Eurasian Journal of Educational Research 117 (2025) 45-63

Eurasian Journal of Educational Research
www.ejer.com.tr

Unlocking Minds: Building and Validating Metacognition Skills Inventory for Elementary Students

Chang Chen[1]*, Li-Jung[2]

ARTICLE INFO

Article History:
Received: 01 December 2024
Received in revised form: 18 January 2025
Accepted: 30 March 2025
DOI: 10.14689/ejer.2025.117.03

Keywords
Metacognitive Skills, Confirmatory Factor Analysis, Structural Equation Model

ABSTRACT

Purpose: This study aimed to revise and validate the State Metacognitive Inventory for application among upper primary school students in Zhejiang Province, China. The objective was to develop an instrument that is both culturally appropriate and developmentally suitable for assessing children's metacognitive skills. **Design/Methodology/Approach:** The original inventory was translated and adapted through expert evaluation and pilot testing. Following linguistic and contextual adjustments, the revised 21-item scale was administered to 231 students in Years 5 and 6 across five schools representing diverse regions within Zhejiang. Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), and Structural Equation Modelling (SEM) were employed to assess the scale's reliability and validity. **Findings:** The revised scale demonstrated high internal consistency (Cronbach's α = .944) and effectively distinguished between varying levels of metacognitive skills. EFA identified a three-factor structure—Cautiousness, Confidence, and Introspection—replacing the original four-factor model. Although the new model showed superior overall model fit (e.g., RMSEA = .044, CFI = .965), challenges remained in terms of convergent and discriminant validity, particularly within the Introspection dimension. These limitations may be attributable to the abstract nature of metacognitive constructs and the developmental stage of the target population. **Implications:** Despite issues related to validity, the revised scale functions as a reliable tool for the rapid assessment of metacognitive skills among Chinese primary school students. It lays the groundwork for future theoretical refinement and supports the localisation of cross-cultural measurement instruments. **Originality/Value:** This study offers a culturally grounded framework for evaluating metacognitive development in Chinese children and provides empirical direction for enhancing the psychometric quality of metacognitive inventories tailored to younger learners.

Introduction

Metacognition, originally conceptualised by Flavell (1979), refers to an individual's capacity to monitor, evaluate, and regulate their own cognitive operations. As scholarly

[1] Department of Education, International College, Krirk University, Krirk University, Bangkok, Thailand, 10220
ORCID: https://orcid.org/0009-0009-6671-9134, Email: ccaiyo0813@outlook.com
[2] Department of Education, International College, Krirk University, Krirk University, Bangkok, Thailand, 10220
ORCID: https://orcid.org/0009-0009-0762-8232, Email: lijungyu@gmail.com
*Correspondence: ccaiyo0813@outlook.com

inquiry into this domain has advanced, considerable attention has been directed towards the refinement and validation of instruments for measuring metacognitive functions. A substantial body of prior research has underscored the pivotal influence of metacognitive competencies on everyday functioning and academic engagement. Within educational contexts, Schuster et al. (2020) identified a strong association between metacognitive proficiency and enhanced academic outcomes. Empirical studies have further demonstrated that metacognitive skills play a vital role in facilitating children's assimilation of novel ideas and the restructuring of existing conceptual frameworks Smortchkova and Shea (2020), improving their problem-solving efficacy Güner and Erbay (2021), fostering autonomous decision-making Moses-Payne et al. (2021), influencing the trajectory of academic development across multiple disciplines (Tibken et al., 2021), and reinforcing effective learning behaviours (An et al., 2024). Beyond formal educational settings, metacognitive abilities contribute meaningfully to the cultivation of psychological resources, the enhancement of interpersonal effectiveness Li et al. (2024), the regulation of emotional responses Kahan and Sullivan (2012), and the development of empathy via socially embedded metacognitive processes (Zawidzki, 2019).

In light of this, educators have consistently placed emphasis on strengthening students' metacognitive capacities. In their attempt to bolster these skills among secondary school students, Bae and Kwon (2019) implemented targeted interventions, although the precise efficacy of such measures in elevating metacognitive functioning remains inconclusive. Therefore, designing reliable metacognitive assessment tools has been one of the directions that researchers have been striving for all along. Azevedo (2020), in his review of the metacognition field, proposed several future research directions and recommendations, including investigations into the differences in metacognitive measurement across various age groups. This indicates the necessity of developing more age-appropriate metacognitive assessment instruments tailored to different populations. Recent research within the Chinese educational landscape has largely concentrated on strategies to support learners in deploying metacognitive tools across subjects and learning environments. However, there remains a conspicuous deficit in theoretical inquiry into metacognition's structural formulation and developmental pathways. Notably, there is limited empirical work examining metacognitive proficiency among younger populations or evaluating the instruments designed to assess such skills. This investigation seeks to address this identified gap.

O'Neil and Abedi (1996), incorporating constructs such as cognitive self-awareness and strategic regulation, expanded the conceptual boundaries of metacognitive skills. Drawing inspiration from Spielberger (1966) state-trait anxiety theory, they proposed a dual-dimensional framework of metacognition, comprising state and trait aspects. To operationalise this framework, they developed the "State Metacognitive Scale," designed to measure American adolescents' metacognitive performance. The instrument contains 20 items distributed across four dimensions: Awareness, Cognitive Strategy, Planning, and Self-checking, each encompassing five items. It utilises a 4-point Likert scale, ranging from "Not at all" (1) to "Very much so" (4), with higher cumulative scores signifying greater metacognitive capability. The scale has demonstrated strong psychometric properties in samples of twelfth-grade students, with Cronbach's α coefficients above .70 for all dimensions (.79 for Awareness, .81 for Cognitive Strategy, .83 for Planning, and .75 for Self-checking). Reliability remained robust in alternative cohorts, such as community college

students, with all α values surpassing .70. However, preliminary application among younger learners, including those in eighth grade or below, revealed the necessity for scale modification to ensure developmental appropriateness.

Presently, available instruments for assessing metacognition include not only Spielberger (1966) original scale but also various tools adapted for specific populations. These include the Metacognitive Skills Scale (MSS) by Altındağ and Senemoğlu (2013) for tertiary-level students, the Turkish Metacognitive Inventory developed by Çetinkaya and Erktin (2002) for sixth-grade pupils, a revised version of the State Measure of Metacognition proposed by Immekus and Imbrie (2008) for college use, and a Spanish-language adaptation created by Saldarriaga et al. (2012). There are relatively few children's metacognitive awareness scales translated into Chinese, and most of them are adaptations or validations based on instruments such as the Junior Metacognitive Awareness Inventory (Jr. MAI) and the MCQ-30 (a simplified version of the Metacognition Questionnaire). For example, Ning (2017) examined the validity and reliability of the Jr. MAI when used with Asian children, while Li et al. (2023) assessed the reliability and validity of the MCQ-30 among Chinese adolescents aged 11 to 18. However, to date, there are few metacognitive skill scales developed specifically from the perspective of state metacognition.

Research Method

Materials and Methods

The researcher-initiated contact via email with one of the original authors of the scale, O'Neil, to secure formal permission for its use. Following this, a panel comprising subject-matter experts and academic professors was convened to undertake the translation of the original instrument into Chinese.

Translation

Two English language specialists were invited by the researcher to independently translate the original scale, resulting in two separate Chinese versions. These translations were subsequently reviewed, compared, and synthesised by an expert panel, leading to the development of a consolidated final Chinese version of the scale.

Item Modification

The researcher enlisted six experts and professors from diverse universities, specialising in education and psychology, to review the scale. Each item in the Chinese version was carefully modified to align with the linguistic conventions and cultural context relevant to Chinese elementary school students.

Pilot Testing

A pilot study was conducted using a randomly selected sample of 44 students from fourth to sixth grade. During this stage, each item on the scale was supplemented with the response option "I do not understand this question" to evaluate item clarity and assess whether the scale was appropriate for the targeted age group. Feedback obtained during this phase revealed that certain fourth-grade pupils experienced difficulty comprehending the metacognitive items, whereas fifth- and sixth-grade participants generally reported

understanding the item content. In light of these findings, the researcher refined the participant pool by limiting the study to fifth and sixth graders. The final pilot sample comprised 33 students, including 15 fifth graders (45.45%) and 18 sixth graders (54.55%), of whom 24 were boys (72.73%) and 9 were girls (27.27%).

Internal consistency and reliability analyses were performed on the pilot data (refer to Table 1). Items that appeared ambiguous or susceptible to multiple interpretations were revised accordingly. The overall reliability coefficient of the scale during the pilot phase was .879, while the Kaiser-Meyer-Olkin (KMO) value stood at .465. The suitability of data for factor analysis can be assessed using the KMO value. According to Shrestha (2021), a KMO value of at least 0.6 is required to proceed with subsequent factor analysis. The deletion of items 10 and 17 led to an increase in Cronbach's α beyond .879, suggesting that these items required revision. The Cronbach's α values for the four dimensions were as follows: Awareness (.693), Cognitive Strategy (.640), Planning (.809), and Self-checking (.477). Following the removal of items 3, 8, 6, 10, and 17, internal consistency across the dimensions improved, warranting continued refinement. Additionally, the researcher observed that item 14 incorporated two distinct questions within a single item, potentially causing confusion among participants. To address this, item 14 was divided into two separate statements, increasing the total number of items from 20 to 21. Detailed modifications to each item are presented in Table 2.

Table 1

Pilot Test Result

Dimension	Items	Cronbach's α if Item Deleted	Dimension's α	Variable's α	KMO
Awareness	M1	.662	.693	.879	.465
	M5	.609			
	M9	.589			
	M13	.640			
	M17	.696			
Cognitive Strategy	M3	.641	.640		
	M7	.613			
	M11	.538			
	M15	.561			
	M19	.570			
Planning	M4	.765	.809		
	M8	.845			
	M12	.742			
	M16	.720			
	M20	.762			
Self-Checking	M2	.419	.477		
	M6	.570			
	M10	.519			
	M14	.265			
	M18	.324			

Table 2

Item Adjustment Comparison (Translated from Chinese)

Before Item Adjustment	After Item Adjustment
3. I will try my best to find the key point to solve a problem.	3. I will try to find out how to solve the problem.
8. I tried to determine what this activity entailed.	8. I'll try to figure out the requirements for this activity.
6. In the process of completing the activity tasks, I corrected some inappropriate methods.	6. During the activity, I will change my approach depending on the situation.
10. I almost always know roughly how long it will take me to complete an activity.	10. I know exactly how much time I have left to complete this activity.
14. I will regularly monitor the progress of my activities and tasks and change my methods or strategies when necessary.	14. I always keep an eye on the progress of my tasks. 15. When necessary, I will change the way I do things.
17. I can see myself trying to understand a problem first before solving it.	18. When I need to solve a problem, I first understand what the problem means.

Note: Item 14 has been split into two separate items, numbered 14 and 15. Consequently, the numbering of subsequent items has been incremented by one.

Research Subjects

The participants in this investigation comprised fifth- and sixth-grade pupils enrolled in primary schools across Zhejiang Province, China, with ages ranging from 10 to 12 years, as determined by the standard age requirements for school admission. Zhejiang Province represents one of China's economically advanced regions and is recognised for its leadership in the advancement of educational modernisation. The population in this province typically benefits from a high standard of living and demonstrates a strong commitment to their children's academic development, rendering it a highly suitable setting for conducting educational research.

From a geographical perspective, Zhejiang is categorised into four principal regions: East, South, West, and North Zhejiang. Among these, the northern and southern regions possess comparatively denser populations, whereas the western region is characterised by the lowest population density. As reported in the "2022 Statistical Bulletin on the Development of Education in Zhejiang Province," the province accommodates a total of 3,204 primary schools. For the purpose of this study, the researcher secured participation from five schools that expressed willingness to be involved, including one school located in East Zhejiang, two in South Zhejiang, one in West Zhejiang, and one in North Zhejiang. The detailed distribution of the sample across these schools is provided in Table 3.

Table 3

Samples for Formal Testing (n=231)

Area	Sample	%
1. Northern Zhejiang	26	11.26
2. Western Zhejiang	24	10.39
3. Eastern Zhejiang	33	14.29
4. Southern Zhejiang-1	88	38.10
5. Southern Zhejiang-2	60	25.97
Total	231	100

Background Variables

This research adopted the Family Affluence Scale (FAS), formulated by Currie et al. (2008), as a reference framework for evaluating the socioeconomic status of participants' families. The FAS is designed to capture indicators of current household affluence, offering two key advantages that underpin its selection in this context. Firstly, the FAS demonstrates strong adaptability for cross-national studies, requiring only minimal contextual adjustments to reflect the economic realities of different countries. Secondly, conventional approaches to assessing socioeconomic status typically demand detailed information regarding parental income, occupation, and marital status. Such data can be difficult for primary school pupils to provide accurately and may pose unnecessary cognitive or emotional burden. In contrast, the FAS streamlines this process by enabling students to simply report the presence or absence of certain household items considered indicative of material well-being.

Administration Method and Process

This study employed a convenience sampling approach for questionnaire distribution. Following initial contact with school principals and securing approval from school administrations, the researcher disseminated the questionnaire to parents' mobile phones via an online platform. Students completed the questionnaire using their parents' devices after engaging in extracurricular activities. Participation was entirely voluntary, and prior notification was provided to parents through the respective schools. As students were required to partake in extracurricular activities before completing the questionnaire, an introductory note was incorporated at the beginning of the instrument, prompting students to reflect on their emotional and cognitive experiences during these activities while responding. The researcher is currently undertaking further investigation into the specific categories of extracurricular activities involved, as well as the nature of metacognitive skills demonstrated in these contexts.

Data Analysis

In line with the recommendations of Brown and Moore (2012), EFA is considered appropriate when the theoretical basis is insufficient to support a predetermined factor structure. In contrast, when a solid theoretical foundation exists, SEM is more suitable for performing CFA. Additionally, following the results of previous research, methods for testing construct validity often employ factor analysis, cause there are strong association between them, which allows for the assessment of the scale's measurement validity based on the extracted common factors (Tavakol & Wetzel, 2020). Accordingly, the data analysis procedures for this study were implemented through the following sequential steps:

1. Item analysis was performed using SPSS 23.0 to determine the suitability of individual items;
2. EFA was applied to extract latent factors and refine the theoretical framework;
3. The revised model underwent reliability and validity testing;
4. The fit between the sample data and the revised model was examined, and a comparative analysis was carried out between the original and the reconstructed models.

Formal Test

The formal data collection yielded a total of 231 valid responses, consisting of 178 fifth-grade pupils (77.06%) and 53 sixth-grade pupils (22.94%). The gender distribution was relatively balanced, with 118 male students (51.08%) and 113 female students (48.92%). Participants' ages fell within the range of 10 to 12 years. Regarding socioeconomic status (SES), 54 students (23.38%) were identified as belonging to low SES households, 170 students (73.59%) were from middle SES backgrounds, and 7 students (3.03%) represented high SES families.

Item Analysis

The objective of conducting item analysis was to assess the degree to which each question functioned consistently in relation to both other items and the overall measurement scale. As outlined in Table 4, the t-values and Pearson correlation coefficients for all items attained statistical significance, with each correlation coefficient surpassing the threshold of .40. These findings demonstrate that the items are capable of effectively distinguishing among different levels of metacognitive skills, while also exhibiting coherence with the construct measured by the full scale. This consistency justifies the progression to subsequent factor analysis.

Table 4

Summary of Item Analysis (n=231)

Item	EGC	Item-Total Correlation	Item	EGC	Item-Total Correlation
	CR	Pearson's r		CR	Pearson's r
M1	6.842***	.503**	M11	12.707***	.694**
M2	10.582***	.645**	M12	13.046***	.761**
M3	12.653***	.691**	M13	10.876***	.632**
M4	13.696***	.710**	M14	9.990***	.653**
M5	11.235***	.701**	M15	12.097***	.699**
M6	12.548***	.722**	M16	11.802***	.722**
M7	11.691***	.657**	M17	14.876***	.760**
M8	10.356***	.694**	M18	14.017***	.737**
M9	13.046***	.664**	M19	12.071***	.699**
M10	11.391***	.653**	M20	14.612***	.737**
			M21	11.545***	.690**

Note: EGC: Extreme Groups Comparison; CR: Critical Ratio.

Exploratory Factor Analysis

Principal component analysis was utilised to conduct exploratory factor analysis. The Kaiser-Meyer-Olkin (KMO) measure was .959***, indicating a high degree of common variance among the observed variables and confirming the suitability of the data for factor analysis. Guided by the theoretical structure of the original instrument, four factors were extracted through forced extraction. As shown in Table 5, the first factor consists of eight items, the second includes seven, and the third and fourth each comprise three items. The scree plot displayed in Figure 1 reveals a distinct inflection point beginning at the fourth factor, indicating the appropriateness of a four-factor model. Collectively, these four factors account for 61.511% of the total variance, reflecting a robust level of explanatory adequacy. All communalities exceed .50, indicating strong correlations between the items and their

underlying factors, thereby justifying the retention of every item. The internal consistency coefficients (Cronbach's α) for the respective factors are as follows: .896 for Factor 1, .870 for Factor 2, .725 for Factor 3, and .698 for Factor 4.

Table 5

Summary of Factor Analysis (n=231)

Item	% of Variance	Cumulative %	Fac 1	Fac 2	Fac 3	Fac 4	Extraction
M19	20.046	20.046	**.732**	.182	.097	.299	.668
M18			**.691**	.259	.258	.195	.649
M21			**.665**	.196	.258	.192	.584
M20			**.630**	.363	.243	.147	.609
M16			**.610**	.240	.417	.137	.622
M17			**.582**	.377	.159	.346	.625
M11			**.530**	.496	-.030	.279	.605
M15			**.469**	.262	.221	.452	.541
M9	17.643	37.689	.330	**.721**	.061	.090	.640
M5			.249	**.680**	.333	.098	.645
M10			.211	**.612**	.020	.423	.599
M7			.121	**.583**	.455	.179	.594
M6			.297	**.535**	.233	.370	.567
M12			.352	**.515**	.213	.444	.631
M4			.311	**.514**	.421	.162	.565
M1	12.575	50.264	.126	-.001	**.811**	.222	.722
M2			.276	.293	**.604**	.176	.558
M3			.457	.332	**.571**	.001	.646
M13	11.247	61.511	.210	.189	.224	**.749**	.692
M14			.471	.143	.136	**.569**	.584
M8			.205	.394	.418	**.448**	.572
Total			4.210	3.705	2.641	2.362	

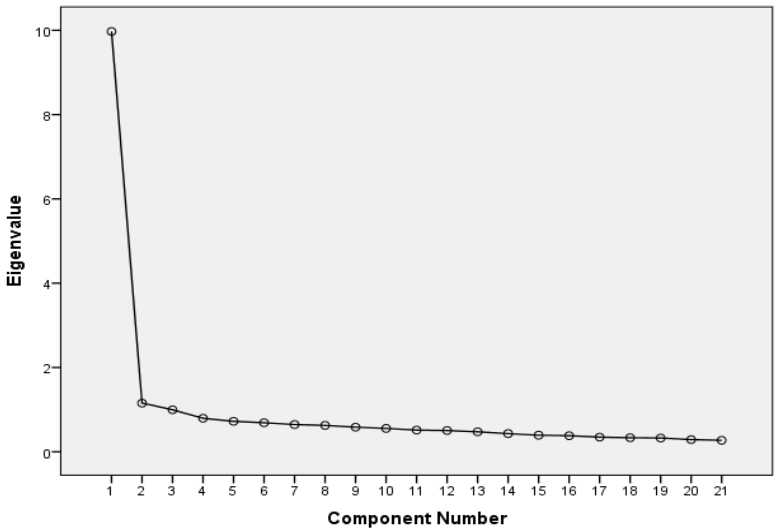

Figure 1: Scree Plot (n=231)

Factor Merging, Naming, Reliability and Validity Testing of the Model

This study aimed to assess and compare the efficacy of the original theoretical model against a revised version, using data obtained from elementary school students in Zhejiang Province, China. The overall reliability coefficient for the 21-item instrument was found to be .944, indicating a robust level of internal consistency. Tables 6 and 7 provide a comparative overview of the reliability and validity analyses for both models. Specifically, Table 6 outlines the reliability and validity findings based on the original framework, while Table 7 presents the corresponding analyses for the revised model, which further explores two alternative configurations—one of which involves combining the third and fourth factors.

A comparison of the results demonstrates that the revised model exhibits higher standardised factor loadings than the original, suggesting an improved alignment between items and their designated constructs. Additionally, the Cronbach's α values for the first and second dimensions in the revised framework surpass those of any dimension within the original model. However, the third and fourth factors in the revised model yield Cronbach's α values of .725 and .698, respectively, which fall below the reliability coefficients observed for any factor in the original framework. As advised by Nunnally (1978), a Cronbach's α of at least .7 is the recommended threshold for questionnaire revision, and increasing the number of items per factor is an effective strategy for enhancing reliability. In light of this, the present study has opted to combine the third and fourth factors, resulting in an improved reliability coefficient of .804.

Table 6

Reliability Analysis of the Original Theory

Dimension	Item	Factor Loading	Cronbach's α if Item Deleted	Dimension's α
Awareness	M1	.464	.734	.737
	M5	.648	.661	
	M9	.620	.699	
	M13	.580	.689	
	M18	.705	.665	
Cognitive Strategies	M4	.689	.796	.829
	M8	.668	.802	
	M12	.746	.781	
	M17	.747	.784	
	M21	.671	.812	
Planning	M3	.668	.768	.810
	M7	.630	.787	
	M11	.676	.787	
	M16	.704	.762	
	M20	.718	.760	
Self-Checking	M2	.622	.809	.821
	M6	.705	.781	
	M10	.628	.800	
	M14	.634	.790	
	M15	.685	.788	
	M19	.688	.789	

Following the classification of the 21 items into three distinct factors, the researcher examined the thematic content of each group to inform the naming of the factors. The items under Factor One primarily depict a behavioural approach marked by "repetitive thinking, thorough preparation, and timely adjustments," and therefore this factor was designated as "Cautiousness." The items within Factor Two reflect a tendency towards "clarity of goals and confidence in methods," justifying its label as "Confidence." Upon merging Factor Three and Factor Four, the unified thematic representation centres on a style characterised by "internal reflection and maintaining oversight," and this composite factor was subsequently named "Introspection." For clarity, the researcher referred to the original structure as Model I, while the revised configuration, resulting from the integration of Factors Three and Four, was termed Model II.

In accordance with the guidelines proposed by Hair et al. (2010), it is recommended that all standardised factor loadings attain a minimum value of .50, with an ideal threshold of .70 or above. As shown in Table 7, under the four-factor configuration, the standardised factor loadings for the 21 items range between .553 (M1) and .772 (M17 and M3). Within the three-factor structure, these loadings range from .513 (M1) to .771 (M17), with eight items reaching the .70 benchmark. However, following the amalgamation of Factor Three and Factor Four, a reduction in standardised factor loadings was observed for 11 items, which may have implications for subsequent validity assessments.

Table 7

Reliability Analysis of the New Theory

Dimension	Item	Factor Loading 4 Factors	Factor Loading 3 Factors	Cronbach's α If Item Deleted	Dimension's α	Cronbach's α
fac1	M11	.688	.686*	.888	.896	.944
(Cautiousness)	M15	.688	.687*	.888		
	M16	.722	.724	.884		
	M17	.772	.771*	.879		
	M18	.756	.757	.880		
	M19	.724	.723*	.881		
	M20	.736	.737	.883		
	M21	.694	.695	.885		
fac2	M4	.709	.711	.851	.870	
(Confidence)	M5	.715	.717	.848		
	M6	.730	.730	.848		
	M7	.657	.659	.856		
	M9	.661	.661	.856		
	M10	.658	.656*	.855		
	M12	.769	.767*	.846		
fac3	M1	.553		.719	.725	
	M2	.733		.577		
	M3	.772		.595		
fac4	M8	.693		.621	.698	
	M13	.633		.578		
	M14	.650		.620		
fac3 + fac4	M1		.513*	.788	.804	
(Introspection)	M2		.664*	.762		
	M3		.694*	.765		
	M8		.696	.764		
	M13		.625*	.778		
	M14		.635*	.786		

Note: Items with smaller Standardized factor loadings have been marked with *.

Tables 8 and 9 present the results concerning the convergent and discriminant validity of Model II. Following the criteria proposed by Cheung et al. (2023), the Average Variance Extracted (AVE) for each latent construct should exceed .50, while the Composite Reliability (CR) should be greater than .70. In Model II, all three dimensions satisfied the CR requirement, indicating acceptable internal consistency among their respective items. However, with respect to AVE, only the dimension labelled Cautiousness surpassed the threshold, registering a value of .523, thereby demonstrating satisfactory convergent validity. The AVE for Confidence was slightly below the benchmark at .492, suggesting marginal adequacy. Conversely, the Introspection dimension recorded a notably lower AVE of .411, which may be attributed to the abstract nature of its associated items. Many of these items originate from the "Awareness" component of the original scale, potentially reducing their clarity and specificity compared to the other two dimensions. Consequently, within the framework of the revised model, convergent validity is confirmed only for the Cautiousness dimension, while the Confidence and Introspection dimensions do not meet the required standard.

Table 8

Convergent Validity of Model II

		Parameter Significance Estimates				Convergent Validity				
		Unstd.	**S.E.**	**T-Value**	**P**	**Std.**	**SMC**	**1-SMC**	**CR**	**AVE**
Metacognitive Skills	Cautiousness	1.000				.947	.932	.068	.967	.908
	Confidence	1.031	.112	9.183	***	.946	.895	.105		
	Introspection	.610	.087	6.988	***	.965	.896	.104		
Cautiousness	M11	1.000				.686	.403*	.597	.898	.523
	M15	.963	.099	9.678	***	.687	.390*	.610		
	M16	1.016	.100	10.159	***	.724	.484*	.516		
	M17	1.146	.106	10.764	***	.771	.482*	.518		
	M18	.986	.093	10.583	***	.757	.441*	.559		
	M19	.987	.097	10.142	***	.723	.263*	.737		
	M20	1.084	.105	10.331	***	.737	.589	.411		
	M21	.976	.100	9.777	***	.695	.430*	.570		
Confidence	M4	1.000				.711	.436*	.564	.871	.492*
	M5	.969	.094	10.349	***	.717	.434*	.566		
	M6	1.001	.095	10.523	***	.730	.532	.468		
	M7	.906	.095	9.512	***	.659	.514	.486		
	M9	1.024	.107	9.542	***	.661	.505	.495		
	M10	.972	.103	9.469	***	.656	.482*	.518		
	M12	1.029	.093	11.058	***	.767	.544	.456		
Introspection	M1	1.000				.513	.522	.478	.806	.411*
	M2	1.398	.197	7.091	***	.664	.573	.427		
	M3	1.623	.224	7.262	***	.694	.595	.405		
	M8	1.496	.206	7.268	***	.696	.524	.476		
	M13	1.446	.211	6.851	***	.625	.472*	.528		
	M14	1.541	.223	6.914	***	.635	.470*	.530		

Note: SMC and AVE value less than .50 have been marked with *.

Discriminant validity is typically confirmed when the correlation coefficient is below .50 and the square root of the AVE exceeds the correlation coefficient. However, the results of the analysis indicated that the lowest correlation coefficient among the three dimensions was .896, considerably higher than the .50 criterion. Moreover, all AVE square root values

were lower than their respective correlation coefficients. Consequently, the new model does not demonstrate discriminant validity.

Table 9

Discriminant Validity of Model II

	AVE	Cautiousness	Confidence	Introspection
Cautiousness	.523	**.723**		
Confidence	.492	.896***	**.701**	
Introspection	.411	.914***	.914***	**.641**

Note: The open root AVE value is in bold.

Construction of the New Model

Despite the limited validity exhibited by the new theoretical model, this study proceeded with an evaluation of its model fit and conducted a comparative analysis with the original theoretical framework to ensure a comprehensive model assessment. Initially, Model I was developed in alignment with the original scale, representing a second-order four-factor structure encompassing the dimensions of awareness, cognitive strategies, planning, and self-checking. This model included a total of 20 items, with each dimension comprising five items (Figure 2). Subsequently, Model II was constructed based on the outcomes of the factor analysis, representing a second-order three-factor model. This revised model consisted of the factors Cautiousness (8 items), Confidence (7 items), and Introspection (6 items), resulting in a total of 21 items (Figure 3).

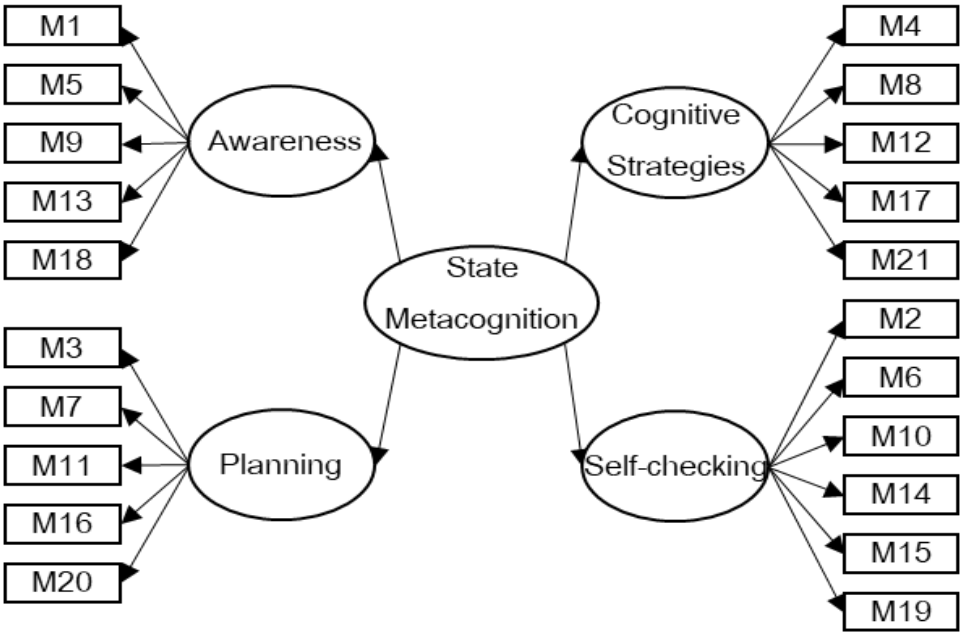

Figure 1: Model I

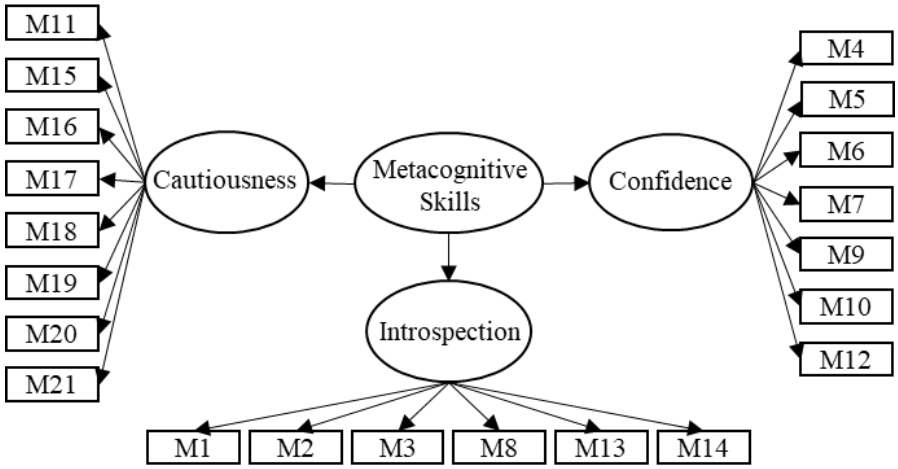

Figure 2: Model II

Comparison of Model Fit

As the sample in this study differed from that of the original research, the structure derived from the EFA varied from the original scale. Therefore, structural equation modelling was employed to evaluate both Model I and Model II in order to determine which model demonstrated superior fit. The model fit indices for both frameworks are presented in Table 10, where it is evident that both models met the criteria for acceptable model fit.

Table 10

Indicators of Fitness between Model I and Model II (n=231)

Index Criteria Value	NC <5	RMSEA <.08	TLI >.90	SRMR < .10	CFI >.90	PRATIO >.70
Model I	1.701	.055	.938	.045	.945	.881
Model II	**1.445**	**.044**	**.960**	**.041**	**.965**	**.886**
Result	pass	pass	pass	pass	pass	pass

Note: Bold indicates better indicators.

In accordance with the framework proposed by Hair et al. (1988), this study evaluated the compatibility of the models with the sample data using three distinct approaches to assess absolute model fit.

(1) Absolute Fit Indices: These indices assess how well the hypothesised model reproduces the observed data. The indices used include the chi-square to degrees of freedom ratio (χ^2/df, also referred to as the Normed Chi-square or NC), the Standardised Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA). A lower NC value suggests a better fitting model. Given that the chi-square statistic is highly sensitive to sample size, the χ^2/df ratio offers a more stable estimate of fit. An acceptable range for NC is generally considered to be between 2 and 5, with values below 2 representing a more desirable fit (Hair et al., 2010). Both models exhibited NC values below 2, suggesting that they offer a good fit to the data and are not overly complex. Notably, Model II yielded a smaller NC value

compared to Model I, implying that it demonstrates superior parsimony and is more effective and succinct in explaining the observed data.

SRMR reflects the average discrepancy between the observed and predicted correlations, while RMSEA estimates the degree of error in model approximation. Lower values for both indicate better model fit. According to the criteria outlined by Hu and Bentler (1999), SRMR values below .10 are acceptable, with values under .08 being more stringent. For RMSEA, values below .08 are considered adequate, and those below .06 denote excellent fit. Both models satisfied the cut-off thresholds for SRMR and RMSEA, although Model II exhibited more favourable values.

(2) Incremental Fit Indices: These indices evaluate the model by comparing its fit to that of a null model, which assumes no relationships among variables. Key indicators include the Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI). Conventionally, values above .90 indicate a good fit, while those exceeding .95 reflect an even stronger fit. However, Hu and Bentler (1999) caution that with small samples ($N \leq 250$), TLI and CFI may incorrectly reject a well-fitting model. To address this, they propose a dual-criterion strategy, requiring SRMR < .08 alongside TLI/CFI ≥ .95. Considering the present study's sample size (N = 231), these fit indices must be interpreted with care. As reported in Table 10, both models achieved SRMR values below .08, but only Model II attained TLI and CFI values above .95, specifically .960 and .965, respectively, indicating that Model II demonstrated a marginally better fit.

(3) Parsimony Fit Indices: The Parsimony Ratio (PRATIO) is used to assess the simplicity of a model by comparing the degrees of freedom of the null model with those of the specified model. The null model assumes complete independence among variables, whereas the specified (target) model is derived from theoretical assumptions. A higher PRATIO value indicates a more efficient model, as it captures the data structure using fewer estimated parameters. According to Mulaik (1998), a PRATIO value of .70 or higher is desirable. As presented in Table 10, both models surpassed this threshold, with Model II displaying a slightly higher PRATIO, reinforcing its greater parsimony relative to Model I.

Although the new model demonstrated limited validity, this study aimed to gain a more holistic perspective by evaluating its model fit in comparison to the original model. The findings revealed that Model II offered superior model fit in comparison to Model I.

Results and Discussion

This study undertook a revision of the State Metacognition Scale originally developed by O'Neil and Abedi (1996), adapting it as an assessment instrument for evaluating metacognitive skills among primary school pupils in Zhejiang Province. Following expert evaluation and modification, a revised theoretical framework was established. The revised model subsequently underwent a comprehensive series of analyses, including item analysis, factor analysis, tests of reliability and validity, as well as structural equation modelling to evaluate model fit.

Results

The findings of the study revealed that the revised items effectively distinguished between students with high and low levels of metacognitive skills. In the subsequent exploratory factor analysis, four factors were extracted in accordance with the structure of

the original model. During the reliability assessment phase, the overall reliability of the scale reached .944. Notably, the reliability coefficients of the first and second factors in the revised model surpassed those of any single factor in the original version. Following Nunnally (1978) recommendation, Factors 3 and 4 were merged to enhance internal consistency, resulting in an improved reliability coefficient of .804. This reconfiguration yielded three dimensions: Cautiousness (8 items), Confidence (7 items), and Introspection (6 items). Although this adjustment led to higher reliability, the standardised factor loadings for 11 items across the scale declined, thereby reducing the effectiveness of subsequent validity analyses. Specifically, only the Cautiousness dimension demonstrated adequate convergent validity, while the remaining two dimensions failed to meet the required thresholds. Furthermore, the revised model did not exhibit discriminant validity. Despite these shortcomings, the researchers continued the analysis by comparing the revised model to the original in terms of model fit. The results indicated that the revised model achieved a superior overall fit relative to the original framework.

The validation results suggest that upper primary school pupils may have encountered difficulties in understanding the vocabulary used in the scale, particularly in item M1, which assesses an individual's self-perception. The abstract nature of this item appeared to influence responses among younger respondents. More broadly, the inherently abstract construct of metacognition, coupled with the ongoing development of cognitive and linguistic abilities at this educational stage, likely hindered pupils' capacity to fully comprehend and accurately respond to complex metacognitive concepts. This may have led to a degree of randomness in answering, thereby contributing to the unsatisfactory convergent and discriminant validity outcomes observed. These findings are consistent with those of O'Neil and Abedi (1996), the original developers of the English version of the scale, who similarly reported limitations in its application among younger student cohorts.

Several potential factors may account for the limited convergent and discriminant validity: (1) Item design represents a key limitation. High inter-item correlations may result in reduced Average Variance Extracted (AVE) values, adversely affecting convergent validity. Similarly, poor alignment between items and their corresponding factors can undermine discriminant validity. (2) Low standardised factor loadings for certain items indicate weak associations with their underlying constructs, further impairing convergent validity. (3) Although the items may appear accessible, the complexity of the underlying constructs could impede elementary pupils' full comprehension, thereby necessitating further revision to improve item clarity and appropriateness. (4) The sample size may have been insufficient, potentially leading to biased estimations of AVE and CR, which would compromise validity outcomes. (5) The structure of the model itself may require adjustment, including a redefinition of the interrelationships among dimensions, to improve its fit and construct validity.

Contribution of the Research

Although the revised model demonstrates limitations in terms of validity, the scale remains effective in distinguishing between higher and lower levels of metacognitive skills, supported by a high reliability coefficient of .944. Therefore, it retains practical utility as a tool for the rapid evaluation of metacognitive competence. Additionally, the validity indices of the revised model may offer insights into the complex structure of children's

metacognition, serving as a basis for future scholarly inquiry. This study contributes to the development of a revised metacognitive skills assessment framework tailored to elementary students within the Chinese cultural context, thereby offering an empirical foundation for the localisation of cross-cultural measurement instruments.

Limitations of the Research

(1) While upper primary school students generally possess the capacity for logical reasoning and can comprehend concrete concepts as well as causal relationships, the findings of this study suggest that their ability to engage in abstract thinking and to reflect on metacognitive constructs remains limited. Additionally, the revised scale's title lacks brevity and contextual specificity, which may hinder students' full understanding of its intent.
(2) Although the original scale remains a succinct and established instrument, its direct application in the Chinese cultural context may not fully capture the cognitive and developmental characteristics of Chinese elementary students. Hence, moderate, contextually informed, and creative modifications could enhance its cultural relevance and measurement precision, thereby aligning it more effectively with the study's objectives.
(3) The use of a 4-point Likert scale may be overly abstract for young respondents, potentially leading to difficulties in interpretation. A more concrete response format might improve clarity and the accuracy of student responses.
(4) The total number of revised items may be excessive, potentially increasing cognitive load and response fatigue among students. A high volume of questions may induce confusion or result in disengaged and superficial answering behaviour, thereby affecting the quality of the data collected.

Future Research Directions

In view of the limitations identified in this study, several directions for future research are proposed. From a theoretical perspective:

(1) There is a need for deeper investigation into the dimensions of metacognitive skills among Chinese elementary school students, as well as the interrelationships among these dimensions. Such an inquiry would offer a stronger theoretical basis for refining the assessment scale.
(2) It is recommended that future evaluations of metacognitive skills incorporate comparative analyses aligned with specific extracurricular activity programmes. This approach could provide a broader understanding of how metacognitive skills vary across differing contexts.

From a methodological standpoint, future research should aim to optimise the scale's content, structure, and format. For example:

(1) Items should be reformulated to minimise cognitive demands and avoid complexity that surpasses students' developmental capacities.
(2) The total number of items could be reduced to lessen students' cognitive load and associated stress.

(3) Alternative response formats, such as colour-coded scales indicating degree, binary choices (e.g., yes or no), or visual emoticon-based indicators (e.g., smiling or sad faces), may facilitate more accessible and accurate self-assessment for young learners.

At the application level:

(1) Educators may integrate the Metacognitive Skills Scale into instructional practice to support targeted teaching and skills training.
(2) Given that the upper primary stage (ages 10–12) represents a critical period for metacognitive development, schools are encouraged to establish a systematic mechanism for regular assessment and feedback. Such systems can track students' developmental progress in metacognition and provide individualised feedback to foster self-reflection and enhance self-regulated learning.

In summary, the development of an assessment system that not only satisfies psychometric standards but also accurately reflects the characteristics of metacognitive development among Chinese children remains a significant challenge and a key objective for future research.

Conflict of interest:

All authors declare no conflict of interest.

Reference

Altındağ, M., & Senemoğlu, N. (2013). Metacognitive Skills Scale. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *28*(28-1), 15-26. https://dergipark.org.tr/en/pub/hunefd/issue/7789/101825

An, D., Ye, C., & Liu, S. (2024). The Influence of Metacognition on Learning Engagement the Mediating Effect of Learning Strategy and Learning Behavior. *Current Psychology*, *43*(40), 31241-31253. https://doi.org/10.1007/s12144-024-06400-y

Azevedo, R. (2020). Reflections on the Field of Metacognition: Issues, Challenges, and Opportunities. *Metacognition and Learning*, *15*(2), 91-98. https://doi.org/10.1007/s11409-020-09231-x

Bae, H., & Kwon, K. (2019). Developing Metacognitive Skills through Class Activities: What Makes Students Use Metacognitive Skills? *Educational Studies*, *47*(4), 456-471. https://doi.org/10.1080/03055698.2019.1707068

Brown, T. A., & Moore, M. T. (2012). Confirmatory Factor Analysis. *Handbook of structural equation modeling*, *361*, 379. https://www.researchgate.net/publication/251573889

Çetinkaya, P., & Erktin, E. (2002). Assessment of Metacognition and Its Relationship with Reading Comprehension, Achievement, and Aptitude. *Boğaziçi Üniversitesi Eğitim Dergisi*, *19*(1), 1-11. https://www.researchgate.net/publication/252277037

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2023). Reporting Reliability, Convergent and Discriminant Validity with Structural Equation Modeling: A Review and Best-Practice Recommendations. *Asia Pacific Journal of Management*, *41*(2), 745-783. https://doi.org/10.1007/s10490-023-09871-y

Currie, C., Molcho, M., Boyce, W., Holstein, B., Torsheim, T., & Richter, M. (2008). Researching Health Inequalities in Adolescents: The Development of the Health

Behaviour in School-Aged Children (Hbsc) Family Affluence Scale. *Social Science & Medicine*, *66*(6), 1429-1436. https://doi.org/10.1016/j.socscimed.2007.11.024

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist*, *34*(10), 906-911. https://doi.org/10.1037//0003-066x.34.10.906

Güner, P., & Erbay, H. N. (2021). Metacognitive Skills and Problem-Solving. *International Journal of Research in Education and Science*, *7*(3), 715-734. https://doi.org/10.46328/ijres.1594

Hair, J. F., Anderson, R. E., & Tatham, R. L. (1988). *Multivariate Date Analysis with Reading* (2nd ed.). Maxwell MacMillan International. https://archive.org/details/multivariatedata0000hair/page/n5

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). https://www.drnishikantjha.com/papersCollection/Multivariate%20Data%20Analysis.pdf

Hu, L. t., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality Assessment Using the Full-Information Item Bifactor Analysis for Graded Response Data. *Educational and Psychological Measurement*, *68*(4), 695-709. https://doi.org/10.1177/0013164407313366

Kahan, T. L., & Sullivan, K. T. (2012). Assessing Metacognitive Skills in Waking and Sleep: A Psychometric Analysis of the Metacognitive, Affective, Cognitive Experience (Mace) Questionnaire. *Consciousness and Cognition*, *21*(1), 340-352. https://doi.org/10.1016/j.concog.2011.11.005

Li, F., Yuan, D., Gao, C., Xiong, K., Geng, F., & Zhang, L. (2023). Validity and Reliability of the Metacognitions Questionnaire-30 (Mcq-30) among Chinese Adolescents. *Child Psychiatry & Human Development*, *56*(4), 1031-1040. https://doi.org/10.1007/s10578-023-01625-7

Li, X., Wang, M., Feng, X., Yin, X., & Liang, J. (2024). An in-Depth Analysis of the Personal Factors and Their Pathways in Shaping Self-Directed Learning Abilities among Undergraduate Nursing Students. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1450462

Moses-Payne, M. E., Habicht, J., Bowler, A., Steinbeis, N., & Hauser, T. U. (2021). I Know Better! Emerging Metacognition Allows Adolescents to Ignore False Advice. *Developmental Science*, *24*(5). https://doi.org/10.1111/desc.13101

Mulaik, S. A. (1998). Parsimony and Model Evaluation. *The Journal of Experimental Education*, *66*(3), 266-273. https://doi.org/10.1080/00220979809604411

Ning, H. K. (2017). The Bifactor Model of the Junior Metacognitive Awareness Inventory (Jr. Mai). *Current Psychology*, *38*(2), 367-375. https://doi.org/10.1007/s12144-017-9619-3

Nunnally, J. C. (1978). An Overview of Psychological Measurement. In *Clinical Diagnosis of Mental Disorders* (pp. 97-146). Springer US. https://doi.org/10.1007/978-1-4684-2490-4_4

O'Neil, H. F., & Abedi, J. (1996). Reliability and Validity of a State Metacognitive Inventory:

Potential for Alternative Assessment. *The Journal of Educational Research*, *89*(4), 234-245. https://doi.org/10.1080/00220671.1996.9941208

Saldarriaga, J. V., Jaimes, C., Polo, E. A., & Merino, M. (2012). Validez, Confiabilidad Y Baremación Del Inventario De Estrategias Metacognitivas En Estudiantes Universitarios. *Revista de Psicología (Trujillo)*, *14*(1), 9-20. https://revistas.ucv.edu.pe/index.php/revpsi/article/view/438

Schuster, C., Stebner, F., Leutner, D., & Wirth, J. (2020). Transfer of Metacognitive Skills in Self-Regulated Learning: An Experimental Training Study. *Metacognition and Learning*, *15*(3), 455-477. https://doi.org/10.1007/s11409-020-09237-5

Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4-11. https://doi.org/10.12691/ajams-9-1-2

Smortchkova, J., & Shea, N. (2020). Metacognitive Development and Conceptual Change in Children. *Review of Philosophy and Psychology*, *11*(4), 745-763. https://doi.org/10.1007/s13164-020-00477-7

Spielberger, C. D. (1966). Theory and Research on Anxiety. In *Anxiety and Behavior* (pp. 3-20). Elsevier. https://doi.org/10.1016/b978-1-4832-3131-0.50006-8

Tavakol, M., & Wetzel, A. (2020). Factor Analysis: A Means for Theory and Instrument Development in Support of Construct Validity. *International Journal of Medical Education*, *11*, 245-247. https://doi.org/10.5116/ijme.5f96.0f4a

Tibken, C., Richter, T., von der Linden, N., Schmiedeler, S., & Schneider, W. (2021). The Role of Metacognitive Competences in the Development of School Achievement among Gifted Adolescents. *Child Development*, *93*(1), 117-133. https://doi.org/10.1111/cdev.13640

Zawidzki, T. W. (2019). A New Perspective on the Relationship between Metacognition and Social Cognition: Metacognitive Concepts as Socio-Cognitive Tools. *Synthese*, *198*(7), 6573-6596. https://doi.org/10.1007/s11229-019-02477-2