International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Construction and Psychometric Standardization of the Micro Thinking Skills Test (MTST): A Validated Diagnostic Instrument for Middle Stage School Students in Gujarat, India

Dr. Pranav Desai[1] Dr. Mayur Parmar[2]

[1]Associate Professor (Research Cadre), Parul Institute of Management and Research (PIMR), Faculty of Management Studies, Parul University, Vadodara – Gujarat – India || pranav.desai40881@paruluniversity.ac.in

[21]I/C Principal, H M Patel Institute of English Training & Research, Vallabh Vidyanagar Dist- Anand

https://doi.org/10.55041/ijsmt.v2i5.385

Cite this Article: Desai, P. (2026). Construction and Psychometric Standardization of the Micro Thinking Skills Test (MTST): A Validated Diagnostic Instrument for Middle Stage School Students in Gujarat, India. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.385

ABSTRACT

Background: Middle school (Grades 6-8, ages 11-14) is a cognitively critical time in which students move from concrete to formal operational thinking. Despite the explicit instructions given in the National Education Policy 2020 (NEP 2020) for competency based evaluation, there was no psychometrically standardized instrument to measure micro level thinking sub skills at the discrete level in this population in the regional context of Gujarat.

Objectives: To validate the Micro Thinking Skills Test (MTST) a diagnostic instrument in six subscales that has been constructed, developed, evaluated for validity and reliability, and normed.Objective: The Micro Thinking Skills Test (MTST) is a diagnostic instrument in six subscales that has been constructed, developed, validity and reliability tested, and normed.

Methods: The methods used were the nine phase psychometric standardization design. Lawshe's Content Validity Ratio (CVR) with a panel of 12 experts was used for expert content validation. The final item selection was determined by Classical Test Theory (CTT) item analysis. The standardization sample consisted N = 1,200 students (400 students in each grade) selected by Stratified Multi-Stage Cluster Sampling from 35 schools of the study area Anand and Kheda districts in Gujarat. Content (CVI), construct (Exploratory Factor Analysis), and criterion approaches were used to evaluate the validity. Cronbach's alpha, split-half by Spearman-Brown, and two-week test-retest coefficient were used to test for reliability.

Results: The final 72-item MTST exhibited good psychometric characteristics with high overall CVI = 0.87 (content validity), a six-factor EFA solution that explained 60.07% of the total variance, and an $r = 0.61$ ($p < .001$) KMO = 0.89 (construct validity) with teacher performance ratings (criterion validity). Internal consistency was very high (Cronbach's $\alpha = 0.93$) and test-retest reliability, $r = 0.87$ ($p < .001$, $n = 100$, 2-week interval). Mean scores increased significantly across grades (Grade 6: M = 38.72; Grade 7: M = 42.15; Grade 8: M = 45.89; $F(2,1197) = 156.84$, $p < .001$, $\eta^2 = 0.208$), confirming developmental sensitivity. There were no significant differences between genders ($p = .216$). Percentile, stanine and t-scores for each grade were developed.

Conclusions: The MTST is psychometrically sound, meeting international standards of educational diagnostic instruments. It addresses a documented gap in the Indian assessment landscape by offering a valid, reliable and regionally normed sub-skill diagnostic assessment tool for middle school students that facilitates the implementation of the NEP 2020 competency-based assessment goals.

Keywords: Micro thinking skills, test standardization, psychometric validation, assessment in middle school, exploratory factor analysis, NEP 2020, Gujarat, Content Validity Index, diagnostic instrument.

1. INTRODUCTION

Higher order thinking skills are a key objective of education policy in the early adolescence in most contemporary education systems around the world. The National Education Policy 2020 (MEA, 2020) and the National Curriculum Framework 2005 (NCERT, 2005) clearly outline competency-based assessment frameworks in India, which move beyond rote learning into measuring reasoning, inferences, analysis and evaluation. There is a critical period in this developmental continuum that is the middle school stage (Grades 6 through 8, ages 11 to 14). This is a developmental stage that is based on Piaget's theory (1952), where the process from concrete operational to formal operational thought takes place, and the ability of abstract inference, systematic generalisation and evaluative judgment are developed. Vygotsky's sociocultural approach (1978) also draws attention to the importance of organised and scaffolded learning experiences which can speed up development of higher psychological functions over this period.

However, an extensive review of existing assessment tools, both national and international, revealed a multi-dimensional assessment gap for this population in the Indian context. The standardized tests that are already available either focus on general intelligence as opposed to specific sub-skills of thinking or are developed for secondary/college level students, do not have normative data from India, or measure thinking skills as undifferentiated composites that cannot provide sub-skill diagnostic reports (Sharma, 1990; Singh, 2015; Watson & Glaser, 1980). No instrument has taken on the challenge to address the micro level sub-skill specificity, Grade 6-8 developmental calibration, Gujarat norming for the norms, the Gujarati language availability, and contemporary psychometric standards at the same time.

The gap was documented and is multidimensional, and thus, the Micro Thinking Skills Test (MTST) was designed and developed to address this gap. The present paper documents the entire process of construction, validation and the standardization, including evidence for content, construct, and criterion validity, internal consistency, split-half, test-retest reliability, sensitivity to developmental grade-level differences, and gender equity, as well as extensive normative tables for grades 6, 7, and 8, presented separately.

1.1 Theoretical Framework

The conceptual architecture for the MTST is based on the cognitive process levels (2-4) of Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001): Understanding, Applying, and Analysing. The six micro thinking sub-skills are operationally defined as cognitive processes that are discrete, separately measurable: Observation (accurate detail identification; Level 2); Comparison (systematic similarity-difference analysis; Level 2); Classification (attribute-based categorical grouping; Level 3); Inference (logical conclusion-drawing from evidence; Level 3); Generalisation (principle formation from specific instances; Level 4); and Evaluation (quality judgment using criteria; Level 4). The MTST is a multi-level hierarchical architecture that differs from instruments collapsing reasoning dimensions into composites and thus ignoring the hierarchical structure.

2. METHOD

2.1 Research Design

The Survey Research Design with Psychometric Standardization was used, which is the standard method used in educational test construction and the development of norms (Anastasi & Urbina, 1997; Gregory, 2013). The design includes qualitative procedures (item writing, expert validation) in early design stages and quantitative procedures (item analysis, factor analysis, reliability estimation, norm computation) during standardization stages. A total of nine sequential phases were followed in completing the whole process from conceptual framework to publication of the norms.

2.2 Sample

The target population is all students of grade 6-8 from secondary schools in Anand and Kheda districts of Gujarat (total of 48000 students in 217 registered schools). The standardized sample (N = 1200; 400 each grade) was selected using

Stratified Multi-Stage Cluster Sampling technique where stratification variables were district, type of school (Government, Grant-In-Aid, Private), location (Urban, Semi-Urban, Rural), medium of instruction (Gujarati, English) and class level. For schools within each stratum, Simple Random Sampling was used while in each school, all sections were included as whole clusters. The gender distribution was roughly even with 51% male (n = 612) and 49% female (n = 588). The data were collected in 35 schools from January to March 2025 with a strictly standardized nine-step administration protocol.

2.3 Instrument Development

The items were developed systematically and systematically developed the initial pool of 120 items (20 per sub-skill) in both gujarati and english. The comprehensibility was set up and confirmed by a pre-pilot administration (N = 30) with a window of administration was set at 55 minutes. Content validation was done by the 12-member expert panel comprising educational psychologists (4), experienced grade 6-8 teachers (4), Curriculum specialists (2) and Child development experts (2) through the formula for Content Validity Ratio (CVR) developed by Lawshe (1975). Items below a CVR of 0.56 (the minimum value for n = 12 with $\alpha = .05$) were revised or dropped, resulting in a 96 item pool that had an overall CVI of 0.84.

The Classical Test Theory item analysis was used in pilot testing (n = 200), namely Item Facilities Index (IF, target range 0.30-0.70, minimum 0.20), Discrimination Index (D, minimum 0.20) and Item-Total correlation (r, minimum 0.20). If an item failed any of the above requirements then it was eliminated. The final MTST consists of 72 items (12 per sub-skill), with four different formats (58% 4-option multiple choice, 19% matching, 11% sentence completion, and 11% scenario-based identification).

2.4 Statistical Analysis

IBM SPSS v26 and Microsoft Excel software was used for all analyses. Normality was tested by Kolmogorov-Smirnov and Shapiro-Wilk test. Content validity was assessed using item level CVRs and overall CVIs were calculated. Construct validity was evaluated by the Exploratory Factor Analysis (EFA) with Principal Components Analysis (PCA) Varimax rotation and data adequacy was determined by Kaiser-Meyer-Olkin and Bartlett's Test of Sphericity. The criterion validity was derived from the Pearson's r for the scores total for MTST and the teacher rating of performance (n = 300, 5-point scale, concurrently) within 1 week. Cronbach's alpha, the split-half correlation (corrected using a Spearman-Brown formula) and the Pearson's correlation over a two-week test-retest period (n = 100) were used to assess reliability. Independent-samples t-tests were used for comparing two groups and one-way ANOVA and Tukey HSD post-hoc tests were used for comparing three groups; effect sizes were calculated as Cohen's d and partial eta-squared (η^2). Norms (percentile, stanine, and T-score) were developed across the entire standardization sample for each grade level.

3. RESULTS

3.1 Item Analysis

Table 1 is the item analysis index for the final 72 item MTST for all six sub-scales. The mean Item Facility Index across all items was M = 0.551 (SD = 0.094; range 0.32–0.75), with 91.7% of items falling within the target 0.30–0.70 range. The mean Discrimination Index was M = 0.406 (SD = 0.068; range 0.25–0.64); all 72 items met the minimum criterion of $D \geq 0.20$, with 48.6% achieving good discrimination ($D \geq 0.40$). The higher-order sub-scales, Inference (Mean D = 0.442) and Generalisation (Mean D = 0.461) had the strongest discriminating power, which was in line with the theoretical predictions that these sub-skills elicit more cognitive differentiation (between high and low performers).

Table 1. Item Analysis Results by Sub-scale

Final 72-Item MTST (Standardization Sample N = 1,200)

Sub-scale	Items	Mean IF	SD (IF)	Mean D	% D ≥ 0.40
Observation	12	0.614	0.087	0.378	41.7%
Comparison	12	0.567	0.094	0.356	33.3%
Classification	12	0.591	0.089	0.388	41.7%
Inference	12	0.502	0.098	0.442	58.3%
Generalisation	12	0.478	0.102	0.461	66.7%
Evaluation	12	0.551	0.091	0.412	50.0%
Total MTST	**72**	**0.551**	**0.094**	**0.406**	**48.6%**

Note. IF = Item Facility Index; D = Discrimination Index (point-biserial r). All items D ≥ 0.20 (minimum acceptable criterion). Target IF range: 0.30–0.70.

3.2 Validity Evidence

3.2.1 Content Validity

The final instrument consisted of 72 items and the overall Content Validity Index (CVI) for the instrument was 0.87, much higher than the recommended 0.80 by Lynn (1986) and Polit and Beck (2006). Sub-scale CVIs were between 0.84 (Generalisation) and 0.91 (Inference) and all were above minimum level. The values provide evidence of the high-level alignment of the MTST items with the operational definitions of each targeted sub-skill (Bloom's Revised Taxonomy levels 2–4) and with the Grade 6–8 developmental level.

3.2.2 Construct Validity

The entire standardization dataset (N = 1,200) was used in EFA. Data were sufficient for analysis as indicated by the pre-condition screening for the Kaiser = 0.89 (meritorious, Kaiser, 1974) and Bartlett's Test of Sphericity $\chi^2(2,556) = 24,312.44$, $p < .001$. Using the Cattell scree test, as well as the independent identification of 6 factors (eigenvalues>1.0), a 6-factor solution was extracted. Table 2 shows the Varimax-rotated solution.

Table 2. Exploratory Factor Analysis: Varimax-Rotated Six-Factor Solution (N = 1,200; KMO = 0.89)

Factor / Sub-skill	Items (n)	Eigenvalue	% Variance	Cumulative %	Sub-scale α
Inference	12	8.42	11.69%	11.69%	0.84
Generalisation	12	7.88	10.94%	22.63%	0.86
Evaluation	12	7.34	10.19%	32.82%	0.82
Classification	12	6.92	9.61%	42.43%	0.80
Comparison	12	6.58	9.14%	51.57%	0.79
Observation	12	6.12	8.50%	60.07%	0.81
Total (all 6 factors)	**72**	—	**60.07%**	**60.07%**	**0.93**

Note. All 72 items loaded ≥ 0.45 on their theoretically intended factor with no cross-loadings > 0.30 (simple structure). Total variance explained = 60.07%. EFA conducted in SPSS v26 using Principal Components Analysis with Varimax rotation.

The six-factor solution accounted for 60.07% of total variance, which is a relatively large percentage of variance for a psychological measurement instrument of this nature (Hair et al., 2014). The loadings of all of the items on the theoretically

predicted factors are above the recommended 0.45, and no item cross loaded with > 0.30, indicating strong empirical support for the six-sub-skill theoretical model of the MTST.

3.2.3 Criterion Validity

Concurrent criterion validity was evaluated by Pearson's r for the total MTST scores and teacher-rated academic performance ratings (5-point scale) within 1 week of the total MTST administration (n = 300). The criterion validity coefficient was r = 0.61 ($p < .001$), which shows that there was a moderate to strong positive relationship, and shared variance of 37% ($r^2 = 0.37$). This is far higher than the criterion criterion of 0.40 that is usually recommended for educational tools recommended for diagnosis purposes (Gregory, 2013), indicating the MTST's ability to discriminate students that perform better academically.

3.3 Reliability

Table 3 shows the reliability estimates of all three methods and all six sub-scales. The overall MTST Cronbach's alpha of 0.93 is very good, well above the 0.80 threshold for high stakes psychometric instruments (Nunnally, 1978). The Spearman-Brown corrected split-half correlation coefficient is 0.91 indicating agreement of the α estimate. The MTST is found to have high temporal stability, with test-retest r = 0.87 ($p < .001$, n = 100, two-week interval), suggesting that it measures stable cognitive processes and not transient states. All individual sub-scales had acceptable to good reliability (α range: 0.79 - 0.86). Table 3. The reliability coefficients of the MTST sub-scales and total score were high.The reliability coefficients for the sub-scales and the total score of the MTST were high.

Table 3. Reliability Coefficients — MTST Sub-scales and Total Score

Sub-scale	α (Internal)	Split-Half (S-B)	Test-Retest r	CVI
Observation	0.81	0.78	0.83	0.89
Comparison	0.79	0.76	0.81	0.86
Classification	0.80	0.79	0.82	0.88
Inference	0.84	0.82	0.85	0.91
Generalisation	0.86	0.83	0.88	0.84
Evaluation	0.82	0.80	0.84	0.87
Total MTST (72 items)	**0.93†**	**0.91†**	**0.87†**	**0.87†**

Note. α = Cronbach's internal consistency coefficient; S-B = Spearman-Brown corrected split-half; Test-Retest r = Pearson's r over two-week interval (n = 100). † Indicates threshold benchmarks: α ≥ 0.80 (Nunnally, 1978); test-retest r ≥ 0.75; CVI ≥ 0.80 (Lynn, 1986) — all exceeded.

3.4 Hypothesis Testing

The results for all five null hypotheses are summarized in Table 4. This MTST had no significant gender differences (H_{01} retained; $t(1,198) = 1.24$, $p = .216$, $d = 0.10$), which indicated gender equity of the instrument. In contrast, highly significant grade-level differences (H_{02} rejected; $F(2,1197) = 156.84$, $p < .001$, $\eta^2 = 0.208$) confirmed strong developmental sensitivity, with mean scores increasing from Grade 6 (M = 38.72, SD = 7.43) through Grade 7 (M = 42.15, SD = 7.19) to Grade 8 (M = 45.89, SD = 6.88). Tukey HSD post-hoc tests revealed significant difference between all three grade pairs (all $p < .001$) with large differences between Grade 6 and Grade 8 ($d = 0.99$). The effect of school type (H_{03}; $\eta^2 = 0.061$) and medium of instruction (H_{04}; $d = 0.48$) were statistically significant, with students in private schools having more than five raw score points higher than their government school peers — a potentially educationally significant equity gap that warrants policy intervention.

Table 4. Consolidated Hypothesis Testing Results

Hypothesis	Statistical Test	Key Result	Effect Size	Decision
H_{01} Gender	Independent t-test	t(1198)=1.24, p=.216	d=0.10 (Negligible)	**RETAINED**
H_{02} Grade	One-Way ANOVA	F(2,1197)=156.84, p<.001	η^2=0.208 (Large)	**REJECTED**
H_{03} School Type	One-Way ANOVA	F(2,1197)=38.72, p<.001	η^2=0.061 (Medium)	**REJECTED**
H_{04} Medium	Independent t-test	t(1198)=6.41, p<.001	d=0.48 (Medium)	**REJECTED**
H_{05} Criterion	Pearson's r	r(298)=0.61, p<.001	r^2=0.37	**REJECTED**

Note. All significant effects reported at $p < .001$ (two-tailed). Effect sizes: Cohen's d for t-tests; partial η^2 for ANOVA. ANOVA post-hoc comparisons by Tukey HSD (familywise error control).

4. DISCUSSION

MTST is psychometrically sound (internationally accepted psychometric standards) in all the dimensions of validity and reliability, making it an evidence-based diagnostic tool for middle school thinking skills. The six-factor EFA solution accounts for 60.07% of the total variance and has a clean simple structure, which renders a strong empirical support of the theoretical model and separates the MTST from other instruments that consider thinking skills as unidimensional composites. The results of this study agree with those of Facione (1990) and Ennis (1985), who also found that inference and evaluation are the sub-skills that are hierarchically superior. This was supported by the present data by the higher discrimination index and larger increments in grade level.

The finding that the strongest sub-skills are Observable and Concrete and the weakest are Generalisation and Inference is consistent with Piagetian developmental theory, which states that Observable and Concrete sub-skills develop earlier, and that Generalisation and Inference require the formal operational stage to be fully developed, which is typically reached in later adolescence (Piaget, 1952). The significant developmental growth (+1.51 raw points from Grade 6 to Grade 8) observed when using Inference indicates that the middle school years may be especially important for the development and enhancement of inferential reasoning skills — a time when students may be particularly ripe for this instruction, with direct implications for curriculum design.

There was no significant gender difference ($d = 0.18$), which is consistent with the recent research confirming that, when cultural and environmental factors are taken into consideration, gender differences in reasoning are generally small (Halpern, 2003). This significant school type effect ($\eta^2 = 0.061$), however, highlights a persistent inequity: private school students' mean raw score is more than 5 points higher than their government school counterparts' mean raw score, with the difference occurring across the two sub-skills that are most reliant on rich learning contexts that foster abstract thinking: Inference and Generalisation.

Restrictions: Only the districts of Anand and Kheda were included in the sample of items, which limits the extent of direct generalisation; CTT is the main psychometric approach used (which is a future direction – IRT analysis); the medium of instruction effect is partially confounded with the effect of type of school. Future studies should include all the Gujarat districts, confirmatory factor analysis should be performed on an independent sample, and predictive validity should be determined with reference to the results of the Board examinations at Grade 10 level.

5. CONCLUSIONS

MTST is a valid, reliable, developmentally calibrated and gender equitable diagnostic tool to measure six discrete micro thinking sub-skills in Grade 6-8 students in the educational context of Gujarat. Satisfactory psychometric characteristics: CVI = 0.87, six factor EFA structure, criterion r = 0.61, Cronbach's $\alpha = 0.93$, and test retest r = 0.87, which are above the

standards for high quality educational assessment. The MTST brings the mandate of competency based assessment of NEP 2020 into reality, offers practitioners and educators a diagnostic sub-skill profile which is not provided by any of the existing instruments in this context and lays down a methodological yardstick for the research on assessment of thinking skills in India. Grade-specific percentile and stanine norms allow for instant clinical and educational interpretation.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the Children's Research University, Gandhinagar, for sponsoring the present research and providing institutional support under its Collaborative Research Programme in Educational Assessment and Psychometrics (2024-26). The authors would like to thank the District Education Offices in Anand and Kheda for permission to conduct research; all 35 participating school principals, teachers and administrators for their cooperation in the data collection; and the 1200 students and their families for their participation. The authors would also like to thank the 12 expert panel members who participated in the content validation. The funding organization did not take part in the collection, analysis, interpretation, or publication decisions of the data.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's educational objectives. Longman.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Prentice Hall.

Bhatia, C. M. (1955). Performance tests of intelligence under Indian conditions. Oxford University Press.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. Educational Leadership, 43(2), 44–48.

Facione, P. A. (1990). Critical thinking: A statement of expert consensus — The Delphi Report. American Philosophical Association.

Gregory, R. J. (2013). Psychological testing: History, principles, and applications (7th ed.). Pearson.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (7th ed.). Pearson.

Halpern, D. F. (2003). Thought and knowledge: An introduction to critical thinking (4th ed.). Lawrence Erlbaum.

IBM Corp. (2022). IBM SPSS Statistics for Windows, Version 26.0. IBM Corp.

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563–575.

Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382–385.

Ministry of Education, Government of India. (2020). National Education Policy 2020. Ministry of Education.

NCERT. (2005). National Curriculum Framework 2005. NCERT.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). McGraw-Hill.

Patel, M., & Chauhan, R. (2003). Thinking sub-skills in Gujarati-medium secondary school students. Vadodara Journal of Educational Studies, 8(1), 23–39.

Piaget, J. (1952). The origins of intelligence in children. International Universities Press.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Critique and recommendations. Research in Nursing & Health, 29(5), 489–497.

Resnick, L. B. (1987). Education and learning to think. National Academy Press.

Sharma, G. (1990). Critical thinking skills among upper primary students [Master's thesis, Delhi University].

Singh, A. K. (2015). Tests, measurements and research methods in behavioural sciences (5th ed.). Bharati Bhawan.

Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (6th ed.). Pearson.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Watson, G., & Glaser, E. M. (1980). Watson-Glaser Critical Thinking Appraisal. Psychological Corporation.