# Creation and Validity and Reliability Testing of an AI Literacy Assessment Scale for Teachers - Based on a Dual-track Empirical Study of Zhejiang Province's Primary and Secondary Schools and Higher Education Institutions

Youyou Xiang*, Feiyu Zhan, Qinlin Cai
Zhejiang Normal University, Jinhua 321004, China
*Corresponding email: 2322892914@qq.com

**Abstract**

Teachers' artificial intelligence (AI) literacy has become a critical determinant of educational digital transformation efficacy amid the deep integration of AI and pedagogy. Nevertheless, systematic, locally grounded, and psychometrically validated assessment instruments remain scarce. Persistent challenges, including cross-level educational misconceptions, the theory-practice gap in ethical evaluation, and disconnection from regional policy contexts - have constrained empirical progress in this domain. Drawing upon the digital transformation practices of an East Chinese province and its officially issued AI literacy frameworks for kindergarten through 12th grade (K-12) and higher education teachers, this study developed two parallel standardized scales. Both models exhibited a hierarchical second-order structure comprising value standards, practical literacy, and cognitive literacy. The K-12 scale yielded five first-order factors: classroom implementation, teaching-research inquiry, class management, cognitive level, and normative level. The higher education scale identified another five first-order factors, namely academic support, academic inquiry, classroom implementation, cognitive level, and normative level. Second-order factor analyses confirmed the theoretically posited hierarchical relationships. Each final scale retained 22 items, with cumulative variance explained rates of 79.30% and 76.90%, respectively. Confirmatory factor analyses demonstrated satisfactory fit comparative fit index (CFI=0.957) and 0.935 and root mean square error of approximation (RMSEA)=0.067 and 0.079. Convergent and discriminant validity met recommended thresholds, with overall Cronbach's α coefficients of 0.948 and 0.936. Group difference analyses revealed a rigid age-related decline in AI competence, with teaching experience effects largely attributable to generational cohort differences, alongside systematic variations across subject area, professional role, and school location. The resulting scales exhibit robust psychometric properties and serve as scientific tools for regional policy evaluation, tiered professional development, and precision diagnosis of teacher AI literacy. The value-integrated design, which embeds ethical principles within specific application scenarios, offers a novel approach to reconciling the attitude-behavior measurement gap. Findings suggest that policy interventions should transition from generalized training to targeted empowerment, prioritizing subject-specific tool development, remedial programs for mid-career and senior teachers, and rural infrastructure enhancement.

**Keywords**

Artificial intelligence, Ethical standards, Practical literacy, Cognitive literacy, Group difference

## Introduction

Teachers are becoming organizers of human-machine collaboration as a result of the deep integration of artificial intelligence (AI) into education and teaching. Teachers must change from being passive technology users to active organizers of human-machine collaborative learning environments, according to the United Nations Educational, Scientific and Cultural Organization (UNESCO)'s 2024 "Framework for Teachers' AI Competencies". Improving teachers' AI literacy is part of China's national plan. The Ministry of

Education and eight other departments released the "Opinions on Accelerating the Digital Transformation of Education" in 2025. In 2026, the Ministry of Education and four other departments released the "AI+Education Action Plan", which called for the creation of contextualized assessment systems and AI literacy standards for educators. In light of this, teachers' AI literacy has developed from a supplement to their information technology expertise to a crucial factor in the digital revolution of education.

However, there are still obstacles in the way of scholarly debates and real-world initiatives pertaining to teachers' AI literacy. There are still issues with the systematization, localization, and psychometric validation of assessment instruments, despite the academic community's recognition of their significance [1]. The majority of current research concentrates on theoretical discussions and descriptions of the current situation, with frameworks that are either modeled after international ones or transplanted from general digital literacy. These frameworks fail to adequately account for the distinctive professional contexts of Chinese teachers and the features of differences across educational levels [2]. Although current frameworks are useful for reference, the majority are still theoretical and have not been converted into useful indicators. More significantly, academics in higher education and kindergarten through 12th grade (K-12) instructors have different professional responsibilities. While academics in higher education perform scientific research and direct students' academic growth, primary and secondary teachers are in charge of classroom administration, home-school collaboration, and the protection of minors.

However, the majority of scales now in use employ generic items and assess samples of both types simultaneously, which leads to measurement material that is significantly detached from teachers' actual work. The fact that the AI ethics dimension is usually handled as a stand-alone "attitude" element with no link to particular application scenarios is another long-ignored problem. Although they may react favorably to ethical attitude questions, teachers find it difficult to put these ideas into practice in actual classrooms. Ethical evaluations become a formality as a result of this gap between knowledge and action. These restrictions put regional education authorities in a difficult position since they lack trustworthy instruments to assess the efficacy of policies and diagnostic data to direct teacher preparation. In conclusion, there are four main issues with the field of teacher AI literacy assessment: a lack of locally adapted, standardized tools; a disconnect between local policies and scales; a confusion of various educational levels and scenarios; a disconnect between ethical measurement knowledge and practice.

As a National Pilot Zone for Digital Economy Innovation and Development, a province in East China has led the country in creating an AI literacy framework for teachers. The province's Department of Education released two documents in 2025: The Framework for Artificial Intelligence Literacy of Primary and Secondary School Teachers (Trial) and the Framework for Artificial Intelligence Literacy of Higher Education Faculty (Trial). These two frameworks were methodically created around three dimensions - foundational literacy, applied literacy, and ethical standards. While the framework for higher education stresses academic ethical standards and digital humanities, the framework for primary and secondary schools places more emphasis on protecting the privacy of minors' data.

Although the publication of these two frameworks offers specific requirements for instructors' AI literacy, their actual use necessitates the use of standardized evaluation instruments that properly match them. However, the majority of scales now in use, both domestically and globally, are based on generic norms or UNESCO frameworks, which leaves the province's policy context severely lacking [3]. There are almost no components in the current scales that relate to the regional requirements highlighted in the frameworks, such as value generation, support for pupils with special needs, and home-school partnership. As a result, there is now a glaring gap between measurement instruments and policy needs.

Based on the aforementioned research, this study is positioned as a policy-oriented and practice-adapted scale development study, utilizing the province's educational digital transformation practices and the two provincial teacher AI literacy frameworks as its foundation. Three main concerns are intended to be addressed by the study: What aspects make up teachers'

AI abilities? What are the validity and reliability levels of the established scale? What are the overall performance of teachers' AI competencies and the features of variances across different groups? The study turns policy-based behavioral descriptions into measurable indicators by anchoring practical competences to their respective core work contexts through the construction of two distinct theoretical models, one for K-12 schools and another for higher education institutions. Additionally, it uses a value-integrated structural design to change ethical measuring from attitude assessments that stand alone to behavioral assessments that are integrated with particular circumstances. Strictly following the full-chain verification procedure of exploratory and confirmatory factor analysis, the study will rely on large-sample, stage-specific empirical validation. The two standardized scores that were eventually created can be useful instruments for assessing the efficacy of policies, implementing training in tiers, and diagnosing instructors' AI literacy in different regions.

## Review of literature

### *The development of the idea of teachers' AI literacy*

The breadth of AI literacy, an emerging academic term, has steadily grown in tandem with the advancement of AI technology, moving from conventional basic literacy abilities to more intricate, all-encompassing competency standards. Early studies concentrated on a technical viewpoint, characterizing it as a person's capacity to comprehend the fundamental ideas, principles, and applications of artificial intelligence, with a primary focus on cognitive comprehension of basic concepts and the operation of basic instruments. AI literacy was first defined as a set of skills that allow people to identify, comprehend, and utilize AI technologies. The academic community then turned its attention to the teaching profession; at this point, the idea was still focused on technology application, highlighting instructors' fundamental abilities to use AI tools to improve instruction.

This idea has quickly developed into a multifaceted, interdisciplinary construct as artificial intelligence's social effect has grown. Academics currently agree that the fundamental component of AI literacy is enabling people to critically assess AI technologies, engage and cooperate with them successfully, and use them to solve practical issues. AI literacy, according to Long and Magerko, is the capacity to assess AI technologies critically, interact and cooperate with them successfully, and use them. A four-dimensional approach embracing cognition and understanding, use and application, assessment and invention, and AI ethics is also put forth by Ng [4]. All things considered, current definitions of AI literacy typically go beyond the purely technical operational level, pointing to a composite construct that incorporates critical analysis, ethical responsibility, technological application, and cognitive comprehension.

The notion of AI literacy in the teaching community has changed over time, moving from a straightforward extension of digital literacy into the AI field to a framework specifically designed for educators. For educators to conduct instruction, student growth, and educational research in intelligent learning settings, teacher AI literacy is a crucial core competency. Three fundamental viewpoints can be used to distill its essence. One dimension focuses on the deep integration of AI technologies with teaching, research, and administrative contexts through technology application and teaching empowerment. A second dimension encompasses comprehension, application, and critical reflection, which emphasizes educators' critical thinking and in-depth knowledge of AI technologies. The final dimension relates to ethical responsibility, which integrates data privacy, educational equity, and AI ethics into its fundamental scope. In conclusion, teachers' AI literacy has developed from a single technical ability into a holistic system that includes three dimensions: ethical standards, contextual application, and fundamental knowledge. Existing systematic sorting of measurement tools further confirms the fragmented design of current AI literacy evaluation scales [5].

### *Studies on the AI literacy framework for teachers*

The foundation for carrying out literacy tests, putting tiered training into practice, and passing legislation is the framework for instructors' AI literacy. Systematic study findings about the framework's components, theoretical models, adaptation to various educational stages, and practical applications have surfaced both locally and globally.

Three basic types of frameworks have emerged on a global scale: regional initiatives, theory-derived frameworks, and standards-driven frameworks. The Framework for Teachers' Artificial Intelligence Competencies, which emphasizes humanism and AI ethics, was published by UNESCO in 2024. It is broken down into three progressive levels: acquisition, deepening, and creation. It encompasses five domains: human-centered thinking, AI ethics, AI fundamentals and applications, AI pedagogy, and AI professional development. The deep integration of AI technology, instructional content, and pedagogy is the micro-level focus of the AI-Technological Pedagogical Content Knowledge (TPACK) and Smart TPACK frameworks, which are based on traditional ideas of educational technology. AI technical knowledge, AI pedagogical knowledge, AI content knowledge, human-machine cooperation knowledge, and fundamental ethical knowledge are the focal points of Celik's Smart TPACK architecture [6].

Additionally, references for grade-level-specific design are provided by regional projects like the European Union (EU)'s DigComp 2.2 digital literacy framework and the U.S. Artificial Intelligence for K-12 Initiative (AI4K12) project guidelines. It is challenging to immediately match international frameworks with China's educational scenarios and local policy requirements due to their lack of practical applicability and limited flexibility to local circumstances, despite their overall comprehensiveness and strong value-driven guidance.

The numerous components and practice-oriented approach of domestic research, which is based on the demands of educational digital transformation, define it. Some researchers presented a three-dimensional framework with fundamental, advanced, and ethical levels. AI literacy was incorporated into a five-dimensional digital literacy framework. Some researchers developed a "1+4" competency framework based on prompt engineering with an emphasis on university professors. Simultaneously, research produced a fundamental theoretical logic that included contextual applications, technological underpinnings, and ethical standards, making it clear that the framework needs to be rooted in the three main work settings of administration, teaching, and research. Nevertheless, there are still three issues with current frameworks.

First, they don't accurately discern the contextual distinctions between K-12 and university teachers, making them inadequately adapted to various educational levels. Based on the extended intelligent-TPACK theoretical system, differentiated implementation paths for teachers at different educational stages can be sorted out [7]. Second, because most frameworks have not been converted into useful assessment measures, actual implementation is lacking. Third, there is a lack of adaptation to regional and local contexts. These frameworks do not adequately emphasize unique components like digital humanities and the security of minors' data, nor do they conform to local policy needs. Specialized frameworks for higher education teachers also ignore localized academic scenario requirements [8].

***Research on teachers' AI literacy assessment tools***

The primary means of carrying out accurate assessments of these competencies are AI literacy assessment tools for educators. An integrated method based on literacy frameworks has been embraced by research both locally and abroad, integrating empirical validation with framework-oriented design.

Using standardized procedures like literature reviews, expert consultations, and empirical research, scale development at the global level is based on UNESCO's AI Teacher Competency Framework and traditional theories like AI-TPACK, with an emphasis on components like AI ethics, technology application, and integration into teaching. A self-evaluation tool for instructors' AI competences was created by Delcker [9]. It covers topics including AI knowledge, AI pedagogy, AI assessment, and AI ethics. International tools, however, typically have inadequate localization and adaptation. In accordance with the "Teacher Digital Literacy" industry standard, domestic research has created scales that are specific to the different roles of instructors in primary and secondary schools vs. higher education institutions.

Nevertheless, the majority of current tools have only undergone small-scale preliminary development, with constrained sample sizes and geographic coverage. Tools targeting university researchers merely focus on academic scenarios without matching K-12 teaching management demands [10]. Additionally, they frequently leave out crucial processes like exploratory and confirmatory factor analysis, and their validity, reliability, and application to other populations have not been thoroughly examined. Some similar scale development research only targets student groups rather than frontline educators [11].

***The focus of this study and the restrictions of current research***

The aforementioned study shows that current research still has a lot of shortcomings. First, there is a significant lack of standardized measurement instruments; there are very few instruments that have undergone rigorous psychometric validation and can be used directly for large-scale evaluations. Second, there is a significant mismatch between the scales and the context of provincial policy. A province in East China was the first in the country to release an AI literacy framework for K-12 and higher education instructors in 2025, yet there are hardly any corresponding items on the scales that are now in use. Third, there is insufficient anchoring to particular educational settings; current scales utilize almost similar items for instructors in K-12 and higher education, failing to recognize basic distinctions between essential contexts like scientific research and classroom management. Fourth, there is a gap between attitudes and behavior because the ethical aspect is handled as a distinct "AI ethics attitudes" portion and is not tied to particular application scenarios. Most general AI literacy measurement tools fail to embed ethical indicators into daily work scenarios [12].

This study leverages the digital transformation methods of a particular East Chinese province as a starting point and the province's AI literacy framework for teachers as its fundamental foundation to overcome these problems. It is positioned as a practice-adapted and policy-oriented scale development study. Its goal is to develop a new scale that precisely aligns with the province's AI literacy framework for teachers, differentiating between versions for K-12 and higher education teachers, and turning the framework's behavioral descriptions into quantifiable items to achieve precise alignment between measurement tools and policy.

**Model building**

***Theoretical foundation***

This study's model creation is based on Bloom's taxonomy of educational objectives and Piaget's theory of cognition. According to Piaget, knowledge is created via action, and application in practical settings deepens cognitive comprehension. The transforming relationship between knowledge and skills within the cognitive domain is revealed by Bloom's Taxonomy: applied abilities reflect contextual practice, foundational abilities relate to cognitive acquisition, and both dynamically change over professional development. The epistemological foundation for the three-dimensional framework of value standards, practical literacy, and cognitive literacy is provided by this logic.

This study also incorporates the human-centered approach to AI application found in UNESCO's "Framework for Teachers' AI Competencies". This approach emphasizes teachers' roles as educators rather than as technology users and centers on ethical awareness, instructional integration, and professional development. A theoretical lens for comprehending the knowledge instructors need to incorporate AI is provided by analyses of frameworks like AI-TPACK, which investigate the deep integration of technology, pedagogy, and content. Building on this basis, this study achieves a methodical translation of basic frameworks into the professional contexts of teachers by further anchoring practical competence in particular work contexts.

***Building a dual-track model: Differentiated design for higher education and K-12***

This study rejects a "one-size-fits-all" competency framework and builds two simultaneous, dual-track teacher AI competency models based on structural distinctions in educational logic between K-12 and higher education. The three-dimensional framework of these models - AI cognitive competencies, AI practical competencies, and AI ethical standards - is the same, but the contextual anchors for practical competencies have distinct designs.

Three main scenarios serve as the foundation for practical competences in the K-12 model: classroom management, teaching and research investigation, and classroom implementation. The foundation of this design is a methodical examination of the primary duties of K-12 educators. In addition to teaching, teachers also administer the classroom, collaborate with the home school, and keep an eye on the physical and mental well-being of their students. Preventing algorithmic bias and safeguarding the privacy of minors' data are key components of establishing classroom management as a stand-alone scenario.

The higher education model properly aligns with the academic nature of higher education by anchoring practical competence in three scenarios: academic inquiry, academic support, and classroom execution. It shifts the ethical focus to preventing academic misconduct by highlighting unique aspects of academic inquiry, such as handling large amounts of material and analyzing high-dimensional data.

Cognitive competency, which includes knowledge and abilities as well as value understanding, forms the basis of both frameworks. Value concepts, which show themselves in the three dimensions of safety, fairness, and benefit, permeate every step of the AI application process. Cognitive and ethical aspects are integrated and activated by practical ability. The three components work together in a dialectical and cohesive way, with practice driving quality development, ethics defining direction, and cognition supporting practice.

***Design with value integration***

By naturally integrating standards like data security and algorithmic fairness into the components of particular application scenarios, this study takes a value-integrated approach, departing from the conventional method of considering ethics as a stand-alone segment. Similar integrated ethical measurement ideas have been adopted in classroom-oriented teacher scales [13].

Teachers may score highly on attitude items but find it difficult to put them into practice since traditional ethical measurement treats ethics as a distinct "AI ethics attitude" part, creating a gap between attitudes and conduct.

In this study, value principles are refined into observable behavioral indicators; teachers' adherence to ethical boundaries is evaluated concurrently with their responses to scenario-based application questions. Ethics infuses teaching, research, and classroom management at the K-12 level, with a focus on safeguarding the privacy of minors' data. Teaching, research, and student issues are all impacted by ethics in higher education, with a focus on academic ethical standards. This methodology achieves unified measurement of attitudes and actions by reflecting the interconnectedness of value judgment and technology application in educational practice.

***Framework for theoretical models***

Both models have three first-level dimensions based on the framework mentioned above.

AI cognitive literacy, which includes knowledge, abilities, and value understanding, is the first dimension. While knowledge and skills cover fundamental knowledge of general AI and the fundamental competences needed for educational applications, value understanding entails acknowledging the role of AI in value creation and digital humanities from the perspective of reconstructing the educational ecosystem.

AI Practical Literacy, which focuses on certain behaviors regarding AI integration across three primary work settings, is the second dimension. These scenarios include classroom management (including daily administration, home-school collaboration, health management, and special needs support), educational research and exploration (including knowledge management, research design, data processing, and writing assistance), and classroom implementation (including student research, curriculum development, instructional interaction, and classroom analysis) in K-12 settings. Classroom implementation, academic inquiry (including knowledge management, research design, data processing, and writing help), and student affairs support (including daily management and career planning) are the main areas of attention in higher education.

AI value principles make up the third dimension and provide the overall ethical framework for the entire process. These can be seen in equity ethics (preventing algorithmic bias and protecting vulnerable groups), safety ethics (protecting data privacy), and benefit ethics (ensuring technology serves the purpose of education and preventing misuse).

This three-dimensional framework provides clear conceptual boundaries for further item creation and empirical validation by establishing a dialectical and unified relationship in which ethical norms lead direction, practice activates ethics, and cognition supports practice. Hierarchical second-order factor structures are widely adopted in mature AI literacy measurement tools [14].

## Research techniques

***Overall procedure and research design***

The theoretical construction, item generation, content validity testing, pilot testing, formal administration, model validation, and reliability and validity testing were all carried out in accordance with the usual paradigm for scale development. The study conducted independent scale development for college professors and elementary and secondary school teachers, integrating expert evaluation and interview-based induction procedures. A study framework with two parallel tracks and independent validation for each educational level was created by designing the two sets of scales differently in terms of theoretical framework and item content depending on the features of their unique professional settings. Throughout the whole procedure, the K-12 and higher education samples, data, and statistical analyses were kept completely apart. The researcher's university's ethics committee examined and approved the research protocol, guaranteeing adherence to academic ethical standards.

***Theoretical model construction***

The theoretical underpinnings of this work are Bloom's Taxonomy and Piaget's epistemology. Additionally, it builds theoretical models of AI competencies for K-12 and higher education teachers, respectively. This research adopts the UNESCO "Framework for Teachers' AI Competencies", which emphasizes the integration of ethical awareness and teaching. It also draws on the Intelligent TPACK framework, which offers an analytical perspective on deep integration of technology, pedagogy, and content. According to Piaget's constructivist theory, a thorough grasp of AI must be developed within real-world education, research, and administrative contexts because individual cognition cannot be isolated from practical application. The cognitive domain's transformative relationship between knowledge and competencies is revealed by Bloom's Taxonomy: applied competencies reflect contextual practice, foundational competencies correspond to cognitive acquisition, and both dynamically change during professional development.

Although the contextual anchoring of practical literacy varies, both models share three top-level dimensions: AI cognitive literacy, AI practical literacy, and AI value principles. Three crucial contexts - classroom implementation, teaching and research investigation, and class management - are the foundation of the K-12 model for practical literacy. Among these, class management is handled as a separate context, emphasizing fundamental principles such as algorithmic fairness and the privacy of minors' data. Three scenarios serve as the foundation for practical competency in the higher education model: classroom implementation, academic inquiry, and academic assistance, emphasizing particular requirements such as data processing and literature management in academic inquiry. Safety ethics, fairness ethics, and benefit ethics are the three dimensions of value concepts that permeate the entire AI application process. These three components work together to create a dialectical and cohesive relationship where ethics directs direction, practice activates ethics, and cognition supports practice. Based on this, this study departs from the conventional approach of addressing ethics as a distinct subject by adopting a value-integration framework. Instead, it addresses the measuring gap brought about by the discrepancy between ethical ideas and actual conduct by integrating safety, fairness, and benefit-based ethical indicators into behavioral items across all work settings.

***Content validity testing and item development***

The concepts of contextuality, relevance, clarity, and universality are rigorously adhered to during item construction. Items must be deeply ingrained in teachers' actual daily work environments in order to be properly considered contextual. Each item must specifically target and measure a particular sub-construct in order to be considered relevant. Concise and unambiguous item phrasing is necessary for clarity. In order to accommodate the disparities between teachers from different disciplinary backgrounds and the quick improvements in technology, universality necessitates that items concentrate on generic competencies and behaviors. Three types of item sources are available: a bank of already published things, newly created items customized to the particular work environments of each educational level, and policy text modifications based on particular behavioral descriptions in two official frameworks from an East Chinese province.

The study team developed a preliminary 30-item initial scale for K-12 teachers and a 28-item initial scale for higher education faculty using a three-step process: mutually defining components, independently creating items, and iteratively discussing and refining them.

Stratified sampling and separated Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) verification are standard operations for K-12 teacher scale development [15].

The content validity of items across four dimensions-clarity, relevance, logic, and adequacy - was then assessed using a 5-point Likert scale by a review panel made up of eight specialists in educational technology, teacher education, and psychometrics. Expert scores showed good consistency, with intraclass correlation values ranging from 0.75 to 0.90. The research team refined broad statements into more particular ones that were in line with instructors' work contexts, eliminated items with inadequate content validity, and combined semantically redundant items based on expert feedback. The team conducted semi-structured interviews with thirty K-12 teachers and thirty faculty members in higher education. The interviews aimed to investigate teachers' practical use of AI tools and their perceptions of core competencies in teaching, research, and administrative contexts. The K-12 teachers represented urban and rural areas and a variety of subject areas, while the higher education faculty represented undergraduate, vocational, and multidisciplinary fields. The team further adjusted item contextual relevance based on interview comments after identifying common psychological traits and behavioral patterns through thematic analysis of interview transcripts. Revised scales with 26 items each for K-12 and higher education were created after iterative optimization. Two pilot scales with 23 items each were produced after invalid items were removed during subsequent pilot testing and item analysis.

***Item analysis and pilot testing***

100 K-12 teachers and 100 faculty members from higher education were chosen to represent a range of geographic locations, educational levels, disciplinary backgrounds, and years of teaching experience. The pilot testing was carried out using an online survey platform. To guarantee the validity of the data, attention-checking questions were incorporated into the survey. The purpose of the pilot test was to assess the scale's practicality, understandability, and structural validity in real-world testing conditions. Based on the results of the pilot test, a systematic item analysis was carried out. The critical ratio method was adopted to test item discrimination, which involved grouping the sample by total score, comparing the score differences for each item between high-score and low-score groups, and eliminating items that did not reach significance. Each item's correlation coefficient with the overall score was computed, and items having a correlation of less than 0.30 were eliminated. Item-by-item deletion of observations was followed by a comparison of changes in the overall Cronbach's α coefficient. If α increased significantly following removal, the item was considered to have insufficient discriminant validity. The scale structure was logically reorganized in accordance with dimensions of cognitive competence, practical competence, and value standards based on analysis results.

Items with unclear wording or inadequate contextual relevance were revised, and items with low discriminant validity or weak correlation were eliminated based on analysis results. In the end, this produced a 23-item pilot scale for college instructors and a 23-item pilot scale for K-12 educators. Differentiated scoring was done using a five-point Likert scale: a degree of agreement scale for the cognitive competence component, a frequency of usage scale for the practical competence dimension, and a behavioral conformity scale for the value standards dimension.

***Data and sample gathering***

This study was carried out in an East Chinese province under the administrative supervision of the Department of Education's Teacher Affairs Office. A large-scale formal assessment based on a particular percentage of the province's total teaching population was conducted using a stratified random sampling approach.

The sampling of K-12 teachers considered main topic categories, urban, county, and rural locations, as well as educational levels, regions, and subject areas, including elementary, middle, and high school. Undergraduate and vocational colleges were included in the sampling of higher education faculty, which took into account institution type, discipline, and professional level. The official survey was done separately for K-12 teachers and higher education faculty using a professional online survey platform, and questionnaires were disseminated under the guidance of university human resources departments and municipal and county education bureaus. Teachers with some knowledge of or expertise with AI technology were specifically addressed by the questionnaire instructions. Over 21,300 questionnaires were gathered from K-12 educators. 18,160 valid surveys with an 85.10% valid response rate were obtained after thorough data cleaning, which eliminated invalid questionnaires because of abnormally short response times or extremely consistent selections.

After data cleaning, 1,508 valid questionnaires with an 89.20% valid response rate were obtained from the more than 1,680 questionnaires that were gathered from higher education faculty. The sample size completely satisfies statistical requirements for both exploratory and confirmatory factor analysis, and the sample structure is highly typical of teacher populations at matching educational levels in a specific East Chinese region.

***Investigative factor evaluation***

This study used hierarchical exploratory factor analysis, which was carried out in two stages, to examine the latent multidimensional structure of the scale and test the theoretically proposed hierarchical model. Initially, the basic first-order factor structure was investigated using exploratory factor analysis on observed items. The hierarchical structure of instructors' AI competencies was then fully revealed by doing a second-order exploratory factor analysis based on dimensional scores aggregated from first-order factors to investigate higher-order competency constructs. Two homogenous subsamples were randomly selected from the formal test samples from K-12 schools and universities: Subsample 2 is used for confirmatory component analysis, whereas Subsample 1 is used for exploratory factor analysis. This prevented possible overfitting problems from using statistical techniques on a single sample and guaranteed the independence of the model exploration and validation stages.

Before analysis, suitability checks were carried out. The Kaiser-Meyer-Olkin (KMO) value for the K-12 sample was 0.952, and Bartlett's sphericity test was significant. Both datasets were appropriate for factor analysis, as evidenced by the higher education sample's KMO value of 0.934 and the significance of Bartlett's sphericity test. Factor extraction was done using principal component analysis, and factor rotation was done using the maximum variance method. Items that did not match the screening criteria were eliminated. The criteria were unidimensional standardized factor loadings of at least 0.5, cross-loadings between dimensions below 0.35, and communality of at least 0.30. Twenty-two items were kept for the elementary and secondary school sample, resulting in five factors that together accounted for 79.30% of the variance. Cognitive Level, Normative Level, Classroom Implementation, Teaching and Research Exploration, and Class Management were the names given to these elements. Additionally, 22 items from the university sample were kept, resulting in five factors that accounted for 76.90% of the variation. These elements were referred to as academic inquiry, classroom implementation, normative level, cognitive level, and academic support.

The mean scores of the five dimensions were used as variables in a second-order exploratory factor analysis based on first-order factor scores. Three factors with eigenvalues larger than 1 were found in the K-12 sample, accounting for 90.70% of the cumulative variance. Among them, classroom management, teaching and research exploration, and classroom implementation loaded heavily on factor 1, also known as "practical competence". Normative dimension loaded only on factor 2, and the cognitive dimension loaded only on factor 3. Three factors were also found in the higher education sample, accounting for 87.50% of the cumulative variance. Among them, Factor 1 was loaded by classroom implementation, academic inquiry, and student affairs support; factors 2 and 3 were loaded by the normative dimension and cognitive dimension, respectively. The three-dimensional theoretical framework of cognitive competence, practical competence, and value norms was validated by the second-order factor structure, demonstrating the empirical validity of the hierarchical linkages proposed by theory.

***Validity and reliability tests and confirmatory factor analysis***

Subsample 2 was used to test the model's goodness of fit using confirmatory factor analysis. A first-order five-factor model was built first, followed by a second-order three-factor model based on theoretical presumptions to evaluate the validity of the higher-order structure, in accordance with a validation process from lower to higher orders. First-order model fit statistics for primary and secondary school data were comparative fit index (CFI)=0.955, root mean square error of approximation (RMSEA)=0.069, and standardized root mean square residual (SRMR)=0.055; second-order model fit values were CFI=0.957, RMSEA=0.067. Higher education's first-order model fit values were CFI=0.934, RMSEA=0.080, and Tucker-Lewis Index (TLI)=0.924; the second-order model fit was similar. The second-order model's fit was similar to the first-order model's, suggesting that the second-order three-factor structure is highly consistent with the theoretical framework and more economical. Further demonstrating hierarchical linkages, a comparison of competing models reveals that the second-order higher-order factor model performs better than both single-factor and first-order models.

Tests for validity and reliability were carried out using the complete sample data. Cronbach's α for the elementary and secondary school scale as a whole was 0.948 for reliability, with α values ranging from 0.870 to 0.938 for each dimension. All internal consistency coefficients meet the standard threshold proposed by instrument research scholars [16]. Cronbach's α for the entire higher education scale was 0.936, with α values ranging from 0.810 to 0.940, all of which were higher above the acceptable cutoff of 0.700. Each dimension's composite reliability (CR) for composite validity was greater than 0.830, surpassing the suggested threshold of 0.700. Each dimension's extracted average variance was greater than 0.570, surpassing the crucial criterion of 0.500 [17]. Each dimension's extracted average variance was greater than 0.570, surpassing the crucial criterion of 0.500. Every observed variable's standardized factor loading on its matching factor was greater than 0.500. The Fornell-Larcker criterion was used to test discriminant validity. Each first-order latent variable's square roots of average variance extracted (AVE) were higher than correlations with other latent variables, suggesting strong discriminant validity. Both scales have outstanding psychometric properties based on thorough data. The final versions of both scales have a clear, stable structure and 22 items that are evenly dispersed among dimensions.

In order to analyze group differences across demographic variables like gender, age, years of teaching experience, educational attainment, teaching level or subject category, job type, professional rank, school type, and school location, one-way analysis of variance (ANOVA) and independent-samples t-tests were used in addition to Tukey's or Games-Howell post hoc tests. Two-way ANOVA and stratified regression were employed for cross-validation to elucidate the independent impacts of age and years of experience in order to address the substantial multicollinearity between these two variables.

***Findings from the group difference analysis***

Teachers in primary and secondary schools had mean AI literacy scores of 3.37 and 3.47, respectively, which were in the upper-middle range. Samples from both educational levels showed a distinct declining tendency in terms of years of experience and age, with younger teachers outperforming middle-aged and older teachers. After adjusting for age, cross-validation using two-way ANOVA and stratified regression revealed that the main effect of years of experience was no longer significant, indicating that variations in years of experience are primarily caused by generational differences rather than simple experience accumulation. While there were no discernible differences between university instructors at different educational levels, K-12 teachers with master's degrees showed significantly higher proficiency than those with bachelor's or associate's degrees.

On the one hand, teachers in information technology showed considerably greater proficiency in terms of subject areas and work responsibilities than teachers in other disciplines. On the other hand, K-12 math and science teachers scored the lowest of all subject areas, even lower than teachers in the humanities and social sciences. While student advisors in higher education beat full-time faculty and laboratory staff in practice, teaching, and student affairs aspects, administrative roles in K-12 schools outperformed teaching roles at every level. This indicates systemic disparities in the frequency of real-world application scenarios and training resource access among positions.

Urban instructors outperformed town and rural teachers in terms of school location. Practical rather than cognitive competencies were where the urban-rural divide was most noticeable, suggesting that rural instructors may not lack a fundamental understanding of AI, but rather lack the setting and circumstances necessary to convert knowledge into useful skills.

Similar disparities in digital practice ability between urban and rural educators have been observed in blended learning research [18]. Gender differences were not statistically significant overall, but there were structural differences in several areas.

**Conversation and consequences**

***Framework design's theoretical contributions***

The development of a dual-track competency framework that is distinguished by educational stage using a value-integrated measuring paradigm is the study's primary theoretical contribution. This method achieves a theoretical breakthrough while addressing the shortcomings of current frameworks.

First, the long-ignored problem of misunderstanding between educational phases is addressed by the dual-track concept. The majority of current frameworks use generic design, ignoring the underlying distinctions in pedagogical reasoning between higher education and K-12. In order to link practical competencies to their different fundamental settings, this study created independent models for the two types of teachers. This design is substantially supported by exploratory factor analysis. The "classroom management" factor independently surfaced in the K-12 scale with a loading greater than 0.620. The "academic inquiry" item independently surfaced in the higher education scale, with a loading greater than 0.560. This confirms that stage-differentiated modeling is necessary.

Second, the conventional conundrum of separating attitudes from behavior is overcome by value-integrated design. Previous research disconnected knowledge from action by treating ethics as a distinct "attitude" element. This study developed observable behavioral indicators for safety, fairness, and benefit ethics and incorporated them into items for each scenario. According to empirical findings, there is a moderate link between the second-order "principles" level and the "application scenarios" component, with correlation values ranging from 0.490 to 0.590. This demonstrates that ethics is a set of values that are intricately linked to the use of technology rather than an add-on, offering a new framework for evaluating ethical literacy.

Third, a dual relationship between continuity and transcendence with current frameworks is established by this study. It achieves two types of advancement while integrating UNESCO's human-centered core philosophy and the AI-TPACK analytical perspective on deep integration of technology and teaching. First, it addresses the difficulty of putting macro-level frameworks into practice by converting practical skill from an abstract dimension into actionable indicators anchored in particular contexts. Second, it counteracts the propensity to marginalize ethics in framework design by elevating value principles from an autonomous component to an aspect pervading the entire process.

***The scales' psychometric benefits***

The Teachers' AI Literacy Scale for K-12 (TLCS-K12) and Teachers' AI Literacy Scale for Higher Education (TLCS-HE) scales exhibit notable psychometric benefits.

First, the study closely followed a standardized, end-to-end procedure that included a review of the literature, semi-structured interviews, two rounds of expert content validity reviews, preliminary item analysis (item-total correlation, critical ratio method, coefficient α variation), large-scale administration, EFA, CFA, and thorough reliability and validity testing. The problem of putting development ahead of validation is avoided by this theory-driven, empirically backed, and cross-validated architecture.

Second, resilience was improved by a large sample size and independent validation across educational levels. The two scales were separately validated using random sample splitting based on 18,160 K-12 and 1,508 higher education samples, guaranteeing the independence of the EFA and CFA phases and preventing overfitting. The sample size and validation methods of this work clearly demonstrate methodological advantages above those of Delcker.

Third, both measures showed outstanding psychometric qualities. Cronbach's α was 0.948, RMSEA=0.067, CFI=0.957, and the cumulative variance explained for the K-12 scale was 79.30%. Cronbach's α was 0.936, RMSEA=0.079, CFI=0.935, and cumulative variance explained was 76.90% for the higher education scale. Both discriminant validity and convergent validity (AVE>0.570, CR>0.830) satisfied suggested criteria. With cumulative variance explained at 90.70% for K-12 and 87.50% for higher education, second-order factor structure confirmed the theoretical hypothesis that first-order five-factor structure aggregates into second-order three-factor structure.

***Theoretical interpretations of group disparities***

Teachers' AI abilities varied systematically between groups, and several of these findings had significant theoretical ramifications.

The most valuable effect in theory is the age-cohort effect. There was a definite, age-dependent declining trend in both educational levels. The primary effect of teaching experience is no longer significant after adjusting for age, according to cross-validation of two-way ANOVA and stratified regression. This implies that the generational effect is the fundamental cause of the disparity in teaching experience. The advantage of younger teachers is explained by their status as "digital natives" rather than by accumulated experience. This result causes generational transition theory to replace accumulated experience theory as the explanation.

Subject-specific variations demonstrate the dual impact of tool appropriateness and technology relevance. Information technology instructors did far better, as was to be expected. Contrary to expectations, teachers of the natural sciences and mathematics received the lowest scores - even lower than those of the humanities and social sciences. This is due to the fact that general-purpose AI tools are designed for language interaction and content creation, which makes it challenging to satisfy the exact requirements of natural sciences and mathematics, which place a strong emphasis on symbolic operations and logical deduction. This suggests the creation of tools tailored in a certain field.

Variations in work responsibilities are a reflection of systemic problems with incentive systems and resource distribution. Frontline teaching personnel, the group most in need of AI empowerment, scored lowest in K-12 schools, whereas administrative roles outperformed teaching positions at all levels, suggesting administrative staff having access to more training resources. Because their duties closely match AI application situations, student counselors in higher education scored better in practice, teaching, and student affairs than full-time academic and laboratory personnel. Because laboratory personnel rely on physical equipment and standardized procedures, the extent of AI integration is comparatively limited.

The availability of possibilities to use information, rather than knowledge itself, is the fundamental cause of the discrepancy between urban and rural areas. While there is a considerable disparity in practical literacy, there is no discernible variation in cognitive literacy between instructors in urban and rural areas. This suggests that rural educators lack the application scenarios and hardware environment necessary to turn their understanding of AI into useful skills. Infrastructure and scenario development should replace universal training as the focus of policy actions.

***Useful consequences***

The two scales offer accurate diagnostic instruments for education administrative divisions. To find flaws based on data, it is advised that they be integrated into provincial monitoring systems. Remedial training should be improved for middle-aged and older instructors in order to counteract generational impacts. Digital infrastructure in rural areas should be prioritized in order to overcome urban-rural inequities. For science and math, unique tools should be created to address subject-specific disparities. AI proficiency could also be used as a benchmark for professional title assessment and hiring.

The measures can identify the need for school-based professional development for school administrators. Assessment results can be used to build differentiated programs. AI-assisted home-school communication training might be given priority in schools that struggle with classroom management. Training in literature reviews and data processing can be strengthened for those who struggle with educational research. Teachers who succeed in a variety of areas might also be chosen as "seed teachers" to act as role models.

Training for teacher education institutions should move away from just technical operations and toward a curriculum that incorporates ethical norms, practical literacy, and cognitive literacy. Educational stage and subject area should be taken into consideration when designing practical modules immersed in real-world scenarios, and ethical considerations should be included in every scenario.

The assessment scale can be used by individual teachers as a foundation for self-diagnosis. Teachers should understand that increasing AI proficiency requires more proactive research and ongoing practice than on fixed characteristics like age or subject. Instructors can find ways to incorporate AI into their regular lessons and learn by doing in practical situations.

## Conclusion

### *Research findings*

The organization of teachers' AI competencies, the creation of a measurement scale, and group differences are the three main topics this study investigates. The following are the main conclusions.

First, the AI abilities of instructors are a hierarchical construct that includes value standards, practical literacy, and cognitive literacy. Certain work environments might serve as the foundation for practical literacy. Practical literacy appears in classroom management, educational research and inquiry, and instruction for elementary and secondary educators. Practical competency is demonstrated by academic investigation, student support, and classroom instruction for higher education professors. The hierarchical structural model - from first-order five-factor to second-order three-factor structure - fits data well, according to cross-validation of EFA and CFA, and theoretical presumptions are empirically validated by a sizable sample.

Second, both the TLCS-K12 and TLCS-HE scales show strong validity and reliability. The cumulative variance explained for each of the 22 items is 79.30% and 76.90%, respectively. For both, the RMSEA is less than 0.080 and the CFI values are all greater than 0.930. The overall scales have Cronbach's α of 0.948 and 0.936, respectively. The suggested criteria are satisfied by convergent and discriminant validity. For accurate diagnosis, training design, and policy evaluation, these two scales can be used as standardized assessment instruments.

Third, across demographic characteristics, teachers' AI competencies show systematic differentiation. Rather than cumulative years of teaching expertise, the "digital native effect" is the primary reason for the rigid declining trend in competencies with age. Mathematics and the natural sciences rank lowest in terms of subject distribution, showing the combined effects of tool compatibility and technology relevance. Information technology areas, on the other hand, greatly surpass others. Disparities across positions are a reflection of systemic problems with incentive systems and resource distribution. The structural imbalance of application opportunities is the fundamental cause of urban-rural inequality.

In conclusion, this study successfully creates two scenario-based scales with outstanding validity and reliability, builds a dual-track framework that is differentiated by educational stage, and identifies group differentiation traits and underlying mechanisms. As a result, it provides empirical support and scientific tools for teacher workforce development during the digital transformation of education.

### *Research restrictions*

First, the sample's regional constraints. Every contestant came from a single East Chinese province. The cross-provincial applicability of scales has not yet been empirically evaluated, and measurement invariance needs to be confirmed, despite this province's forward-thinking status as a leader in educational reform.

Second, self-reporting social desirability bias. The value standards dimension is especially vulnerable to social desirability bias, because all data come from instructors' self-evaluations. This may result in decreased measurement precision for this dimension compared to objective behavioral measurements, and it is impossible to totally avoid common-method bias.

Third, causal inference is constrained by cross-sectional design. Since all of the data is gathered at a single moment in time, it is not possible to trace the dynamic growth of competences, determine causal links, or differentiate between age and generational effects.

Fourth, there is currently no recognized normative reference system. The scale's precise diagnostic capabilities are limited because normative standards by educational stage, area, and years of teaching experience have not yet been defined, despite the completion of reliability and validity testing and group differences studies.

### *Future research directions*

Testing cross-regional measurement equivalency comes first. Multi-group CFA might test measurement invariance, setting the stage for broader use; cooperative pilot studies could be carried out with regions like Jiangsu and Guangdong.

Second, long-term monitoring to identify pathways of development. By using latent variable growth models in conjunction with longitudinal surveys conducted over time among the same cohort of teachers, it is possible to investigate developmental trajectories and influencing factors while differentiating between age and generational effects.

Third, combining data from several sources to improve objectivity. To verify the validity of self-assessments and investigate connections between subjective and objective indicators, scale data might be connected to behavioral logs from training platforms and classroom AI use records.

Fourth, validating the closed-loop intervention procedure and setting normative norms. The practical efficacy of the closed-loop process - which includes diagnosis, attribution, intervention, and re-assessment - can be validated through quasi-experimental studies based on diagnostic results. Percentile and standard score norms can be established for various subgroups, thereby facilitating the translation of research findings into educational practice.

**Funding**

This work was supported by the National Undergraduate Training Program on Innovation and Entrepreneurship (Grant No. 202610345036) and the Scholastic Undergraduate Training Program on Innovation and Entrepreneurship (Grant No. X202610345030).

**Acknowledgements**

The authors would like to show sincere thanks to those techniques who have contributed to this research.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

[1] Mikeladze, T., Meijer, P. C., Verhoeff, R. P. (2024) A comprehensive exploration of artificial intelligence competence frameworks for educators: a critical review. *European Journal of Education*, 59(3), e12663.

[2] Ouyang, F., Jiao, P. (2021) Artificial intelligence in education: the three paradigms. *Computers and Education: Artificial Intelligence*, 2, 100020.

[3] Tsioukas, K., Kostas, A. (2026) Assessing teachers' AI Literacy for lifelong learning: a systematic review and framework alignment. *Artificial Intelligence Advances in Education*, 1(1), 1-15.

[4] Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., Qiao, M. S. (2021) Conceptualizing AI literacy: an exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041.

[5] Lintner, T. (2024) A systematic review of AI literacy scales. *Npj Science of Learning*, 9(1), 50.

[6] Celik, I. (2023) Towards intelligent-TPACK: an empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, 138, 107468.

[7] Chiu, T. K. (2025) Developing intelligent-TPACK (I-TPACK) framework from unpacking AI literacy and competency: implementation strategies and future research direction. *Interactive Learning Environments*, 33(7), 4189-4192.

[8] Williams, A. (2026) An integrated AI literacy framework for higher-education teachers: a critical review and conceptual synthesis. *Cogent Education*, 13(1), 2698137.

[9] Delcker, J., Heil, J., Ifenthaler, D. (2025) Evidence-based development of an instrument for the assessment of teachers' self-perceptions of their artificial intelligence competence. *Educational Technology Research and Development*, 73(1), 115-133.

[10] Xia, Q., Yang, Y., Weng, X., Cilsalar-Sagnak, H., Cheng, W. K., Chiu, T. K. (2025) Generative artificial intelligence competency self-efficacy scale for teachers and researchers (GAICS-TR) in higher education. *Journal of Computers in Education*, 1-25.

[11] Zhong, B., Liu, X. (2025) Evaluating AI literacy of secondary students: framework and scale development. *Computers & Education*, 227, 105230.

[12] Nong, Y., Cui, J., He, Y., Zhang, P., Zhang, T. (2024) Development and validation of an AI literacy scale. *Journal of Artificial Intelligence Research*, 1(1), 17-26.

[13] Alqarni, A. (2025) Artificial intelligence-critical pedagogic: design and psychologic validation of a teacher-specific scale for enhancing critical thinking in classrooms. *Journal of Computer Assisted Learning*, 41(3), e70039.

[14] Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., Wienrich, C. (2023) MAILS-Meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change-and meta-competencies. *Computers in Human Behavior: Artificial Humans*, 1(2), 100014.

[15] Luo, S., Zou, D. (2025) K-12 English Teachers' ChatGPT-based digital literacy: scale development and validation. *European Journal of Education*, 60(4), e70273.

[16] Taber, K. S. (2018) The use of Cronbach's α when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273-1296.

[17] Dos Santos, P. M., Cirillo, M. Â. (2023) Construction of the average variance extracted index for construct validation in structural equation models with adaptive regressions. *Communications in Statistics-Simulation and Computation*, 52(4), 1639-1650.

[18] Li, Y., Cheng, H., Qin, Q. (2025) Evaluations and improvement methods of deep learning ability in blended learning. *International Journal of e-Collaboration (IJeC)*, 21(1), 1-17.