

# Validity and reliability of the questionnaire of Chinese self-efficacy for Chinese as a second language learners in China

Yulan Peng<sup>1,2</sup>, Muhammad Zuhair Zainal<sup>1</sup>

<sup>1</sup>School of Educational Studies, Universiti Sains Malaysia, George Town, Malaysia

<sup>2</sup>College of Chinese Literature, Gannan Normal University, Ganzhou, China

## Article Info

### Article history:

Received Nov 8, 2024

Revised Mar 25, 2025

Accepted Apr 2, 2025

### Keywords:

Chinese as a second language

Confirmatory factor analysis

Reliability

Self-efficacy

Validity

## ABSTRACT

Self-efficacy is a pivotal predictor of academic success in second language learning. With the growing enrollment of international students in Chinese language programs in China, there is a pressing need for validated tools to assess self-efficacy among Chinese as a second language (CSL) learners. This study addresses this need by adapting and validating the English-Chinese questionnaire of Chinese self-efficacy (QCSE) from the original questionnaire of English self-efficacy (QESE). Using a cross-sectional design, data were collected from 174 international college students in Jiangxi Province for psychometric evaluation. Confirmatory factor analysis (CFA) assessed the instrument's reliability and validity, with results showing a high overall Cronbach's alpha of 0.942 and strong dimension-specific values (listening, speaking, reading, and writing) ranging from 0.915 to 0.927. All item outer loadings exceeded 0.7, and the average variance extracted (AVE) values were above 0.5, confirming the robustness of the QCSE. This validated instrument addresses a critical gap in CSL self-efficacy assessment and provides valuable insights for learners, educators, and policymakers.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Muhammad Zuhair Zainal

School of Educational Studies, Universiti Sains Malaysia

George Town, Penang, Malaysia

Email: zuhair@usm.my

## 1. INTRODUCTION

Chinese, regarded as the second most important international language after English [1], has gained increasing global significance with China's rising international influence and the belt and road initiative (BRI) [2]. This growth has spurred numerous countries to explore Chinese language and culture, leading to a surge in international students opting to study Chinese in China. However, the lack of self-efficacy measurement tools specifically tailored for Chinese as a second language (CSL) learners limits educators and researchers in assessing their self-confidence.

Self-efficacy, a fundamental concept in social cognitive theory, refers to an individual's belief in their capacity to successfully complete a particular task [3]. Research showed that self-efficacy significantly influences learning outcomes and academic success, particularly in second language acquisition [4], [5]. This highlights the importance of having accurate tools to measure self-efficacy in diverse language learning contexts, including CSL learners. Without a validated tool for CSL learners, it remains challenging to gauge their confidence levels and the impact on language acquisition, hindering effective instructional strategies.

Although Wang [6] questionnaire of English self-efficacy (QESE) has been validated across various languages and cultural contexts [7]–[9], a comparable tool for CSL learners remains unavailable. Therefore, this study aims to address this gap by translating and adapting the QESE into an English-Chinese version

questionnaire of Chinese self-efficacy (QCSE) and validating its reliability and validity among CSL learners to support its use across the four language skills of listening, speaking, reading, and writing. The QCSE offers a meaningful tool for advancing research on CSL learning, helping educators and policymakers better understand learners' confidence in key language skills.

In this study, 174 international students studying Chinese in China were recruited through a cross-sectional design, with data collected via a questionnaire. Confirmatory factor analysis (CFA) was applied to evaluate the scale's reliability and validity, confirming its structural consistency across listening, speaking, reading, and writing skills. The QCSE developed in this study not only provides a reliable tool for assessing self-efficacy among CSL learners but also supports educators and policymakers in optimizing instructional strategies to enhance learning outcomes. Additionally, this scale offers a foundation for future research across diverse cultural contexts, advancing cross-cultural studies on self-efficacy in language learning. Future studies can further explore the QCSE's applicability across diverse cultural backgrounds and age groups, broadening its cross-cultural utility.

## 2. METHOD

### 2.1. Sample

Bentler and Chou [10] recommended a minimum sample-to-item ratio of 5:1 for statistical analyses, with a minimum of 100 participants required for factor analysis [11]. Following this guideline, the current study determined that a minimum sample size of 160, or five times the scale's 32 items, was needed. This study involved 174 international students (82 males and 92 females) from three universities in Jiangxi Province, China. The participants' ages ranged from 18-43 years ( $M=23.52$ ,  $SD=4.223$ ) and had studied Chinese for periods varying from under 1 year to beyond 5 years (57 participants with less than 1 year, 68 participants with 1-3 years, 26 participants with 3-5 years, and 23 participants with more than 5 years). Among the participants, 65.5% were from Asian countries, 25.9% were from African countries, 4.6% were from American countries, and 4% were from European countries. All these participants' mother tongues were other than Chinese. The Chinese language is a compulsory course for all participants in these universities.

### 2.2. Sampling method

Convenience sampling was employed in this study, and participants were chosen based on their accessibility and willingness to participate. In this study, the target population was CSL learners at Chinese universities, which are large and widely distributed. It is relatively feasible for researchers to obtain data from universities geographically close to them. Although convenience sampling has some statistical limitations, especially in terms of broad generalization of the results, this method has significant advantages in terms of feasibility and time cost [12], [13].

### 2.3. Data collection procedure

The researcher contacted potential participants through Chinese language teachers and administrators at the three universities and friended them on WeChat (WeChat is the most extensively used social networking platform in China). Subsequently, the researcher introduced the study to the participants through WeChat and made it clear that they might opt out of the survey at any moment. Following this introduction, a link to the questionnaire hosted by Wenjuanxing (one of China's leading online survey platforms) was provided via WeChat. Completing the questionnaire took participants approximately five minutes, and they could do so at any time of the day.

Data collection was completed between September 4th and September 18th, 2024. A total of 305 questionnaires were administered across three universities, resulting in 185 responses. After data cleaning (which involved removing responses that were completed in less than 120 seconds or where more than 10 consecutive items received the same score), 174 valid responses were retained for analysis.

### 2.4. Instrument

The English-Chinese version of the QCSE was translated and adapted from the original QESE, which Wang [6] developed through verbal procedures, observations, and interviews involving young Chinese learners studying English in the USA. The QESE comprises 32 items, each asking students to evaluate their perceived ability to complete specific tasks in English, encompassing listening, speaking, reading, and writing, with each skill corresponding to a designated set of eight items. To mitigate the participants' awareness of the specific aspects being assessed, the items were presented in a randomized order [14]. The participants used a 7-point Likert scale to score their perceived abilities, with 1 denoting "I cannot do it at all" and 7 denoting "I can do it very well."

Based on expert opinions, scale modifications focus on small-range adjustments, such as minor adjustments to individual items or language, to improve the applicability or comprehension of the scale.

Scale adaptation, on the other hand, usually involves extensive changes aimed at adapting the scale to different cultural, linguistic, or social contexts. This may include not only language translation but also the reconstruction or replacement of items on the scale [15], [16].

In this study, to form a new English-Chinese version of the QCSE, some adaptations were applied to the original QESE. Primarily, the title QCSE was selected to capture the traits of the intended group and the specific Chinese setting more accurately. The researcher also modified the wording of all the items. As the QCSE is specifically for Chinese language learning, to make it clear and straightforward, “English” in the original questionnaire was modified to “Chinese.”

Subsequently, we substituted a 5-point Likert scale, with 1 denoting “very inconsistent” and 5 denoting “very consistent,” for the original 7-point Likert scale. Lewis [17] evaluated short-form versions of the user experience usability indicator questionnaire with various response options (3, 5, 7, and 11) and reported that versions with 5, 7, or 11 options maintain strong reliability, whereas the 3-option format shows low reliability and weak correlation with recommendation likelihood, advising against its use. Revilla *et al.* [18] examined the validity of different scales and respondents’ reaction time. The results indicated that the 5-point scale effectively reduced respondents’ reaction time and fatigue, while maintaining the validity of the questionnaire. Hartley and Betts [19] explored the effects of 5- and 7-point Likert scales on response accuracy, validity, and user preference. Their findings suggested that the 5-point scale generally provides participants with a more straightforward response process, avoiding too much ambiguity, reducing cognitive loading, and resulting in faster decision-making. In contrast, the 7-point scale can theoretically offer more precision; this additional nuance may not always be necessary but rather complicates the decision-making process of the respondent and creates “choice fatigue.”

Furthermore, the questionnaire was presented in both Chinese and English, with each English item paired with its Chinese equivalent to guarantee that participants understood each item accurately. To ensure linguistic, conceptual, operational, and metric equivalence between the English-Chinese version and the original scale, an iterative procedure involving individual translation followed by independent back-translation [20] was applied. Four bilingual language experts participated in this procedure: a pair of translators converted the items from English to Chinese, whereas another pair translated them back from Chinese to English. The researcher, along with the translators mentioned above, reviewed the English-Chinese version for sentence structure, meaning, and expression to produce the final scale.

## 2.5. Data analytical procedure

Validity can indicate whether the constructs being investigated are properly measured using different items. Two types of validity are popular in quantitative research: content and construct [21]. Reliability describes the degree of consistency or repeatability of a scale. For instruments with multiple items, internal consistency is most important; this means that the scale items should collectively measure the same underlying construct [22].

Content validity reflects how comprehensively and accurately the items within a scale represent a study’s concept or constructs. Typically, domain experts evaluate a scale to ensure that each item is reasonable, appropriate, and reflective of the primary concept under investigation [23]. In this study, content validity was assessed using expert reviews, with the results reported as a content validity index (CVI). As recommended by Lynn [24], the minimum number of experts needed to evaluate an instrument’s content validity is 3. Using a 4-point Likert scale, with 1-4 denoting “strongly disagree” to “strongly agree,” the correlation between each QCSE item and its corresponding dimension was evaluated.

Each QCSE item’s correlation with its respective dimension was evaluated using a 4-point Likert scale, with 1-4 representing “strongly disagree” to “strongly agree.” The item-level content validity index (I-CVI) and scale-level content validity index (S-CVI) were computed. The I-CVI was obtained by dividing the number of experts who rated each item with a score of 3 or 4 by the overall quantity of experts. According to Lynn [24], when the number of experts is 5 or fewer, an I-CVI of 1.00 is needed, indicating that all the experts concur that the item aligns well with the construct being measured for satisfactory content validity. The S-CVI was computed by dividing the number of items that all the experts scored as 3 or 4 by the total item count. An S-CVI of 0.8 or above suggests strong content validity for the instrument [25].

After content validity was approved, the scale was administered to the target population for data collection. IBM SPSS version 27.0 and SmartPLS version 4.0 were used for statistical analysis. SPSS 27.0 was employed for the initial data processing and descriptive analysis of the demographics of the sample. SmartPLS was employed to execute CFA on the QCSE, assessing the reliability and validity of the scale in a sample of CSL learners at Chinese universities. The partial least squares (PLS) algorithm in SmartPLS 4.0 was utilized to compute and obtain values for internal consistency reliability, indicator reliability, convergent validity, and discriminant validity, as shown in Table 1.

Table 1. Criteria for assessing the reliability and validity of the measurement model

Guidance	Acceptance level	Sources
Internal consistency	Cronbach's $\alpha \geq 0.7$ CR $\geq 0.6$	Nunnally and Bernstein [26] Hair <i>et al.</i> [27]
Indicator reliability	Standardized factor loading $> 0.7$ or $> 0.5$	Hair <i>et al.</i> [27] Hulland [28]
Convergent validity	AVE $> 0.5$	Fornell and Larcker [29] Hair <i>et al.</i> [27]
Discriminant validity	Square root of AVE $>$ interconstruct Correlation HTMT $\leq 0.85$	Fornell and Larcker [29] Kline [30]

Note: CR=composite reliability; AVE=average variance extracted; HTMT=heterotrait-monotrait ratio.

Cronbach's alpha and composite reliability (CR) values are important indicators for assessing internal consistency. Cronbach's alpha ranges from 0 to 1, with optimal values between 0.7 and 0.9 reflecting strong internal consistency [26]. CR values above 0.60 were deemed acceptable, values ranging from 0.70 to 0.90 were classified as satisfactory to good, and values surpassing 0.90 were viewed as less desirable [27].

Indicator reliability refers to the communality among indicators (items) belonging to the same construct. Hair *et al.* [27] suggested standardized factor loadings  $\geq 0.70$  or higher, with p values  $< 0.05$ , indicating statistical significance. However, standardized factor loadings greater than 0.7 are not easy to achieve in practice; thus, Hulland [28] suggested that standardized factor loadings above 0.5 are sufficient to establish indicator reliability.

The average variance extracted (AVE) is one of the principal methods used to evaluate convergent validity. The AVE represents the sum of the squared standardized factor loadings divided by the total number of indicators, which represents the average explanatory power of the indicators in the construct [31]. For adequate convergent validity, an AVE value of 0.50 or higher is recommended [29].

The definition of discriminant validity is a low correlation between indicators of one construct and those of other constructs. Discriminant validity is defined as a low correlation between indicators belonging to one construct and those belonging to other constructs [32], [33]. As an important method for assessing discriminant validity, the Fornell-Larcker criterion requires that the square root of the AVE for each construct should be greater than the correlation coefficient between that construct and the other constructs in the model [29]. To address the limitations associated with the Fornell-Larcker method, Henseler *et al.* [34] proposed the heterotrait-monotrait (HTMT) ratio criterion as a complementary criterion for discriminant validity assessment. An HTMT value less than 0.85 suggests satisfactory discriminant validity [30].

### 3. RESULTS AND DISCUSSION

#### 3.1. Content validity

Ensuring the validity of a scale begins with content validity, a crucial initial step in developing or adapting a measurement tool. The English-Chinese version of the QCSE consists of 32 items divided equally among four constructs: self-efficacy for listening, speaking, reading, and writing (8 items per construct). Five experts in linguistics, education, and psychology evaluated the scale's content validity, with a focus on item-level (I-CVI) and scale-level (S-CVI) content validity indices to confirm the comprehensiveness and accuracy of the items.

According to Lynn [24], an I-CVI value of 1.0 indicates strong validity when assessed by five experts. In this study, all 32 items in the QCSE achieved an I-CVI score of 1.0, thereby confirming their individual validity. Additionally, the S-CVI, calculated using the S-CVI/Ave method, also resulted in a perfect score of 1.0. This high content validity demonstrates a strong expert consensus, suggesting that the items comprehensively and accurately reflect self-efficacy in listening, speaking, reading, and writing skills.

Several factors may contribute to this high content validity. Each of these factors uniquely strengthened the scale's content validity. Firstly, the scale was adapted from a well-established tool (the QESE) that has already demonstrated robust reliability and validity across multiple languages and populations. This foundation likely provided a reliable starting point for adaptation [35]. Secondly, a rigorous back-translation process was employed to preserve the core structure and conceptual equivalence across linguistic and cultural contexts [20]. This approach reduces the risk of losing essential meaning, thereby enhancing content accuracy. Additionally, careful cultural adaptation, including adjustments in wording to prevent potential misunderstandings or bias due to cultural differences, likely reinforced the scale's validity [36]. Lastly, the clear, concise, and accessible wording of each item may have further facilitated high content validity, as experts found the items easy to evaluate in terms of relevance and clarity [22].

However, while this high content validity demonstrates a strong expert consensus on the scale's relevance and clarity across constructs, expert evaluations are influenced by personal expertise and contextual understanding, and this homogeneity of the expert panel may also limit the diversity of perspectives. Moreover, the sample size of five experts, though adequate by conventional standards, the generalizability of the results to a broader expert community may be limited [37]. Expanding future panels to include specialists from diverse educational and cultural backgrounds could enhance the comprehensiveness of content validity evaluations, reducing potential bias linked to localized interpretations of CSL learning competencies.

### 3.2. Reliability and validity

Factor analysis aims to interpret and summarize the relationships among numerous variables by identifying a reduced set of underlying factors. It clusters items that measure the same underlying factors into distinct groups, with each group named according to the characteristics of the items it contains. This process helps simplify the structure of the data, allowing the researcher to explain more complex phenomena with fewer factors [31]. CFA is one of the primary methods used to evaluate construct validity in scale-adaptation research. Measurement models present the relationship between a latent variable and its associated observed variables, specifying how these observed variables work together to portray the construct they represent. Measurement models can be categorized into two types: reflective and formative. In this study, both first- and second-order constructs were reflective measurement models. Indicators in a reflective model are caused by latent variables (factors); therefore, these indicators should be highly correlated and internally consistent [27]. Accordingly, reliability was evaluated using Cronbach's alpha, rho\_a, and CR. Indicator reliability was examined using standardized factor loading. Convergent validity was evaluated by the AVE, whereas the Fornell-Larcker and HTMT criteria were employed to evaluate the discriminant validity among the constructs.

#### 3.2.1. Internal consistency

Internal consistency is essential for any measurement tool designed to evaluate distinct but related constructs. Nunnally and Bernstein [26] recommended a Cronbach's alpha value above 0.7, demonstrating strong internal consistency within the measurement model. Table 2 shows the reliability results of the model.

The results demonstrated that the Cronbach's alpha for overall self-efficacy was 0.942, with subscale alphas for listening, speaking, reading, and writing at 0.927, 0.920, 0.917, and 0.915, respectively, all well above the 0.7 threshold, demonstrating high internal consistency. The rho\_a value for overall self-efficacy was 0.943, with subscale alphas for listening, speaking, reading, and writing at 0.929, 0.922, 0.919, and 0.916, respectively, again exceeding 0.7. Furthermore, Hair *et al.* [27] demonstrated that a CR value above 0.7 indicates satisfactory reliability. As shown in Table 2, the CR value for self-efficacy was 0.947, whereas the CR values for listening, speaking, reading, and writing were 0.940, 0.935, 0.932, and 0.930, respectively—all surpassing the recommended threshold. Collectively, these findings suggest high Cronbach's alpha, rho\_a, and CR values across both first- and second-order constructs, confirming strong internal consistency among the items within each construct.

These findings demonstrate robust internal consistency for the QCSE scale and its subscales, confirming the scale's reliability in assessing self-efficacy in Chinese-language learning. This is in line with prior research [7]–[9], which reported similarly high internal consistency coefficients for the QESE across Chinese, Korean, and Vietnamese contexts. This consistency across different cultural adaptations suggests that our adapted scale effectively retains the original instrument's reliability characteristics. While the high internal consistency values confirm the QCSE's reliability, it is worth noting that extremely high Cronbach's alpha values may also imply potential item redundancy within each subscale, which may affect the scale's sensitivity in capturing nuanced variations within each construct [38]. Future studies could explore the refinement of item content to ensure comprehensive coverage while minimizing redundancy, thereby enhancing the scale's precision in differentiating subtle aspects of self-efficacy across language skills.

#### 3.2.2. Indicator reliability

Indicator reliability is a critical component in evaluating a scale's measurement properties. To evaluate indicator reliability, factor loadings for the second-order construct were not calculated due to overlap with first-order construct indicators. Instead, only the standardized factor loadings of the four first-order constructs (listening, speaking, reading, and writing) were assessed [27]. Indicator reliability was determined through outer loadings, as seen in Table 2, where all item loadings ranged from 0.752 to 0.854, surpassing the 0.7 cut-off and reaching statistical significance, indicating that each indicator significantly reflects its intended construct. These high outer loadings confirm the strong reliability coefficients for the indicators within each first-order construct. Figure 1 shows the path diagram and factor loadings of the validated model.

Collectively, these findings are significant, as they indicate that each indicator contributes meaningfully to the latent construct it is intended to measure. The high outer loadings suggest that the scale items are well-aligned with the theoretical construct of self-efficacy in Chinese language learning, with each item demonstrating a strong correlation to its respective construct. This alignment supports the internal structure of the scale, ensuring that the items within listening, speaking, reading, and writing subscales are logically grouped and exhibit consistency with recommended criteria for factor analysis outcomes [22].

While these high loadings confirm robust indicator reliability for the QCSE, it is important to acknowledge potential limitations. For instance, the reliability coefficients may be sample-specific, as all participants were international students in China, which could influence item interpretation due to cultural or linguistic nuances. Future research should examine indicator reliability across varied cultural contexts, ensuring that items retain their intended meaning and relevance for CSL learners globally, thus enhancing cross-cultural robustness.

Table 2. Parameter estimates for the measurement model

Constructs	Items	Loading	Cronbach's $\alpha$	rho a	CR	AVE
SE			0.942	0.943	0.947	0.359
SEL	SE01	0.826	0.927	0.929	0.940	0.663
	SE03	0.809				
	SE09	0.854				
	SE10	0.805				
	SE15	0.821				
	SE22	0.774				
	SE24	0.789				
SES	SE27	0.832	0.920	0.922	0.935	0.642
	SE04	0.807				
	SE06	0.784				
	SE08	0.773				
	SE17	0.826				
	SE19	0.795				
	SE20	0.769				
SER	SE23	0.841	0.917	0.919	0.932	0.633
	SE30	0.813				
	SE02	0.803				
	SE12	0.757				
	SE16	0.819				
	SE21	0.794				
	SE25	0.808				
SEW	SE26	0.819	0.915	0.916	0.930	0.626
	SE29	0.788				
	SE32	0.774				
	SE05	0.820				
	SE07	0.813				
	SE11	0.752				
	SE13	0.800				
	SE14	0.796				
	SE18	0.792				
	SE28	0.766				
	SE31	0.789				

Note: SE=self-efficacy; SEL=self-efficacy for listening; SES=self-efficacy for speaking; SER=self-efficacy for reading; SEW=self-efficacy for writing

### 3.2.3. Convergent validity

Convergent validity is a crucial component of construct validity, referring to whether multiple indicators within a measurement instrument consistently measure the same underlying construct or latent variable. This ensures that items or constructs share common variances theoretically related to the construct being measured [31]. The AVE was computed to assess convergent validity. Fornell and Larcker [29] suggest that an AVE above 0.5 is acceptable, indicating that the variance explained by the construct's indicators exceeds that attributed to error. Since the second-order construct shares indicators with the first-order constructs and most variance is accounted for at the first-order level, the AVE for the second-order construct is not included in this analysis. Table 2 shows that the AVE values for all first-order constructs exceeded 0.5, ranging from 0.626 to 0.663, confirming the high convergent validity for each first-order construct.

The findings confirm that each construct effectively captures the variance among its indicators, with high AVE values supporting the scale's reliability in measuring the distinct aspects of self-efficacy in language learning. Specifically, the well-defined separation and strong validity of the constructs for listening, speaking, reading, and writing self-efficacy align with the theoretical framework, suggesting that self-efficacy in language

learning can be differentiated into specific skill areas, each with its own contributing factor. This confirms that the QCSE is a robust instrument for assessing Chinese learners’ self-efficacy across multiple dimensions.

The QCSE’s high AVE values affirm that items within each construct capture related aspects of self-efficacy, supporting the scale’s convergent validity. Nonetheless, as self-efficacy perceptions can be shaped by cultural factors, learners from different backgrounds might prioritize or interpret certain items differently. Future studies should examine convergent validity in broader cultural samples to ensure that scale’s constructs consistently resonate with diverse learner populations, thereby enhancing the scale’s generalizability.

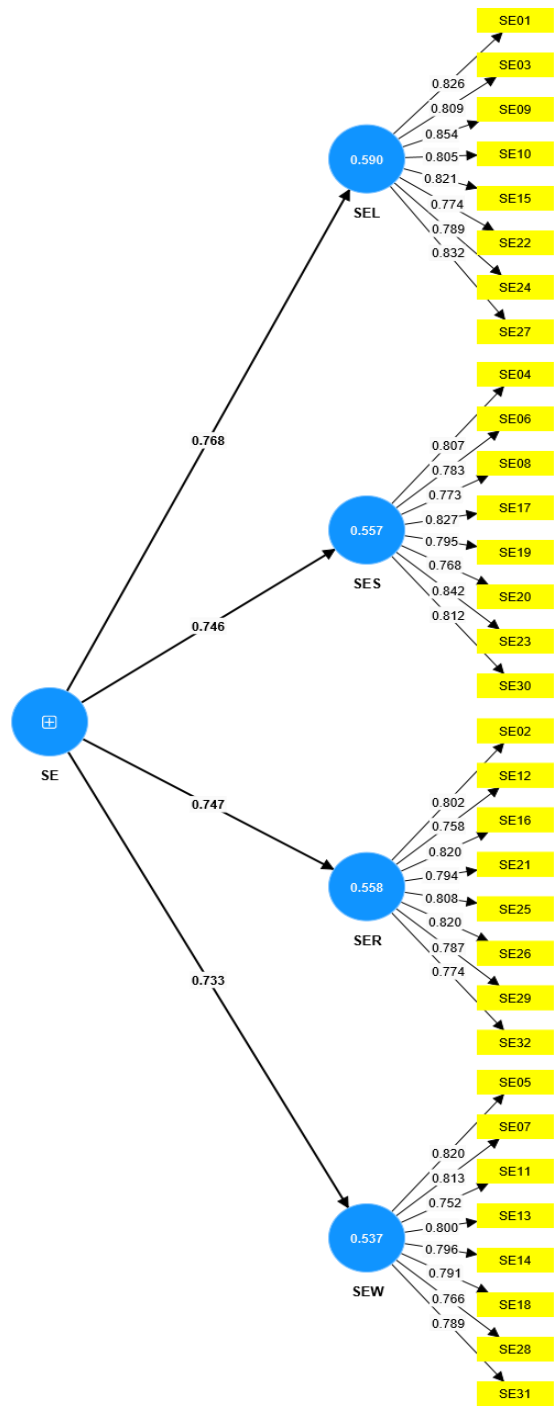


Figure 1. Path diagram and factor loadings of the Chinese self-efficacy model

### 3.2.4. Discriminant validity

Discriminant validity was assessed to determine distinctions among all latent variables, which is essential because indicators of a specific construct should show low correlation coefficients with those of other constructs to confirm distinctiveness [31]. This study initially employed the Fornell-Larcker criterion to evaluate discriminant validity by examining the square roots of the AVE values in relation to the correlation coefficients among latent variables. As the measurement model included a second-order construct (self-efficacy), the Fornell-Larcker criterion did not necessitate assessing discriminant validity between the second-order construct and each associated first-order construct [29]. Table 3 shows that the square roots of the AVE values for the four first-order constructs exceeded their inter-construct correlation coefficients, indicating sufficient discriminant validity. Discriminant validity was also evaluated by the HTMT ratio, which reflects the relationship between the correlations of different constructs (traits) and those within the same construct [34]. As presented in Table 4, the HTMT values for the constructs were all less than 0.85, satisfying the criteria for adequate discriminant validity.

Table 3. Fornell–Larcker criterion

	Self-efficacy for listening	Self-efficacy for speaking	Self-efficacy for reading	Self-efficacy for writing
Self-efficacy for listening	0.814			
Self-efficacy for speaking	0.438	0.801		
Self-efficacy for reading	0.426	0.400	0.796	
Self-efficacy for writing	0.405	0.396	0.417	0.791

Table 4. HTMT criterion

	Self-efficacy for listening	Self-efficacy for speaking	Self-efficacy for reading	Self-efficacy for writing
Self-efficacy for listening				
Self-efficacy for speaking	0.468			
Self-efficacy for reading	0.459	0.430		
Self-efficacy for writing	0.436	0.427	0.450	

These findings indicate that each first-order construct shares more variance with its own indicators than with those of other constructs, thereby confirming adequate discriminant validity. The low inter-construct correlations, confirmed by both Fornell-Larcker and HTMT criteria, suggest that the constructs for self-efficacy in listening, speaking, reading, and writing are distinct despite being part of the broader second-order construct of self-efficacy in language learning. This outcome enhances the robustness of the QCSE by ensuring that it can distinguish between specific dimensions of Chinese self-efficacy without overlap, thereby providing a clearer understanding of CSL learners' self-efficacy in each language skill area.

The QCSE's discriminant validity confirms that each construct uniquely contributes to understanding self-efficacy in CSL. However, certain language skills may naturally overlap in specific cultural or educational contexts, potentially affecting construct distinctiveness. Investigating the scale's discriminant validity across varied instructional settings and cultural backgrounds can help refine its ability to differentiate between skill-based self-efficacy dimensions, particularly in environments where integrated language skills are emphasized.

In conclusion, this study provides compelling evidence that the QCSE is a reliable and valid instrument for assessing self-efficacy among CSL learners in Chinese universities. Consistent with prior research [39], [40], the findings confirm the scale's robustness, identifying a second-order common factor, self-efficacy, and four distinct dimensions: self-efficacy for listening, speaking, reading, and writing. This multidimensional structure aligns with Zimmerman and Cleary [41] assertion that self-efficacy measures should reflect the specific tasks within a domain to capture the full scope of learners' abilities. In the context of second language acquisition, self-efficacy is indeed a multifaceted construct. Interestingly, while this study validated a multidimensional structure, prior research on the QESE scale has sometimes identified a unidimensional structure when assessing ESL learners, potentially due to sample characteristics such as limited sampling methods or unbalanced gender distribution [7]–[9]. This contrast underscores the importance of sample diversity and contextual factors in shaping the measurement of self-efficacy beliefs. Future studies could explore these dimensions across varied cultural and linguistic backgrounds to further confirm the QCSE's generalizability and adaptability, enhancing its utility as a comprehensive tool for CSL educators and researchers in diverse settings.



#### 4. CONCLUSION

This study establishes the QCSE as a reliable, domain-specific instrument for CSL learners, addressing a significant gap in research on learning CSL. By encompassing the key domains of listening, speaking, reading, and writing, the QCSE's multidimensional structure offers educators a clearer picture of learners' specific strengths and areas for growth. The scale's practical implications are substantial: it allows educators to conduct targeted self-efficacy assessments, supporting the development of customized instructional methods that enhance learners' confidence in each skill area. It also provides policymakers with data-driven insights for creating CSL frameworks that foster self-efficacy and improve language learning outcomes. However, this study is a cross-sectional design with a predominantly Asian and African sample. Future research could extend the QCSE's applicability by testing it across diverse cultural contexts, and exploring the longitudinal impact of self-efficacy on language outcomes, ultimately contributing to more effective and sustainable strategies in CSL education globally.

#### ACKNOWLEDGEMENTS

The authors express their sincere appreciation to all those who provided support for this research.

#### FUNDING INFORMATION

Authors state no funding involved.

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yulan Peng	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Muhammad Zuhair Zainal	✓	✓					✓	✓	✓	✓		✓	✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

#### CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

#### DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author [MZZ], upon reasonable request.




#### REFERENCES

- [1] N. L. Murtadhoh and W. Arini, "The Existence of Chinese Language in The Globalization Era," *Journal of Maobi*, vol. 1, no. 1, p. 7, Nov. 2023, doi: 10.20961/maobi.v1i1.79731.
- [2] Y. Gao, "How the Belt and Road Initiative Informs Language Planning Policies in China and among the Countries along the Road," *Sustainability*, vol. 12, no. 14, p. 5506, Jul. 2020, doi: 10.3390/su12145506.
- [3] A. Bandura, *Self-efficacy: the exercise of control*. New York, USA: Freeman, 1997.
- [4] S. H. Kim and H. W. Shin, "Second Language Learners' Self-efficacy and English Achievement: the Mediating Role of Integrative Motivation," *English Teaching & Learning*, vol. 45, no. 3, pp. 325–338, Oct. 2021, doi: 10.1007/s42321-021-00083-5.
- [5] J. Goetze and M. Driver, "Is learning really just believing? A meta-analysis of self-efficacy and achievement in SLA," *Studies in Second Language Learning and Teaching*, vol. 12, no. 2, pp. 233–259, Jun. 2022, doi: 10.14746/sslt.2022.12.2.4.
- [6] C. Wang, *Self-Regulated Learning Strategies and Self-Efficacy Beliefs of Children Learning English as a Second Language*. Columbus, USA: The Ohio State University, 2004.
- [7] D.-H. Kim, C. Wang, and T. N. N. Truong, "Psychometric properties of a self-efficacy scale for English language learners in Vietnam," *Language Teaching Research*, vol. 28, no. 4, pp. 1412–1427, Jul. 2024, doi: 10.1177/13621688211027852.
- [8] C. Wang, D.-H. Kim, R. Bai, and J. Hu, "Psychometric properties of a self-efficacy scale for English language learners in China," *System*, vol. 44, pp. 24–33, Jun. 2014, doi: 10.1016/j.system.2014.01.015.




- [9] C. Wang, D.-H. Kim, M. Bong, and H. S. Ahn, "Examining measurement properties of an English Self-Efficacy scale for English language learners in Korea," *International Journal of Educational Research*, vol. 59, pp. 24–34, 2013, doi: 10.1016/j.ijer.2013.02.004.
- [10] P. M. Bentler and C.-C. Chou, "Practical Issues in Structural Modeling," *Sociological Methods & Research*, vol. 16, no. 1, pp. 78–117, 1987, doi: 10.1177/0049124187016001004
- [11] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong, "Sample size in factor analysis," *Psychological Methods*, vol. 4, no. 1, pp. 84–99, Mar. 1999, doi: 10.1037/1082-989X.4.1.84.
- [12] I. Etikan, "Comparison of Convenience Sampling and Purposive Sampling," *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 1–4, 2016, doi: 10.11648/j.ajtas.20160501.11.
- [13] M. H. Bornstein, J. Jager, and D. L. Putnick, "Sampling in developmental science: Situations, shortcomings, solutions, and standards," *Developmental Review*, vol. 33, no. 4, pp. 357–370, Dec. 2013, doi: 10.1016/j.dr.2013.08.003.
- [14] R. Tourangeau, L. J. Rips, and K. Rasinski, *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press, 2000, doi: 10.1017/cbo9780511819322.
- [15] D. E. Beaton, C. Bombardier, F. Guillemin, and M. B. Ferraz, "Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures," *Spine*, vol. 25, no. 24, pp. 3186–3191, 2000.
- [16] J. Epstein, R. M. Santo, and F. Guillemin, "A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus," *Journal of Clinical Epidemiology*, vol. 68, no. 4, pp. 435–441, Apr. 2015, doi: 10.1016/j.jclinepi.2014.11.021.
- [17] J. R. Lewis, "Measuring User Experience With 3, 5, 7, or 11 Points," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 63, no. 6, pp. 999–1011, Sep. 2021, doi: 10.1177/0018720819881312.
- [18] M. A. Revilla, W. E. Saris, and J. A. Krosnick, "Choosing the Number of Categories in Agree–Disagree Scales," *Sociological Methods & Research*, vol. 43, no. 1, pp. 73–97, Feb. 2014, doi: 10.1177/0049124113509605.
- [19] J. Hartley and L. R. Betts, "Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on Likert-type scales," *International Journal of Social Research Methodology*, vol. 13, no. 1, pp. 17–27, Feb. 2010, doi: 10.1080/13645570802648077.
- [20] R. W. Brislin, "Back-Translation for Cross-Cultural Research," *Journal of Cross-Cultural Psychology*, vol. 1, no. 3, pp. 185–216, Sep. 1970, doi: 10.1177/135910457000100301.
- [21] R. Adcock and D. Collier, "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research," *American Political Science Review*, vol. 95, no. 3, pp. 529–546, Sep. 2001, doi: 10.1017/S0003055401003100.
- [22] R. F. DeVellis, *Scale Development: Theory and Applications*, 4th ed. Thousand Oaks, CA: SAGE Publications, Inc., 2017.
- [23] D. F. Polit and C. T. Beck, "The content validity index: Are you sure you know what's being reported? critique and recommendations," *Research in Nursing & Health*, vol. 29, no. 5, pp. 489–497, Oct. 2006, doi: 10.1002/nur.20147.
- [24] M. R. Lynn, "Determination and Quantification of Content Validity," *Nursing Research*, vol. 35, no. 6, pp. 382–386, Nov. 1986, doi: 10.1097/00006199-198611000-00017.
- [25] L. L. Davis, "Instrument review: Getting the most from a panel of experts," *Applied Nursing Research*, vol. 5, no. 4, pp. 194–197, Nov. 1992, doi: 10.1016/S0897-1897(05)80008-4.
- [26] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd ed. New York: McGraw-Hill, 1994.
- [27] J. F. Hair Jr., G. T. M. Hult, C. Ringle, and M. Sarstedt, *A primer on partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks, CA: SAGE Publications, Inc., 2014.
- [28] J. Hulland, "Use of partial least squares (PLS) in strategic management research: a review of four recent studies," *Strategic Management Journal*, vol. 20, no. 2, p. 195, 1999, doi: 10.1002/(SICI)1097-0266(199902)20:2<195::AID-SMJ13>3.0.CO;2-7.
- [29] C. Fornell and D. F. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, Feb. 1981, doi: 10.1177/002224378101800104.
- [30] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 5th ed. New York: The Guilford Press, 2023.
- [31] J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 7th ed. London, UK: Pearson, 2014, doi: 10.1002/9781119409137.ch4.
- [32] G. A. Churchill, "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, vol. 16, no. 1, pp. 64–73, Feb. 1979, doi: 10.1177/002224377901600110.
- [33] J. C. Anderson and D. W. Gerbing, "Structural equation modeling in practice: A review and recommended two-step approach," *Psychological Bulletin*, vol. 103, no. 3, pp. 411–423, May 1988, doi: 10.1037/0033-2909.103.3.411.
- [34] J. Henseler, C. M. Ringle, and M. Sarstedt, "A New Criterion for Assessing Discriminant Validity in Variance-based Structural Equation Modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, Jan. 2015, doi: 10.1007/s11747-014-0403-8.
- [35] G. Efsthathiou, "Translation, Adaptation and Validation Process of Research Instruments," in *Individualized Care: Theory, Measurement, Research and Practice*, R. Suhonen, M. Stolt, and E. Papastavrou, Eds., Cham: Springer, 2019, pp. 65–78, doi: 10.1007/978-3-319-89899-5\_7.
- [36] R. K. Hambleton and L. Patsula, "Adapting tests for use in multiple languages and cultures," *Social Indicators Research*, vol. 45, no. 1–3, pp. 153–171, 1998, doi: 10.1023/a:1006941729637.
- [37] C. P. Friedman, J. C. Wyatt, and J. S. Ash, "Measurement Fundamentals: Reliability and Validity," in *Evaluation Methods in Biomedical and Health Informatics*, C. P. Friedman, J. C. Wyatt, J. S. Ash, Eds., Cham: Springer, 2022, pp. 129–154, doi: 10.1007/978-3-030-86453-8\_7.
- [38] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education*, vol. 48, no. 6, pp. 1273–1296, Dec. 2018, doi: 10.1007/s11165-016-9602-2.
- [39] C. Wang and B. Bai, "Validating the Instruments to Measure ESL/EFL Learners' Self-Efficacy Beliefs and Self-Regulated Learning Strategies," *TESOL Quarterly*, vol. 51, no. 4, pp. 931–947, Dec. 2017, doi: 10.1002/tesq.355.
- [40] C. Wang, G. Schwab, P. Fenn, and M. Chang, "Self-Efficacy and Self-Regulated Learning Strategies for English Language Learners: Comparison between Chinese and German College Students," *Journal of Educational and Developmental Psychology*, vol. 3, no. 1, pp. 173–191, Apr. 2013, doi: 10.5539/jedp.v3n1p173.
- [41] B. J. Zimmerman and T. J. Cleary, "Adolescents' Development of Personal Agency: The Role of Self-Efficacy Beliefs and Self-Regulatory Skill," in *Self-Efficacy Beliefs of Adolescents*, F. Pajares and T. Urdan, Eds., Greenwich, CT: Information Age Publishing, 1998, pp. 45–69.

---

**BIOGRAPHIES OF AUTHORS**

**Yulan Peng**    is a Ph.D. candidate, School of Educational Studies, Universiti Sains Malaysia, Malaysia and associate professor at College of Chinese Literature, Gannan Normal University, China. Her research focuses on language education and measurement, and international Chinese language education. She can be contacted at email: pengyulan6526@student.usm.my; 13766333844@163.com.



**Muhammad Zuhair Zainal**    is a senior lecturer at the School of Educational Studies, Universiti Sains Malaysia, Malaysia. His research interests include curriculum, pedagogy and teacher development, particularly in the field of language education. He is also interested in studying the phenomenon of language use in education as he has expertise in the field of pragmatics. He can be contacted at email: zuhair@usm.my.