. Validity testing confirmed satisfactory content validity and structural validity. The findings indicate that the scale possesses robust psychometric properties and practical feasibility, providing a valid instrument for evaluating the effectiveness of generative AI-enhanced English writing instruction.

Keywords

generative AI, English writing, learning effectiveness, scale development, reliability and validity

1. Introduction

In the era of deep integration between educational digitalization and artificial intelligence, generative artificial intelligence (Generative AI) is rapidly reshaping the landscape of higher education. At the national policy level, there has been a clear mandate to leverage AI to drive educational transformation and promote the digital transition of education. Generative AI has been integrated into university English writing instruction, offering support across writing assistance, feedback generation, error correction, and personalized resource provision [1]. However, existing research has predominantly focused on descriptive accounts of technological application models and the construction of theoretical frameworks [2]. From a learner-centred perspective,

there is a significant gap in systematic empirical evidence regarding the actual learning outcomes of generative AI-assisted English writing.

Learning effectiveness assessment constitutes the core link for examining the value of educational technology applications. Currently, foreign language learning effectiveness assessment instruments are largely oriented toward traditional classroom teaching contexts, with limited incorporation of key variables such as learners' acceptance attitudes toward AI technology and multidimensional learning engagement into a unified evaluation framework. Consequently, these instruments fail to accurately capture the authentic effectiveness and underlying mechanisms of university students' English writing learning in AI-empowered environments. Therefore, developing a scientifically grounded and contextually applicable assessment scale has become an urgent need to accurately diagnose the strengths and weaknesses of technology empowerment and to drive the transformation of educational technology applications from "experience-driven" to "data-driven."

In view of this, the present study draws upon the Technology Acceptance Model (TAM), Kirkpatrick's Evaluation Model, and student engagement theory. Referencing established scales from domestic and international literature [3, 4], and integrating the contextual characteristics of generative AI-assisted English writing among Chinese university students, this study developed the "Scale for Assessing the Effectiveness of Generative AI-Enhanced English Writing Learning." Through questionnaire surveys administered across multiple universities nationwide, and employing SPSS for item analysis, exploratory factor analysis, and reliability and validity testing, this study aims to construct a measurement instrument with clear structure, robust psychometric properties, and direct applicability for empirical research, thereby providing metrological support for subsequent investigations into the precise assessment and optimization pathways of generative AI-enhanced English writing instruction.

2. Research Subjects and Methods

2.1 Research Subjects

This study employed convenience sampling to distribute questionnaires via an online survey platform to university students across 28 provinces in China. Participants completed the survey autonomously through mobile devices. A total of 282 valid questionnaires were collected. The demographic characteristics of the sample are presented in Table 1. The sample comprised 120 males (42.6%) and 162 females (57.4%). By academic year, the majority were third-year students (41.1%), followed by second-year students (23.4%). Regarding academic disciplines, students from humanities and social sciences constituted the largest group (40.8%), followed by science and engineering (31.6%), with the remainder distributed across education, arts, business, law, and other fields (27.7%). In terms of generative AI usage frequency, the majority used AI tools 1–2 times per week (35.5%) or 3–5 times per week (29.8%).

Table 1: Demographic Characteristics of the Sample (n = 282)

Category	Frequency (n)	Percentage (%)
Gender		
Male	120	42.6
Female	162	57.4
Academic Year		
First Year	33	11.7
Second Year	66	23.4
Third Year	116	41.1
Fourth Year	41	14.5
Graduate and above	26	9.2
Academic Discipline		
Humanities and Social Sciences	115	40.8
Science and Engineering	89	31.6
Others (Education, Arts, Business, Law, etc.)	78	27.7
Generative AI Usage Frequency		
1–2 times per week	100	35.5
3–5 times per week	84	29.8
Other frequencies	98	34.8

Note: The "Others" category in Academic Discipline includes students from education, arts, business administration, and law. The "Other frequencies" category encompasses all usage frequencies beyond the two primary categories listed above.

2.2 Scale Design

This study integrated the Technology Acceptance Model (TAM), Kirkpatrick's Evaluation Model, and student engagement theory to construct the theoretical framework of the scale. The Technology Acceptance Model, proposed by Davis based on the Theory of Reasoned Action, posits that perceived usefulness and perceived ease of use are core variables influencing technology adoption, with both factors affecting behavioral intention through attitude toward use. In the field of AI-enhanced education, this model has been empirically validated as effective in explaining learners' acceptance mechanisms for intelligent technologies. Kirkpatrick's Evaluation Model is an internationally applied training assessment framework comprising four hierarchical levels: reaction, learning, behavior, and results, encompassing both process-oriented and summative evaluation, and capable of comprehensively assessing learning behaviors, psychological performance, and learning outcomes [5]. Given that generative AI-assisted English writing emphasizes not only linguistic knowledge acquisition but also the comprehensive enhancement of writing literacy, learning engagement, and tangible outcomes—characterized by flexible formats and rich content that defy simple application of traditional writing assessment criteria—the hierarchical structure and cross-dimensional evaluation features of Kirkpatrick's Model align closely with this context.

Regarding learning engagement, human-AI collaborative writing engagement represents a novel construct in the era of generative artificial intelligence, referring to the state of active behavioral, cognitive, and emotional participation exhibited by foreign language learners during dynamic, bidirectional interaction with generative AI to co-construct textual meaning [6]. Previous studies have shown that this construct is context sensitive, and there are differences in learners' levels of engagement in different types of writing tasks. AI comprehension, learners' AI literacy, and teacher scaffolding are key factors affecting engagement.

Synthesizing the aforementioned theoretical frameworks, this study integrated AI Use Attitude (corresponding to perceived usefulness, perceived ease of use, and behavioral intention in TAM), Learning Engagement, and Learning Outcomes into a three-dimensional assessment framework. The Learning Engagement dimension was expanded beyond Xu's [6] tripartite structure of behavioral, cognitive, and emotional engagement to include a social engagement sub-dimension based on Fredricks et al.'s student engagement theory, encompassing learners' social interactions such as exchanging AI-assisted writing experiences with peers and consulting instructors.

Drawing upon established scales from domestic and international literature, and through multiple rounds of student interviews and expert consultation, the items were adapted to the disciplinary context of generative AI-assisted English writing. The scale employs a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The initial scale comprised three dimensions: AI Use Attitude (10 items), Learning Engagement (17 items, including 7 cognitive engagement items, 3 emotional engagement items, 4 behavioral engagement items, and 3 social engagement items), and Learning Outcomes (15 items), totaling 42 items. Items 8, 15, 27, 35, and 39 were reverse-scored items, which were recoded prior to data analysis.

2.3 Statistical Methods

Following data collection, SPSS Statistics 27 was employed for statistical processing, sequentially conducting item analysis, EFA, and reliability and validity testing.

Item analysis adopted a combined strategy of the critical ratio method and homogeneity testing: participants were divided into high-scoring (top 27%) and low-scoring (bottom 27%) groups based on total scale scores, and independent samples t-tests were conducted for each item. Items with critical ratios (CR, i.e., t-values) ≥ 3.000 and $p < 0.05$ were considered to possess significant discriminatory power. Concurrently, corrected item-total correlation (CITC), Cronbach's α if item deleted, communality, and factor loading were examined, with CITC ≥ 0.40, communality ≥ 0.40, and factor loading ≥ 0.50 (loading on the hypothesized dimension) serving as reference criteria for item retention, based on which comprehensive judgments regarding item retention or deletion were made.

We performed exploratory factor analysis separately for each dimension, with KMO > 0.80 and Bartlett's test of sphericity $p < 0.05$ serving as criteria for factor analysis suitability. Principal component analysis was employed to extract components with eigenvalues greater than 1, thereby examining the structural validity of the scale. For reliability and validity testing, reliability was assessed using Cronbach's α coefficient (> 0.80 considered ideal) and split-half reliability (> 0.80 considered high quality), while validity was evaluated through content validity and structural validity.

3. Research Results

3.1 Item Analysis

We examined all 42 items through item analysis. The results are presented in Table 2.

Table 2: Item Analysis Results

Item	Critical Ratio (CR)	CITC	α if Item Deleted	Communality	Decision
AI Use Attitude Dimension (Initial α = 0.779)					
Q1	12.291	0.643	0.738	0.626	Retain
Q2	14.952	0.665	0.735	0.635	Retain
Q3	8.004	0.538	0.751	0.477	Retain
Q4	12.785	0.569	0.744	0.503	Retain
Q5	11.015	0.558	0.752	0.515	Retain
Q6	11.015	0.541	0.748	0.534	Retain
Q7	11.427	0.623	0.740	0.531	Retain
Q8	1.267	-0.123	0.873	0.755	Delete
Q9	9.161	0.529	0.755	0.574	Retain
Q10	11.296	0.572	0.745	0.504	Retain
Learning Engagement Dimension (Initial α = 0.830)					
Q11	11.238	0.597	0.817	0.575	Retain
Q12	12.337	0.575	0.815	0.487	Retain
Q13	10.912	0.571	0.815	0.526	Retain
Q14	10.706	0.519	0.818	0.605	Retain
Q15	0.772	-0.070	0.866	0.747	Delete
Q16	8.431	0.455	0.820	0.545	Retain
Q17	12.144	0.614	0.814	0.577	Retain
Q18	15.599	0.684	0.808	0.616	Retain
Q19	13.104	0.636	0.812	0.518	Retain
Q20	13.611	0.643	0.813	0.579	Retain
Q21	10.100	0.486	0.818	0.515	Retain
Q22	12.805	0.643	0.810	0.620	Retain
Q23	10.938	0.577	0.814	0.493	Retain
Q24	8.846	0.445	0.821	0.566	Retain
Q25	11.169	0.580	0.814	0.550	Retain
Q26	11.914	0.555	0.813	0.587	Retain
Q27	0.775	-0.095	0.865	0.760	Delete
Learning Outcomes Dimension (Initial α = 0.831)					
Q28	12.842	0.643	0.812	0.650	Retain
Q29	16.268	0.706	0.807	0.640	Retain
Q30	13.851	0.660	0.811	0.587	Retain
Q31	12.354	0.637	0.811	0.539	Retain
Q32	11.541	0.604	0.812	0.519	Retain
Q33	14.245	0.660	0.809	0.589	Retain
Q34	13.408	0.642	0.811	0.592	Retain
Q35	0.546	-0.083	0.875	0.817	Delete
Q36	12.428	0.612	0.812	0.563	Retain
Q37	16.151	0.734	0.805	0.663	Retain
Q38	12.250	0.681	0.809	0.607	Retain
Q39	0.877	-0.027	0.868	0.800	Delete

Q40	12.718	0.697	0.808	0.647	Retain
Q41	13.965	0.672	0.808	0.577	Retain
Q42	5.705	0.235	0.837	0.331	Delete

Note: Shaded rows indicate deleted items. Items 8, 15, 27, 35, and 39 were reverse-scored. CITC = corrected item-total correlation; "α if Item Deleted" = Cronbach's alpha of the dimension if the item is removed.

3.1.1 AI Use Attitude Dimension

The initial Cronbach's α coefficient for this dimension was 0.779. Item 8 demonstrated a negative corrected item-total correlation ($r = -.12$), indicating that it did not correlate positively with the dimension construct. Furthermore, in factor analysis, Item 8 failed to load on the principal component of this dimension, instead loading with other reverse-scored items on a separate component. The deletion of Item 8 increased the dimension's α coefficient from 0.779 to 0.873, warranting its removal. The remaining nine items (Items 1, 2, 3, 4, 5, 6, 7, 9, 10) demonstrated corrected item-total correlations ranging from 0.529 to 0.665 (all > 0.40), with critical ratios ranging from 8.00 to 14.95 (all $p < .001$), and communalities exceeding .40 ($M = .52$, $SD = .08$) and satisfactory discriminatory power and homogeneity, thus being retained.

3.1.2 Learning Engagement Dimension

The initial Cronbach's α coefficient for this dimension was 0.830. Items 15 and 27 (both reverse-scored) exhibited corrected item-total correlations of −0.070 and −0.095 (both < 0), with critical ratios of 0.772 and 0.775 (both < 3.000), indicating insufficient discriminatory power. Neither item loaded on the principal component of this dimension. The deletion of these items increased the dimension's α coefficient to 0.866 and 0.865, respectively, warranting their removal. The remaining 15 items (Items 11–14, 16–26) demonstrated corrected item-total correlations ranging from 0.445 to 0.684 (all > 0.40), critical ratios from 8.431 to 15.599 (all > 3.000), and satisfactory discriminatory power and homogeneity, thus being retained.

3.1.3 Learning Outcomes Dimension

The initial Cronbach's α coefficient for this dimension was 0.831. Items 35 and 39 (both reverse-scored) exhibited corrected item-total correlations of −0.083 and −0.027 (both < 0), with critical ratios of 0.546 and 0.877 (both < 3.000). The deletion of these items increased the dimension's α coefficient to 0.875 and 0.868, respectively, indicating insufficient discriminatory power, thus warranting their removal. Item 42 exhibited a corrected item-total correlation of 0.235 (< 0.40), communality of merely 0.331 (< 0.40), and factor loading below 0.50. Although its critical ratio reached 5.705, its homogeneity was insufficient, and its deletion increased the α coefficient from 0.831 to 0.837, thus warranting concurrent removal. For the remaining 12 items (items 28-34, 36-38, 40, 41), the total correlation range of corrected items is 0.604 to 0.734 (all>0.40), the critical ratio range is 11.541 to 16.268 (all>3.000), and the factor load range is 0.719 to 0.798 (all>0.50), so it is retained.

In summary, item analysis resulted in the deletion of six items (Items 8, 15, 27, 35, 39, 42), with 36 items retained in the final scale. Notably, the five reverse-scored items (Items 8, 15, 27, 35, 39) collectively loaded on a separate component in factor analysis, exhibiting a typical reverse-wording effect—an artifact attributable to item wording direction rather than substantive content differences. Their deletion is statistically and substantively justified.

3.2 Factor Analysis

We performed exploratory factor analysis separately for each dimension following the deletion of unsatisfactory items. The results are summarized in Table 3, and the factor loading matrix for retained items is presented in Table 4.

Table 3: Summary of Exploratory Factor Analysis by Dimension

Dimension	KMO	Bartlett's χ^2	df	Components	Cumulative Variance Explained (%)
AI Use Attitude	0.902	973.483	36	1	50.234
Learning Engagement	0.918	1927.218	105	2	54.677
Learning Outcomes	0.953	2016.767	66	1	58.707
Overall Scale (36 items)	0.956	6251.359	630	—	—

Note: All Bartlett's tests of sphericity were significant at $p < 0.001$. The overall scale analysis was conducted to examine structural validity (see Section 4.2).

3.2.1 AI Use Attitude Dimension

The KMO value was 0.902, and Bartlett's test of sphericity yielded an approximate χ^2 of 973.483 (df = 36, $p < 0.001$), indicating suitability for factor analysis. One component with eigenvalue greater than 1 was extracted, explaining 50.234% of the variance. The factor loadings ranged from 0.669 to 0.747 (all > 0.50), confirming a unidimensional structure.

3.2.2 Learning Engagement Dimension

The KMO value was 0.918, and Bartlett's test of sphericity yielded an approximate χ^2 of 1927.218 (df = 105, $p < 0.001$), indicating suitability for factor analysis. Two components with eigenvalues greater than 1 were extracted, with a cumulative variance explained of 54.677%. The factor loadings ranged from 0.522 to 0.773. It should be noted that while the theoretical framework posited four sub-dimensions (cognitive, emotional, behavioral, and social engagement), the empirical data converged into two components: Factor 1 primarily comprised behavioral engagement, social engagement, and partial emotional engagement items, reflecting learners' behavioral participation and social interaction in human-AI collaborative writing; Factor 2 primarily comprised cognitive engagement items, reflecting learners' cognitive processing and strategy deployment, with partial emotional engagement items also loading on this component. This finding suggests that in the context of generative AI-assisted writing, behavioral-social participation and cognitive processing constitute two relatively independent core facets of learning engagement, while emotional engagement exhibits cross-loadings with both (see Section 5.2 for discussion).

3.2.3 Learning Outcomes Dimension

The KMO value was 0.953, and Bartlett's test of sphericity yielded an approximate χ^2 of 2016.767 (df = 66, $p < 0.001$), indicating suitability for factor analysis. One component with eigenvalue greater than 1 was extracted, explaining 58.707% of the variance. The factor loadings ranged from 0.719 to 0.798 (all > 0.50), confirming a unidimensional structure.

Table 4: Factor Loading Matrix for Retained Items

Item	Factor 1	Factor 2
AI Use Attitude (Unidimensional, 50.234% variance explained)		
Q1	0.747	—
Q2	0.745	—
Q3	0.669	—
Q4	0.699	—
Q5	0.706	—
Q6	0.696	—
Q7	0.726	—
Q9	0.691	—
Q10	0.694	—
Learning Engagement (Two-factor, 54.677% cumulative variance)		
Q11	0.321	0.677
Q12	0.303	0.624
Q13	0.287	0.663
Q14	0.108	0.773
Q16	0.106	0.712
Q17	0.242	0.712
Q18	0.639	0.458
Q19	0.458	0.522
Q20	0.470	0.597
Q21	0.689	0.182
Q22	0.725	0.286
Q23	0.621	0.333
Q24	0.735	0.099
Q25	0.696	0.215
Q26	0.735	0.187
Learning Outcomes (Unidimensional, 58.707% variance explained)		
Q28	0.791	—

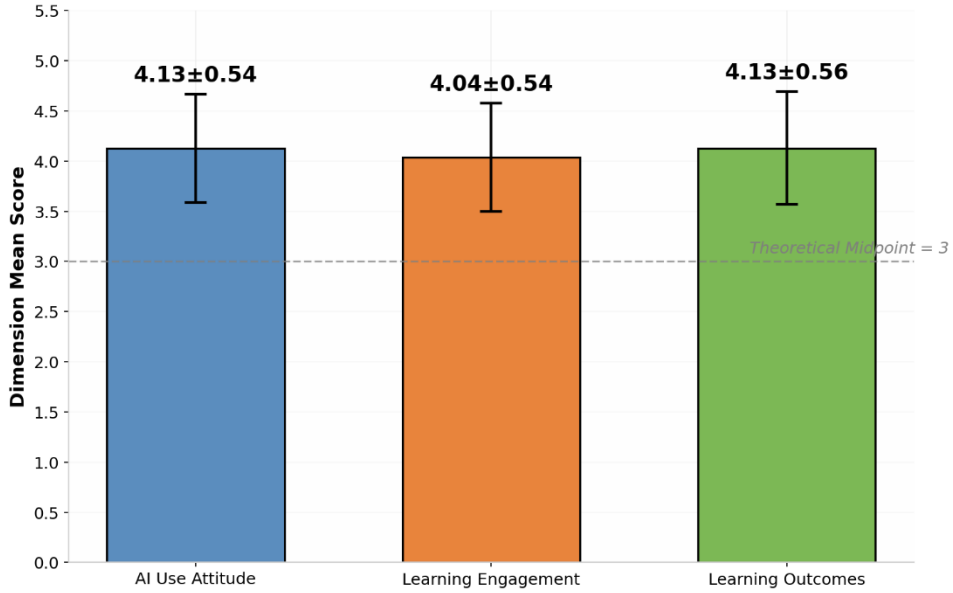
Item	Factor 1	Factor 2
Q29	0.785	—
Q30	0.768	—
Q31	0.733	—
Q32	0.719	—
Q33	0.767	—
Q34	0.760	—
Q36	0.750	—
Q37	0.797	—
Q38	0.779	—
Q40	0.798	—
Q41	0.742	—

Note: Factor 1 and Factor 2 represent loadings after varimax rotation; bold values indicate the primary loading for each item. AI Use Attitude and Learning Outcomes are unidimensional structures, with Factor 2 indicated as "—".

3.3 Descriptive Statistics

Descriptive statistics were made on the scores of 282 subjects in the three dimensions.

Figure 1: Comparison of Mean Scores and Standard Deviations Across Dimensions (n = 282)



The mean scores for all three dimensions exceeded the theoretical midpoint of 3.00, with AI Use Attitude (M = 4.13, SD = 0.54) and Learning Outcomes (M = 4.13, SD = 0.56) marginally higher than Learning Engagement (M = 4.04, SD = 0.54), indicating that university students generally hold positive attitudes toward generative AI-assisted English writing.

4. Reliability and Validity Testing

4.1 Reliability Testing

The reliability testing results are presented in Table 5. Following the deletion of unsatisfactory items, the Cronbach's α coefficients for the AI Use Attitude, Learning Engagement, and Learning Outcomes dimensions were 0.873, 0.907, and 0.935, respectively, with the overall scale (36 items) α coefficient reaching 0.963, all exceeding 0.80 and indicating ideal internal consistency. For split-half reliability, both the Spearman-Brown coefficient and the Guttman Split-Half coefficient were 0.947 (> 0.80), demonstrating stable measurement results across different item groupings and high reliability quality.

Table 5: Reliability Testing Results

Dimension / Overall	Number of Items	Cronbach's α	Split-Half Reliability

AI Use Attitude	9	0.873	—
Learning Engagement	15	0.907	—
Learning Outcomes	12	0.935	—
Overall Scale	36	0.963	0.947

Note: Split-half reliability is reported as the Spearman-Brown coefficient; the Guttman Split-Half coefficient was also 0.947. Split-half reliability was computed for the overall scale only.

4.2 Validity Testing

Content Validity: The scale items were derived from established scales in the literature, refined through multiple rounds of student interviews and expert consultation, and adapted to the disciplinary characteristics of generative AI-assisted English writing. Content validity is thereby assured.

Structural Validity: Exploratory factor analysis was conducted on the final 36 items. The KMO value was 0.956 (> 0.50), and Bartlett's test of sphericity yielded an approximate χ^2 of 6251.359 (df = 630, $p < 0.001$), indicating suitability for factor analysis. The extracted factor structure was largely consistent with the theoretical framework: AI Use Attitude and Learning Outcomes exhibited unidimensional structures, while Learning Engagement exhibited a two-factor structure. The factor loadings were substantial, indicating satisfactory structural validity of the scale. Furthermore, confirmatory factor analysis (CFA) was employed to test the three-dimensional structural model. The fit indices are presented in Table 6.

Table 6: Confirmatory Factor Analysis Fit Indices (Four-Factor Model, n = 282)

Index	χ^2/df	CFI	TLI	RMSEA	SRMR	AIC
Value	2.400	0.861	0.852	0.071	0.134	145.991
Criterion	< 3	> 0.90	> 0.90	< 0.08	< 0.08	—
Evaluation	Ideal	Acceptable	Acceptable	Ideal	Requires attention	—

Note: χ^2/df = chi-square to degrees of freedom ratio; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; AIC = Akaike information criterion.

5. Conclusion and Discussion

5.1 Reliability Discussion

Following the deletion of unsatisfactory items, the Cronbach's α coefficients for all dimensions (AI Use Attitude 0.873, Learning Engagement 0.907, Learning Outcomes 0.935) and the overall scale (0.963) substantially exceeded 0.80, with the split-half reliability coefficient reaching 0.947, indicating excellent internal consistency and stable measurement results. Notably, the deletion of the reverse-scored Item 8 in the AI Use Attitude dimension increased the α coefficient from 0.779 to 0.873, further corroborating the detrimental effect of this item on dimension internal consistency and substantiating the necessity of its removal.

5.2 Validity Discussion

The item design was grounded in classical theoretical frameworks and underwent expert review, ensuring content validity. The overall KMO value of 0.956 substantially exceeded 0.50, and Bartlett's test was significant, indicating satisfactory structural validity. The factor structure of each dimension was largely consistent with the theoretical framework: AI Use Attitude and Learning Outcomes exhibited unidimensional structures, consistent with the scale design; the Learning Engagement dimension extracted two components, diverging from the posited four-sub-dimension structure. This result may be attributable to the intrinsic interconnections among the components of learning engagement in the context of generative AI-assisted writing—behavioral participation, social interaction, and emotional experience are highly intertwined, while cognitive processing remains relatively independent, causing the four sub-dimensions to converge empirically into two core facets: “behavioral-social participation” and “cognitive processing.” This finding reflects the context-specific nature of human-AI collaborative writing engagement and suggests that future research may employ confirmatory factor analysis to further test the dimensional structure of learning engagement.

5.3 Group Differences Analysis

To further examine the applicability of the scale across different demographic groups, this study conducted comparisons by academic year, AI usage frequency, English proficiency level, and gender. The results are presented in Table 7.

Table 7: Scale Score Differences by Demographic Characteristics (n = 282)

Grouping Variable	Category	n	AI Attitude (M±SD)	Learning Engagement (M±SD)	Learning Outcomes (M±SD)	F/t	p
Academic Year	First Year	33	4.047±0.643	4.002±0.593	4.043±0.596	0.423	0.792
	Second Year	66	4.175±0.491	4.089±0.538	4.207±0.514		
	Third Year	116	4.142±0.555	4.059±0.533	4.146±0.562		
	Fourth Year	41	4.152±0.501	4.078±0.520	4.124±0.545		
	Graduate	26	4.068±0.499	3.864±0.526	3.981±0.587		
AI Usage Frequency	Never/Rarely	21	4.111±0.819	3.937±0.835	3.996±0.872	2.557	0.056
	1-3 times/month	84	4.071±0.575	3.941±0.551	4.073±0.568		
	1-2 times/week	100	4.080±0.456	4.023±0.434	4.093±0.464		
	3+ times/week	77	4.276±0.479	4.213±0.519	4.276±0.525		
English Proficiency	No certification	37	4.129±0.687	4.085±0.705	4.137±0.690	0.080	0.971
	CET-4	106	4.153±0.551	4.032±0.566	4.148±0.588		
	CET-6	80	4.119±0.464	4.013±0.467	4.125±0.481		
	Advanced (TEM/IELTS)	59	4.119±0.510	4.081±0.467	4.099±0.506		
Gender	Male	120	4.113±0.571	4.039±0.598	4.130±0.602	-0.543	0.587
	Female	162	4.148±0.512	4.048±0.493	4.130±0.520		

Note: CET-4 = College English Test Band 4; CET-6 = College English Test Band 6; TEM = Test for English Majors; IELTS = International English Language Testing System. No post-hoc comparisons were conducted as no significant differences were found.

The results indicated no significant differences across academic year, English proficiency level, or gender on any of the three dimensions (all $p > 0.05$), demonstrating satisfactory cross-group applicability of the scale. AI usage frequency exhibited marginal significance across the three dimensions ($p = 0.056$), with high-frequency users scoring marginally higher than low-frequency users, consistent with theoretical expectations.

5.4 Comprehensive Conclusion

This study developed a psychometric instrument for assessing the effectiveness of generative AI-assisted English writing learning among Chinese university students. Through item analysis, six items with poor discriminatory power and insufficient homogeneity (Items 8, 15, 27, 35, 39, 42) were deleted. Exploratory factor analysis validated the construct validity of the scale. The final instrument comprises three dimensions—AI Use Attitude, Learning Engagement, and Learning Outcomes—with 36 items. Reliability and validity testing demonstrated Cronbach's α coefficients exceeding 0.80 for all dimensions and the overall scale, split-half reliability of 0.947, and an overall KMO value of 0.956, indicating robust psychometric properties and satisfactory practical feasibility. This study addresses the gap in assessment tools for evaluating English writing effectiveness in generative AI contexts and provides a valid instrument for subsequent empirical research.

5.5 Limitations and Future Directions

This study's sample was primarily drawn from university students across 28 provinces, with a concentration of third-year students (41.1%) and students from humanities and social sciences (40.8%) and science and engineering (31.6%). The sample exhibited some concentration in academic year and discipline distribution, and the convenience sampling method may have affected representativeness to some extent. Future research should expand to include students from different academic years, diverse institution types (e.g., vocational colleges, art institutes), and broader disciplinary backgrounds, and employ confirmatory factor analysis to further test the structural stability and cross-group applicability of the scale. Furthermore, as generative AI technology continues to evolve rapidly, the scale content should be periodically updated to maintain temporal validity.

References

[1] Yang, J. (2025). Research on the application of artificial intelligence in optimizing personalized English learning paths for university students. *Information and Computer*, 2025(5), 224-226.

[2] Zuo, X. Y., & Tan, X. Y. (2025). The effects of enjoyment, cognitive appraisal, and behavioral engagement on academic English learning effectiveness. *Foreign Language Research*, 209(1), 30-37.

[3] Chu, J. J. (2025). Research on university students' attitudes toward generative artificial intelligence-assisted academic English writing. *China Educational Technology*, 2025(8), 123-130.

[4] Parsons, S. A., Ives, S. T., Fields, R. S., & et al. (2023). The writing engagement scale: A formative assessment tool. *The Reading Teacher*, 77(3), 278-289.

[5] Wu, P. Z., Hu, X. L., & Li, B. (2025). Construction and application of an interdisciplinary thematic learning effectiveness assessment model. *China Educational Technology*, 2025(4), 66-74.

[6] Xu, C. G. (2025). Research on human-AI collaborative writing engagement of foreign language learners based on generative artificial intelligence. *Foreign Languages and Their Teaching*, 2025(2), 61-73.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights