Development of a Situational Judgment Test for Assessing Digital Literacy Among Preschool Teachers

Jun Zhang[a], Pei Zhao[b *]

School of Education, Beijing City University, Beijing, China

[a]zhangjun_bcu@bcu.edu.cn, [*b]zhaop_psy@foxmail.com

Abstract. Conventional self-report instruments are susceptible to social desirability bias and response distortion, limiting their validity for assessing teacher professional competencies. The Situational Judgment Test (SJT) addresses this limitation by embedding respondents in work-relevant scenarios to elicit behavioral tendencies. The present study developed and preliminarily validated a Preschool Teacher Digital Literacy SJT grounded in China's Digital literacy of teachers (JY/T 0646—2022). A total of 405 valid responses were obtained from preschool teachers across four provinces. A theory-driven CFA approach revealed that two dimensions—digital technology knowledge and skills (D2) and digital social responsibility (D4)—were not effectively operationalized through SJT scenarios. The final three-dimensional model retaining digital awareness (D1), digital application (D3), and professional development (D5) yielded acceptable fit ($\chi^2(167) = 270.055$, CFI = 0.948, Robust RMSEA = 0.039, SRMR = 0.059) with Cronbach's $\alpha = 0.721$. Content validity was supported by Delphi consultation (Kendall's W = 0.366). These findings provide preliminary evidence of the promise and boundaries of SJT-based measurement of preschool teacher digital literacy.

Additional Keywords and Phrases: Situational Judgment Test, Preschool teacher, Questionnaire development, Digital literacy of teachers

1 Introduction

Preschool teachers' digital literacy has become a policy priority in China. The Ministry of Education promulgated the Digital literacy of teachers (JY/T 0646—2022) [9] in 2022, establishing a five-dimensional framework to guide teacher training and assessment. Yet measuring digital literacy validly remains difficult. Conventional self-report instruments are prone to social desirability bias and response distortion, limiting their ecological validity. An instrument that captures what teachers actually do in digital contexts, rather than what they report, would therefore address a genuine gap in the field.

To address these measurement challenges, the present study developed a preschool teacher digital literacy instrument grounded in Situational Judgment Test (SJT) theory.

Z. Ge et al. (eds.), *Proceedings of the 2026 6th International Conference on Education, Information Management and Service Science (EIMSS 2026)*, Atlantis Highlights in Computer Sciences 25,
https://doi.org/10.2991/978-94-6239-766-8_40

A theory-driven CFA approach was adopted, using China's Digital Literacy of Teachers standard as the a priori structural model. It should be noted at the outset that SJT is designed to elicit behavioral responses to situational cues rather than to assess declarative knowledge or normative awareness; accordingly, dimensions of digital literacy that are primarily knowledge-based or rule-governed may be more difficult to operationalize through SJT scenarios. This methodological consideration shapes both the design and interpretation of findings in the present study.

1.1 SJT Theory and Teacher Digital Literacy Frameworks

SJT has been applied to teacher professional competency assessment since Klassen and colleagues introduced it to educational research [5, 6, 7]. By presenting respondents with authentic occupational scenarios, SJT elicits behavioral tendencies rather than declarative knowledge, offering higher ecological validity than self-report measures [16, 17]. Specialized SJT variants have been developed for both primary and preschool teachers, demonstrating the approach's adaptability across educational contexts [10].

International Digital Literacy of Teachers frameworks have evolved through three generations: ICT tool integration [14], deepening instructional competence [13], and AI ethics integration [15]. China's Digital literacy of teachers (JY/T 0646—2022) synthesizes these developments into five dimensions: digital awareness (D1), knowledge and skills of digital technology (D2), digital application (D3), digital social responsibility (D4), and professional development (D5) [9]. This framework serves as the theoretical foundation for the present study.

1.2 Limitations of Existing Measurement Instruments

Multiple instruments for assessing teacher digital literacy have been developed internationally. Existing instruments can broadly be categorized into three types.

The first and most common type comprises self-report scales based on the DigCompEdu framework, such as the Digital Competency Scale for Teachers [3] and the TDC-S [1]. The second type consists of TPACK-based self-report scales, which have been criticized for leaving certain competency dimensions inadequately operationalized. The third type comprises performance-based instruments [e.g., 11], though these have been developed primarily for student rather than teacher populations.

Across these instruments, self-report assessment predominates, yet self-assessed digital literacy correlates weakly with actual performance [4] and is readily influenced by social desirability and response-style effects [12]. By presenting simulated work situations, SJT can more authentically reflect teachers' behavioral tendencies in instructional contexts, though it is inherently better suited to behavioral dispositions than to declarative knowledge or normative rule-following [16]. Systematic application of SJT to preschool teacher digital literacy nonetheless remains unexplored, and the present study represents a preliminary attempt to examine both its promise and its boundaries in this context.

2 Materials and Methods

2.1 Participants

A total of 580 questionnaires were distributed to preschool teachers in Beijing, Shandong, Sichuan, and Guangdong provinces of mainland China. Prior to analysis, data cleaning was conducted using two criteria: (1) responses with a total completion time of less than 600 seconds were excluded to eliminate perfunctory responding; and (2) records in which all item responses were identical (i.e., straight-line responding) were removed. After applying these screening criteria, 405 valid questionnaires were retained for analysis. The demographic characteristics of the sample are presented in Table 1. The sample was predominantly female (97.3%) and highly educated (98.2% holding an associate degree or above). This profile closely mirrors that of the Chinese preschool teaching workforce rather than reflecting sampling bias: according to the Educational Statistics Yearbook of China 2024 [2], 97.64% of preschool full-time teachers nationwide are female and 94.56% hold an associate degree or above. The sample is therefore broadly representative of the target population on these characteristics.

Table 1. Demographic characteristics of the participant sample (N = 405).

Variable	Category	n	%
Gender	Male	11	2.7
	Female	394	97.3
Age	< 18	0	0
	18–25	65	16.1
	26–30	123	30.4
	31–40	166	41.0
	41–50	47	11.6
	51–60	3	0.7
	> 60	1	0.2
Position	Lead teacher	311	76.8
	Assistant teacher	24	5.9
	Middle management	11	2.7
	Principal/Vice-principal	57	14.1
	Other staff	2	0.5
Education	Master's or above	9	2.2
	Bachelor's	286	70.6
	Associate	103	25.5
	High school or below	7	1.7
Title	Senior	0	0
	Advanced	26	6.4
	Level 1	85	21.0
	Level 2	163	40.3
	Level 3	32	7.9
	Unranked	99	24.4
Experience	< 3 years	63	15.6
	3–10 years	206	50.9
	11–20 years	110	27.2
	≥ 20 years	26	6.4

2.2 Instrument Development

Step 1 – Collection and identification of typical scenarios. Three backbone preschool teachers with over ten years of practical experience and background in digital education were selected to construct scenario events. Following training in the Digital Literacy of Teachers standard and SJT theory, they collected scenarios through interviews and written descriptions. Through research team discussion, 33 scenario events were selected with one-to-one correspondence to the standard's third-level indicators.

The Delphi method was then employed to evaluate the representativeness and typicality of the 33 scenarios. A panel of 15 experts—preschool teachers, teaching research specialists, and university faculty—rated each scenario's typicality on a 10-point scale (higher scores indicating greater alignment with the Teacher Digital Literacy indicators).

Step 2 – Determination of response options. Four response options were developed for each scenario, all conforming to professional ethics and preschool operational guidelines, distinguished by whether and to what degree the response reflected teacher digital literacy. Response options were reviewed in two rounds by the 15-member expert panel.

Step 3 – Determination of scoring criteria. A mixed-method approach was used. Behaviorally Anchored Rating Scales (BARS) were first applied to arrange key behaviors from best to worst: Response A (optimal), B (second-best), C (third-best), D (least preferred). Expert ratings (1–10 scale) were then used to validate the reasonableness of the preset rankings.

Step 4 – Determination of instructions. Behavioral tendency-oriented instructions were adopted. Each scenario item presented respondents with two questions: (1) "I believe the optimal approach is:" (knowledge-oriented) and (2) "I would actually do:" (behavioral tendency-oriented). Although both questions were included in the administered instrument to facilitate ecological engagement with each scenario, only the responses to the behavioral tendency question (Question 2) were used in all statistical analyses, in order to assess teachers' actual behavioral dispositions rather than their normative knowledge. This choice reflects a deliberate methodological commitment: behavioral-tendency instructions in SJTs are theorized to better capture typical performance and to be less susceptible to social desirability bias than knowledge-based instructions [8, 16]. The behavioral-tendency responses were therefore the construct-relevant data for the present aim of assessing digital literacy as enacted in practice. A systematic comparison of the two question types—for example, examining the gap between what teachers regard as optimal and what they would actually do—addresses a distinct research question that lies beyond the scope of this instrument-development study and is noted below as a direction for future research.

Following the above steps, a preliminary Preschool Teacher Digital Literacy SJT was developed. A sample item reads: 'I would use digital technology to support my teaching in the following way. Teacher A: Use an AI robot as a teaching assistant, controlling electronic screens through AI agent technology. Teacher B: Use a smart speaker to play audio, an interactive whiteboard to display animations and images, and other devices as needed during the lesson. Teacher C: Embed all teaching resources into a

Seewo Whiteboard presentation in the order of instructional steps, to be played sequentially. Teacher D: Use presentation software to display images and video content.' Respondents were then asked to indicate: (1) I believe the optimal approach is: (2) I would actually do:

2.3 Instrument Validation

After finalizing all items, the questionnaire was distributed via the Wenjuanxing (SurveyStar) platform with response options presented in randomized order to prevent perfunctory responding. Following distribution of 580 questionnaires and data cleaning, 405 valid questionnaires served as the basis for CFA and reliability testing.

2.4 Statistical Analysis

Content validity was assessed using the Delphi method. Descriptive statistics (mean, median, SD, CV) were computed in Excel; Kendall's W and chi-square tests were conducted using the irr package in R. Item-level (I-CVI) and scale-level (S-CVI/Ave and S-CVI/UA) content validity indices were computed for both Delphi rounds, with expert ratings ≥ 7 on the 10-point typicality scale treated as indicating content relevance.

A theory-driven CFA approach was adopted, using the five-dimensional structure of China's Teacher Digital Literacy standard as the a priori measurement model. Given that the theoretical framework was specified in advance, EFA was not conducted; CFA was deemed more appropriate for evaluating the degree to which the data conform to the theoretically derived structure.

CFA was conducted using the lavaan package in R with the DWLS estimator and WLSMV correction, appropriate for ordered categorical data. Model fit was evaluated using CFI and TLI (acceptable ≥ 0.90), Robust RMSEA (acceptable < 0.08), and SRMR (acceptable < 0.08). Items with non-significant loadings ($p > .05$) or standardized loadings below 0.30 were candidates for removal following standard scale purification procedures. Both the initial and revised models are reported. In addition, modification indices (MIs) were examined to assess potential sources of local misfit, although—consistent with the theory-driven approach—the model was not modified on the basis of these indices.

Reliability of the revised three-dimensional scale was assessed using Cronbach's α and mean inter-item correlation, computed using the psych package in R.

3 Results

3.1 Content Validity

In Round 1, items with a mean or median below 8.0, or a CV above 0.25, were flagged for revision; Item 27 was identified as requiring priority attention (CV = 0.27). Following revision of the scenario wording, Round 2 results showed that medians for all 33

items exceeded 9.0 and all CV values fell below 0.25. Item 27 showed continued improvement (Mdn = 9.5, CV = 0.16) but its mean (M = 8.88) remained marginally below the 9.0 threshold, suggesting that some residual ambiguity in the scenario wording persisted. Item-level and scale-level content validity indices were also computed, with ratings ≥ 7 on the 10-point scale treated as indicating content relevance (see Tables 2 and 3). In both rounds, all 33 items met the I-CVI ≥ .78 criterion, and the scale-level index improved from Round 1 to Round 2 (S-CVI/Ave = .970 and .989, respectively). The consistently high values indicate strong expert endorsement of the content relevance of every scenario, and the improvement across rounds further demonstrates the effectiveness of the iterative revision process.

Table 2. Descriptive statistics of scenario typicality ratings — Round 1 and Round 2 Delphi Survey.

Item	Round 1					Round 2				
	M	Mdn	SD	CV	I-CVI	M	Mdn	SD	CV	I-CVI
1	8.97	10.00	1.49	0.17	0.933	9.38	10	0.92	0.10	1.000
2	8.27	8.00	1.53	0.19	0.867	9.25	10	1.16	0.13	1.000
3	8.87	9.00	1.19	0.13	0.933	9.13	10	2.10	0.23	0.875
4	9.23	9.50	0.94	0.10	1.000	9.75	10	0.46	0.05	1.000
5	9.00	9.00	1.31	0.15	0.933	9.38	10	1.06	0.11	1.000
6	9.27	9.00	0.80	0.09	1.000	9.50	10	0.76	0.08	1.000
7	9.20	10.00	1.01	0.11	1.000	9.50	10	0.76	0.08	1.000
8	9.13	9.00	1.06	0.12	1.000	9.75	10	0.46	0.05	1.000
9	9.33	10.00	0.82	0.09	1.000	9.75	10	0.46	0.05	1.000
10	8.97	9.00	1.23	0.14	0.933	9.63	10	0.52	0.05	1.000
11	9.33	10.00	0.82	0.09	1.000	9.63	10	0.74	0.08	1.000
12	9.47	10.00	0.64	0.07	1.000	9.75	10	0.46	0.05	1.000
13	9.60	10.00	0.63	0.07	1.000	9.88	10	0.35	0.04	1.000
14	9.53	10.00	0.64	0.07	1.000	9.88	10	0.35	0.04	1.000
15	9.40	10.00	0.83	0.09	1.000	9.50	10	0.76	0.08	1.000
16	8.67	9.00	1.80	0.21	0.867	9.75	10	0.46	0.05	1.000
17	9.33	10.00	0.98	0.10	1.000	9.13	9	0.83	0.09	1.000
18	8.87	9.00	1.36	0.15	0.933	9.63	10	0.74	0.08	1.000
19	9.47	10.00	0.74	0.08	1.000	9.25	10	1.16	0.13	1.000
20	9.53	10.00	0.64	0.07	1.000	9.38	9.5	0.74	0.08	1.000
21	9.47	10.00	0.74	0.08	1.000	9.50	10	0.76	0.08	1.000
22	9.47	10.00	0.74	0.08	1.000	9.38	9.5	0.74	0.08	1.000
23	9.60	10.00	0.51	0.05	1.000	9.88	10	0.35	0.04	1.000
24	9.40	10.00	0.91	0.10	1.000	9.63	10	0.52	0.05	1.000
25	9.47	10.00	0.74	0.08	1.000	9.75	10	0.46	0.05	1.000
26	9.27	10.00	1.33	0.14	0.933	9.13	9.5	1.36	0.15	0.875
27	8.60	10.00	2.35	0.27	0.933	8.88	9.5	1.46	0.16	0.875
28	8.60	9.00	1.68	0.20	0.867	9.63	10	0.74	0.08	1.000
29	8.93	9.00	1.44	0.16	0.933	9.75	10	0.46	0.05	1.000
30	9.20	10.00	0.94	0.10	1.000	9.71	10	0.49	0.05	1.000
31	9.53	10.00	0.64	0.07	1.000	9.63	10	0.52	0.05	1.000
32	9.60	10.00	0.51	0.05	1.000	9.25	10	1.16	0.13	1.000
33	8.93	9.00	1.16	0.13	0.933	9.63	10	0.74	0.08	1.000

Table 3 presents the Kendall's W coefficients from both rounds of the Delphi consultation. Round 1 results yielded W = 0.145, $\chi^2(31) = 41.794$, p = .115, indicating very low overall concordance that did not reach statistical significance. Following revision of the scenario bank, Round 2 results showed W = 0.366, $\chi^2(31) = 84.500$, p < .001, indicating that expert agreement reached statistical significance. According to the interpretive guidelines for Kendall's W, values in the range $0.3 \leq W < 0.5$ reflect a low-to-moderate level of concordance. The present W = 0.366 falls within this range, suggesting that the expert panel had reached partial consensus on the core content of the instrument while some divergence of opinion remained. The moderate level of agreement is attributable to reasonable cognitive divergence arising from the diverse professional backgrounds of the expert panel (preschool teachers, teaching research specialists, university faculty), which in turn reflects the cross-domain coverage of the test content.

Table 3. Expert agreement test on scenario typicality — Round 1 and Round 2 Delphi Survey.

Round	Kendall's W	Chi-square (χ^2)	p-value	S-CVI/Ave	S-CVI/UA
Round 1	0.145	41.794	.115	0.970	0.636
Round 2	0.366	84.500***	< .001	0.989	0.909

3.2 Confirmatory Factor Analysis

A theory-driven CFA was used to examine whether the five-dimensional structure of the Digital Literacy of Teachers standard was supported by the data. An initial five-dimensional model comprising all 33 items was estimated first. Inspection of the parameter estimates revealed systematic problems with two dimensions: the D2 (digital technology knowledge and skills) items produced consistently low factor loadings (Q14 = 0.250, Q28 = 0.520, Q31 = 0.188) accompanied by inter-factor correlations exceeding 1.0, indicating a Heywood case and suggesting that D2 could not be identified as a stable latent construct in the present data; the D4 (digital social responsibility) items produced predominantly non-significant or near-zero loadings (Q12 = −0.077, p = .183; Q13 = −0.098, p = .133; Q18 = −0.001, p = .983; Q33 = −0.038, p = .594), with only Q11 and Q29 reaching statistical significance. These patterns indicated that D2 and D4 did not function as coherent measurable dimensions under the SJT format, likely because the content of both dimensions is primarily knowledge-based and normative in character, leaving insufficient behavioral variance for SJT-based measurement. Following scale purification, all D2 items (Q14, Q28, Q31) and all D4 items (Q11, Q12, Q13, Q18, Q29, Q33) were therefore removed. In addition, four items from D1 and D3 were removed due to low or non-significant factor loadings: Q1 (D1, $\lambda = 0.180$), Q4 (D3, $\lambda = 0.088$, p = .160), Q5 (D3, $\lambda = -0.056$, p = .384), and Q10 (D3, $\lambda = 0.137$, p = .058). The resulting revised model retained 20 items across three dimensions. Fit indices for both the initial and revised models are presented in Table 4.

Table 4. CFA model fit indices: initial five-dimensional model versus revised three-dimensional model.

Index	Model 1 (5-factor) Raw	Model 1 Robust	Model 2 (3-factor) Raw	Model 2 Robust
χ^2 (df)	760.291 (485)	735.119 (485)	241.698 (167)	270.055 (167)
p	—	0.000	0.000	0.000
CFI	0.887	0.815	0.948	0.893
TLI	0.877	0.799	0.940	0.878
RMSEA [90% CI]	0.037 [0.032, 0.043]	0.036 [0.030, 0.041]	0.033 [0.024, 0.042]	0.039 [0.030, 0.047]
Robust RMSEA [90% CI]	—	0.053 [0.047, 0.059]	—	0.050 [0.039, 0.060]
SRMR	0.068	0.068	0.059	0.059

Note. Model 1 = initial five-dimensional model (33 items, D1–D5). Model 2 = revised three-dimensional model (20 items): D1 (Q2, Q3, Q8, Q24), D3 (Q6, Q7, Q9, Q15–Q17, Q19, Q21–Q23, Q27), D5 (Q20, Q25, Q26, Q30, Q32). Raw = DWLS; Robust = WLSMV correction.

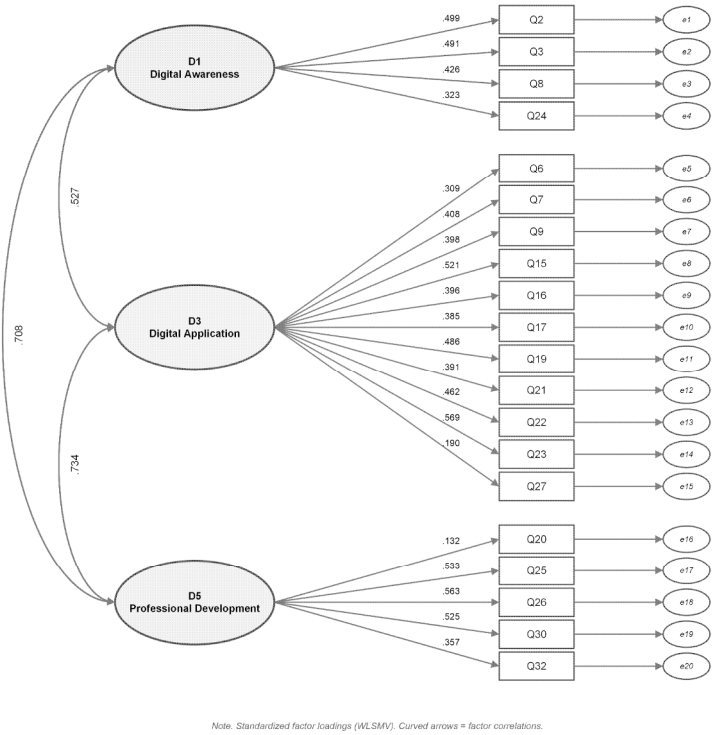

Fig. 1. CFA path diagram of the revised three-dimensional SJT model. Standardized factor loadings (WLSMV) shown.

Figure 1 presents the CFA path diagram of the revised three-dimensional model. Standardized factor loadings ranged from 0.132 (Q20, D5) to 0.569 (Q23, D3), with D3

items generally showing the most consistent loadings across the scale. The two lowest-loading items were Q20 within D5 and Q27 within D3, both of which nonetheless reached statistical significance. Curved arrows indicate factor inter-correlations (D1–D3 = 0.527, D1–D5 = 0.708, D3–D5 = 0.734), reflecting the theoretically expected pattern of related but distinguishable dimensions.

Modification indices (MIs) were inspected to assess local misfit. The MIs were moderate in magnitude (largest MI = 25.54) and indicated no severe misspecification, given that the model already showed acceptable fit. The largest MIs corresponded to cross-loadings of individual items on non-target factors (e.g., Q15 on D5, MI = 25.54; Q15 on D1, MI = 19.36) and to a small number of correlated residuals among items within the same dimension (e.g., Q19 with Q23, MI = 19.50; Q2 with Q3, MI = 16.56). These patterns are consistent with the integrative nature of SJT scenarios, in which a single realistic situation often draws simultaneously on competencies from more than one dimension. In keeping with the theory-driven approach, the model was not modified to accommodate these indices.

3.3 Reliability

Internal consistency reliability of the revised 20-item three-dimensional scale was assessed; results are presented in Table 5. Cronbach's α for the total scale was 0.721, meeting the commonly accepted minimum threshold of 0.70. The mean inter-item correlation was 0.115, which is low relative to conventional scales but consistent with the contextual heterogeneity inherent in SJT item design—each item represents a distinct instructional scenario, so low inter-item correlations are expected rather than indicative of poor reliability [8]. These findings suggest that the revised scale has acceptable, if modest, internal consistency for a preliminary SJT instrument. As shown in Table 5, the dimension-level Cronbach's α values were modest and uneven. This pattern is consistent with, rather than contrary to, the nature of SJT measurement: because each scenario constitutes an independent behavioral sample that taps somewhat different competencies, within-dimension homogeneity is expected to be limited, and Cronbach's α—which assumes item homogeneity—is not an ideal reliability index for SJT-based scores. The uneven dimension-level alphas therefore reinforce, rather than undermine, the interpretation of SJT scenarios as inherently heterogeneous.

Table 5. Reliability analysis of the revised three-dimensional Situational Judgment Test (20 items).

Index	Value
Cronbach's Alpha	0.721
Dimension D1 (digital awareness)	0.386
Dimension D3 (digital application)	0.635
Dimension D5 (professional development)	0.419
Standardized Alpha	0.722
Number of items	20
Sample size	405
Mean inter-item correlation	0.115

4 Discussion

The present study set out to develop and validate an SJT for assessing preschool teacher digital literacy in the context of China's national Digital Literacy of Teachers standard (JY/T 0646—2022). The results both support the feasibility of this approach and clarify where its limitations lie. The following discussion addresses each of these points in turn.

4.1 Why D2 and D4 Were Not Retained: SJT Measurement Logic

The failure of D2 and D4 to form coherent latent constructs is most parsimoniously explained by SJT's fundamental measurement logic. SJT assesses behavioral tendencies in response to situational cues—what people are inclined to do, not what they know or believe they should do [16]. D2 items primarily assess whether teachers know correct technical concepts and tool selection strategies; D4 items assess awareness of and compliance with legal and ethical norms. Both dimensions are knowledge-based and normative. Generally, there is one distinctively correct answer, and the majority of teachers are aware of it. SJT scenarios constructed around such content give rise to ceiling effects and near-zero inter-item covariance, which is precisely the pattern observed in the present data. The implication is not that these dimensions are unimportant, but that SJT—as a behavioral tendency measure—is the wrong tool for assessing them. Alternative instruments—knowledge tests for D2, vignette-based ethical reasoning tasks for D4—may be more appropriate and could complement the SJT in a multi-method battery.

4.2 The Revised Three-Dimensional Structure

The three retained dimensions share a common feature: each manifests in observable, context-specific behavioral patterns that differ meaningfully across individual teachers. Digital awareness concerns teachers' motivated engagement with and orientation toward digital tools in professional contexts. Digital application covers how teachers orchestrate digital resources across instructional design, classroom delivery, assessment, and home-school communication. Professional development concerns the degree to which teachers draw on digital resources for ongoing reflection, inquiry, and career learning. Because each dimension produces behaviorally differentiated responses in scenario-based tasks, all three are well suited to SJT measurement.

Factor inter-correlations (D1–D3 = 0.527, D1–D5 = 0.708, D3–D5 = 0.734) indicate the three dimensions are related but distinguishable, consistent with the Digital literacy of teachers conceptualization of these as interrelated rather than orthogonal dimensions [9].

The substantial factor inter-correlations raise the question of whether alternative model specifications might better represent the data. Two possibilities warrant comment. A higher-order factor model is not identified with only three first-order factors, as a minimum of four is required; it therefore could not be estimated for the present

three-dimensional structure. A bifactor model, although conceptually attractive, imposes considerable demands on item count and sample size and is prone to improper solutions with the present 20 items and sample of 405; moreover, fitting such a model would depart from the study's theory-driven aim of validating the nationally specified dimensional framework rather than searching for an alternative structure. For these reasons, and consistent with the rationale for not conducting EFA, we retained the theoretically specified model and did not pursue data-driven respecification. Examining higher-order and bifactor structures with larger samples represents a worthwhile direction for future research.

4.3 Content Validity

Two rounds of Delphi consultation provided preliminary support for content validity. Kendall's W increased from 0.145 (Round 1) to 0.366 (Round 2, $\chi^2 = 84.500$, $p < .001$). The moderate final concordance level reflects reasonable cognitive heterogeneity among experts from different professional backgrounds rather than insufficient content validity—itself evidence that the scenario bank spans multiple professional domains. Although the final Kendall's W was moderate, the item- and scale-level content validity indices were uniformly high (S-CVI/Ave = .989 in Round 2), indicating that experts strongly endorsed the items as content-relevant even while differing somewhat in their precise typicality rankings. The two indices thus capture complementary facets of content validity: agreement in rank ordering (Kendall's W) versus endorsement of relevance (CVI).

4.4 Limitations and Future Directions

Several limitations of the present study should be acknowledged. The instrument covers only three of the five dimensions specified in the Digital Literacy of Teachers standard, and criterion-related validity against observed classroom practice has not yet been examined. Future work should prioritize testing whether SJT scores predict actual technology integration behavior, developing complementary assessment tools for D2 and D4 so that the full construct can be covered, and replicating the findings across more geographically and demographically diverse samples. Until such evidence is available, the instrument should not be used for individual-level diagnosis or high-stakes decision-making. In addition, although the sample is broadly representative of the national workforce in terms of gender and education (see Section on Participants), the very small numbers of male teachers (n = 11) and teachers with high-school-or-below qualifications (n = 7) precluded formal tests of measurement invariance and differential item functioning across these subgroups. The generalizability of the factor structure and item functioning to these underrepresented groups therefore cannot be assumed, and establishing measurement invariance with more balanced samples is an important direction for future research.

5 Conclusion

The present study developed and conducted preliminary validation of a Preschool Teacher Digital Literacy SJT grounded in China's Digital Literacy of Teachers standard. The study contributes to the literature in three respects.

First, SJT can yield interpretable measurement of preschool teacher digital literacy for behaviorally observable dimensions. The revised three-dimensional model (D1, D3, D5) achieved acceptable fit with all loadings statistically significant, providing preliminary evidence of structural validity.

Second, two dimensions—digital technology knowledge and skills (D2) and digital social responsibility (D4)—were not effectively captured by the SJT format. This finding delineates the boundaries of SJT-based measurement and points toward the need for complementary assessment approaches.

Third, content validity was supported through expert consultation, with Kendall's W improving from 0.145 to 0.366 across two Delphi rounds. The instrument represents a promising but preliminary research tool. Criterion validity, item revision, and replication in diverse samples are needed before wider application.

Acknowledgments

This study was supported by the Beijing Educational Science Planning General Project (Grant No. CDGB24275). The authors declare no conflicts of interest.

References

1. Aydin MK, Yildirim T and Kus M. 2024. Teachers' digital competences: A scale construction and validation study. Front. Psychol. 15, 1356573. https://doi.org/10.3389/fpsyg.2024.1356573
2. Department of Development Planning, Ministry of Education of the People's Republic of China. 2024. Educational Statistics Yearbook of China 2024. China Statistics Press, Beijing.
3. Gümüş MM and Kukul V. 2023. Developing a digital competence scale for teachers: validity and reliability study. Educ. Inf. Technol. 28, 3, 2747–2765. https://doi.org/10.1007/s10639-022-11213-2
4. Hatlevik, O.E., Throndsen, I., Loi, M., Gudmundsdottir, G.B. 2018. Students' ICT self-efficacy and computer and information literacy: Determinants and relationships. Computers & Education 118, 107–119.https://doi.org/10.1016/j.compedu.2017.11.011
5. Klassen RM, Durksen T, Rowett E and Patterson F. 2014. Applicant reactions to a situational judgment test used for selection into initial teacher training. Int. J. Educ. Psychol. 3, 2, 104–124.
6. Klassen, Robert, Durksen, Tracy, Kim, Lisa, Patterson, Fiona, Morley, Emma, Warwick, Jane, Warwick, Paul, and Wolpert, Mary. 2016. Developing a Proof-of-Concept Selection Test for Entry into Primary Teacher Education Programs. International Journal of Assessment Tools in Education 4, 96–114.https://doi.org/10.21449/ijate.275772
7. Klassen RM and Kim LE. 2021. Situational judgment tests and their use for teacher selection. In Teacher Selection: Evidence-Based Practices. Springer, Cham.

8. McDaniel MA and Nguyen NT. 2001. Situational judgment tests: A review of practice and constructs assessed. Int. J. Sel. Assess. 9, 1–2, 103–113. https://doi.org/10.1111/1468-2389.00167
9. Ministry of Education of China. 2022. Digital literacy of teachers (JY/T 0646—2022). MoE, Beijing.
10. Nadmilail MS et al. 2022. Measurement of non-academic attributes in the situational judgment test as part of school teacher selection: A systematic literature review. International Journal of Learning, Teaching and Educational Research 21, 5, 1–18. https://doi.org/10.26803/ijlter.21.5.14
11. Pan Q, Reichert F, Liang Q, et al. 2025. Measuring digital literacy across ages and over time. Educ. Inf. Technol. 30, 22065–22100.https://doi.org/10.1007/s10639-025-13592-8
12. Paulhus, Delroy. 1984. Two-component models of socially desirable responding. Journal of Personality and Social Psychology 46, 598–609. https://doi.org/10.1037/0022-3514.46.3.598
13. Redecker C. 2017. European Framework for the Digital Competence of Educators: DigCompEdu. European Commission, Joint Research Centre, Luxembourg.
14. UNESCO. 2018. UNESCO ICT Competency Framework for Teachers, Version 3. UNESCO, Paris.
15. UNESCO. 2024. AI Competency Framework for Teachers. UNESCO, Paris.
16. Weekley JA and Ployhart RE (Eds.). 2013. Situational Judgment Tests: Theory, Measurement, and Application. Psychology Press, New York.
17. Whetzel DL and McDaniel MA. 2009. Situational judgment tests: An overview of current research. Hum. Resour. Manag. Rev. 19, 3, 188–202. https://doi.org/10.1016/j.hrmr.2009.01.002