

Development, Pilot Testing and Preliminary Psychometric Evaluation of an Achievement Test for Assessing Pre-Service Teachers' Knowledge of Inclusive Education in Nagaland

*Aseto Iralu Meyase, Prof. Gyanendra Nath Tiwari

Department of Teacher Education, Nagaland University: Kohima Campus, Nagaland, India

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.100801133>

Received: 09 September 2026; Accepted: 14 September 2026; Published: 25 September 2026

ABSTRACT

Most existing empirical instruments in inclusive education evaluate teacher attitudes and self-efficacy rather than actual pedagogical content knowledge. This gap is particularly pronounced in Nagaland due to a dearth of psychometrically validated achievement tests which evaluates pre-service teachers' content mastery within the state's prescribed curriculum. To address this deficit, this study developed and pilot-tested an objective achievement test using an instrumentation design grounded in Classical Test Theory and Bloom's Revised Taxonomy. An initial pool of 60 items, based on Nagaland University's B.Ed. Course-10 curriculum underwent expert review for content validity prior to pilot administration. The pilot sample comprised of 42 M.Ed. students from Nagaland University's Kohima Campus, the only institution offering the M.Ed. programme in the state. Item analysis evaluated difficulty indices, item discrimination and internal consistency. This resulted in removal of 10 poorly performing items, raising Cronbach's alpha from .641 to .727 for the refined 50-item version. The refined 50-item instrument maintains balanced representation across five content domains and five cognitive levels, with items operating at the application and analysis levels showing the strongest discrimination indices. The resulting instrument provides teacher educators with a preliminary diagnostic measure to identify specific conceptual gaps in inclusive education preparation, with implications for B.Ed. program design under NEP 2020. These findings offer preliminary evidence of item functioning and internal consistency rather than full validation and require further validation with a larger and more representative sample. It also establishes a replicable methodological framework for similar tool development in other Northeast Indian states.

Keywords: Inclusive education, achievement test development, pilot testing, psychometric evaluation, pre-service teachers, item analysis

INTRODUCTION

The global movement towards inclusive classrooms urges teacher training programs to emphasize specific abilities to handle diversity instead of generic pedagogy (Forlin et al., 2011). In India, the National Education Policy 2020 has reoriented the teacher accountability frameworks to address the needs of historically underprivileged children with special needs (Sarkar & Yadav, 2023). But there is still a disconnect between these national policies and practical classrooms. Pre-service teachers commonly have positive attitudes towards inclusion, but often lack the precise cognitive know-how to properly employ modified teaching practices (Kielblock et al., 2026; Mistry, 2022).

The existing literature also indicates that there is a dearth of specifically designed diagnostic tools to measure pre-service teachers' demonstrated knowledge of inclusive education. Current instruments in inclusive education research focus primarily on teacher attitudes and self-perceived competence instead of verified knowledge (Forlin et al., 2011; Sharma & Sokal, 2016). For instance, attitudinal scales like the SACIE-R (Forlin et al., 2011) measure predispositions but cannot isolate a teacher's actual pedagogical knowledge from their professional motivation. This structural limitation leaves a distinct measurement gap. Currently, no

psychometrically validated achievement test evaluates pre-service teachers' knowledge within the curricular framework of Nagaland.

To address this issue, a new assessment instrument offers a context-specific solution. The initial formulation of 60 items drew on Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001), while Classical Test Theory (Crocker & Algina, 1986) guided the psychometric design. Following a pilot administration with 42 M.Ed. students at Nagaland University, item analysis dictated the reduction of the test to 50 items. Evaluation of this final version focused on internal consistency reliability, item discrimination and item difficulty. The resulting tool provides an empirical measure for evaluating teacher preparation programs in Nagaland.

LITERATURE REVIEW

For a long time, psychometric evaluations of inclusive education have targeted teachers' feeling and motivations. This focus on affective and conative elements has left the actual cognitive knowledge of educators largely unmeasured. On a global scale, tools like the teacher efficacy for Inclusive Practices (TEIP) scale (Sharma et al., 2012) and the Attitudes Towards Inclusion for All (ATIFA) scale (Kielblock et al., 2026) measure teacher philosophies and anxieties internationally. The Sentiments, Attitudes and Concerns about Inclusive Education Revised (SACIE-R) scale (Forlin et al., 2011), the Opinions Relative to Integration of Students with Disabilities (ORI) scale (Antonak & Larrivée, 1995) and the Multidimensional Attitudes Scale towards Persons with Disabilities (Findler et al., 2007) scale have had meaningful contributions to understanding how teachers position themselves towards inclusion. The more recent Teacher Inclusive Education Literacy Scale (Gültekin & Çetin, 2025)) has expanded upon the construct to some extent. However, none of these instruments prompt teachers to demonstrate knowledge but rather to report perceptions of their own knowledge, a methodically critical distinction (Sharma & Sokal, 2016; DeVellis, 2016).

In India, assessment research has largely concentrated on institutional barriers, lack of resources, or general teacher worries. Although, the National Education Policy (NEP) 2020 calls for structural changes in teacher training to build equitable classrooms (Sarkar & Yadav, 2023), this policy commitment has not been sufficiently matched by rigorous and context-specific assessment tools yet. Still, domestic evaluation tools remain largely descriptive and are not designed to psychometrically distinguish whether pre-service teachers have the required cognitive knowledge to effectively plan and execute differentiated instruction. (Pradhan & Gochhayat, 2023).

METHODOLOGY

Research Design

This study used an instrumentation research design to develop and pilot-test the achievement test. The overall approach follows Classical Test Theory (CTT), where test quality depends on how individual items perform and how reliable the scale is as a whole (Crocker & Algina, 1986; Nunnally, 1978). The work progressed in three distinct stages: instrument development, expert content validation and a pilot test with item analysis.

Theoretical and Curricular Grounding

The achievement test was designed to assess the knowledge of pre-service teachers regarding inclusive education in Nagaland. Two frameworks guided the development of these items. First, the test drew directly from Course-10: Creating an Inclusive School, which is a compulsory module in the Nagaland University B.Ed. curriculum. This course details the core principles of inclusive education, policy provisions, classroom strategies and support systems. Second, the cognitive design relied on Bloom's Revised Taxonomy. This structure distributed the questions across different levels of thinking rather than limiting them to simple recall.

Test Blueprint

To ensure proper content representation and cognitive balance, a two-dimensional test blueprint was constructed before drafting any items. The resulting matrix distributed the 60 test items across five content areas and five cognitive domains.

Table 1: Content Area Distribution

Unit	Topics Covered	Weightage	No. of Questions	Marks
Unit I	Special Needs and Education	20%	12	12
Unit II	Policies and Provisions	20%	12	12
Unit III	Inclusive Practices	30%	18	18
Unit IV	Developing Support Networks	20%	12	12
Unit V	Miscellaneous	10%	6	6
Total		100%	60	60

Table 2: Cognitive Level Distribution

Cognitive Level	Description	Weightage	No. of Questions	Distribution by Unit
Remembering	Recall facts, concepts, principles	20%	12	Unit I: 3, Unit II: 3, Unit III: 4, Unit IV: 2, Unit V: 0
Understanding	Explain ideas, interpret concepts	30%	18	Unit I: 3, Unit II: 3, Unit III: 5, Unit IV: 4, Unit V: 3
Applying	Use knowledge in new situations	25%	15	Unit I: 3, Unit II: 3, Unit III: 5, Unit IV: 3, Unit V: 1
Analyzing	Break down information, compare	15%	9	Unit I: 2, Unit II: 2, Unit III: 3, Unit IV: 1, Unit V: 1
Evaluating	Make judgements, critique decisions	10%	6	Unit I: 1, Unit II: 1, Unit III: 1, Unit IV: 2, Unit V: 1
Total		100%	60	

Table 3: Question Type Distribution

Question Type	Number	Percentage	Distribution
Multiple Choice	48	80%	All Units
True/False	4	6.67%	Units I, II, III, IV
Matching Type	4	6.67%	Units I, II, III, IV
Assertion-Reason	4	6.67%	Units III, IV, V
Total	60	100%	

Most of the test consisted of multiple-choice items (80%), which allow for objective scoring across various cognitive levels. To measure higher-level thinking, specifically analysis and evaluation, matching-type and assertion-reason items were also included.

Item Writing

Drafting began with an initial pool of 60 items based on the blueprint. These items were written to be clear, culturally neutral and directly related to the Nagaland University B.Ed. curriculum. Before the expert review, every item was mapped to its specific curricular unit and cognitive level.

Content Validity

To establish content validity, the entire 60 item pool was given in print to five experts in inclusive pedagogy and psychometric test development. Rather than completing a structured numerical rating form, each expert reviewed the instrument directly, marking concerns on the page and communicating feedback in person and through direct discussion with the researchers. The qualitative review examined each item for accuracy, clarity,

relevance to the B.Ed. Course-10 curriculum and appropriateness of the assigned cognitive level.

Feedback gathered through this process was consolidated across the five experts and used to revise item wording, sharpen conceptual precision and correct structural or organizational issues in specific items. Where experts raised the same concern about an item, that item was prioritized for revision; where concerns were isolated to a single reviewer, the researchers judged each case individually against the curriculum and the item's intended cognitive level. No items were removed at this stage. After revision, the panel considered all 60 items suitable for pilot administration.

Participants

Forty-two M.Ed. students from the Kohima Campus of Nagaland University participated in the pilot study. Nagaland University is the only institution in the state offering the M.Ed. programme and its annual intake capacity is 50 seats. The 42 participants therefore represent a substantial proportion of the current M.Ed. cohort, though this figure describes a single-institution, a single-cohort sample rather than a state-wide validation population.

Selection relied on purposive sampling based on two criteria: prior completion of a B.Ed. programme featuring the inclusive education module (Course-10), and current enrolment in the M.Ed. programme. These criteria established that participants had direct exposure to the assessed subject matter. Students with a D.El.Ed. background were excluded because that curriculum does not cover the required content.

Although a sample of 42 is modest, psychometric literature confirms that it is sufficient for item-level analysis during the initial phases of test development, especially when the target population is small and specific (Crocker & Algina, 1986; Nunally, 1978).

Statistical Analysis

Evaluation of the psychometric properties at both the item and scale levels involved several specific steps:

1. Descriptive Statistics: Mean scores and standard deviations were calculated to evaluate the score distribution and check for sufficient variance among the participants.
2. Item Difficulty Index (p-value): The difficulty of each item was measured by calculating the proportion of correct answers. Items with p-values between .30 and .70 fell within the acceptable difficulty range. Questions outside these boundaries were marked for potential adjustment or exclusion.
3. Item Discrimination Index: To evaluate item discrimination, the response patterns of the top 27% and bottom 27% of scores were compared. Items with discrimination values below .20 were classified as weak discriminators and marked for revision.
4. Item-Total Statistics: Corrected item-total correlations were calculated to determine how closely individual items related to the overall score. This served as a secondary check on item discrimination.
5. Internal Consistency Reliability: Cronbach's α was calculated to measure the overall consistency of the scale. Reliability was deemed acceptable at a threshold of .70 or higher, which matches standard educational measurement practices.

These combined calculations established a clear psychometric profile for the draft test. The resulting values guided the choices to keep, modify or drop specific items before administering the final test.

RESULTS AND DISCUSSION

Results

Descriptive Statistics

Analysis of the score distributions for both the initial and refined versions of the test provided the baseline data. On the initial 60 item test, the mean score was 45.6 (SD = 4.63), with scores ranging from 34 to 52 (N = 42). After the deletion and modification of poorly performing items, the final 50 item test produced a mean

score of 37.26 (SD = 4.91), with individual scores ranging from 24 to 45. Moderate variability in both sets of scores provided the necessary variance for reliable item-level analyses.

Table 4: Descriptive Statistics – Initial 60 item

	N	Minimum	Maximum	Mean	Std. Deviation
Initial Total	42	34	52	45.619	4.63261
Valid N (listwise)	42				

Table 5: Descriptive Statistic – Refined 50 item

	N	Minimum	Maximum	Mean	Std. Deviation
Refined Total	42	24	45	37.2619	4.90922
Valid N (listwise)	42				

Reliability Analysis

The initial 60 item instrument had a Cronbach's alpha coefficient of $\alpha = .641$. This value fell below the standard threshold of .70 for acceptable internal consistency. Item-total statistics pointed to two main categories of problematic items. First, items Q4, Q44, Q55 and Q58 produced negative corrected item-total correlations ($r = -.138, -.258, -.101$ and $-.065$ respectively). Second, items Q26, Q28 and Q36 showed zero response variance, where $p = 1.00$. These seven items were discarded during the initial refinement step. An evaluation of the remaining items flagged Q9, Q15 and Q23 as underperforming elements with negative corrected item-total correlations of $-.192, -.226$ and $-.075$ respectively when computed against the 60-item total. These three items were also excluded. After removing these ten items, the final 50 item instrument achieved a Cronbach's alpha of $\alpha = .727$. This result satisfies the standard threshold for internal consistency reliability. As the test spans five distinct content domains, this alpha is best treated as preliminary evidence of internal consistency rather than conclusive proof of scale-wide reliability. Domain-wise reliability has not been computed and is recommended for the planned larger study.

Table 6: Reliability Statistics – Initial 60 Item

Reliability Statistics	
Cronbach's Alpha	N of Items
.641	60

Table 7: Reliability Statistics – Refined 50 Item

Reliability Statistics	
Cronbach's Alpha	N of Items
.727	50

Item Discrimination Analysis

Corrected item-total correlations served as the primary measure of item discrimination, calculated separately against the initial 60-item total and the refined 50-item total (Tables 8 and 9). For the initial 60-item instrument, correlations ranged from $r = -.258$ (Q44) to $r = .505$ (Q34). After removing the poorly performing items, the discrimination values across the remaining 50 items improved. The strongest discriminators in this final pool were Q34 ($r = .548$), Q41 ($r = .473$), Q47 ($r = .450$), Q32 ($r = .436$), Q38 ($r = .402$), Q40 ($r = .401$),

Q43 ($r = .401$) and Q52 ($r = .402$).

Two retained items showed negative discrimination against the refined 50-item total: Q18 ($r = -.006$) and Q39 ($r = -.011$). Both were kept to preserve content coverage in their respective units. Three items, Q21, Q30 and Q39, differ in sign between the two totals (Table 8 vs. Table 9), reflecting the recalculation of item-total correlation after item removal rather than a data error.

Table 8: Item-Total Statistics – Initial 60-Items

Item	Difficulty Index (p)	Corrected Item-Total Correlation (vs. 60-item total)	Decision	Reason
Q1	0.93	0.103	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q2	0.88	0.358	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q3	0.31	0.231	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q4	0.95	-0.138	Delete	Negative corrected item-total correlation ($r = -.138$)
Q5	0.26	0.133	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q6	0.71	0.333	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q7	0.29	0.093	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q8	0.86	0.361	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q9	0.67	-0.192	Delete	Negative corrected item-total correlation ($r = -.192$)
Q10	0.95	0.306	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q11	0.69	0.173	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q12	0.6	0.178	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q13	0.95	0.256	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q14	0.55	0.208	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q15	0.52	-0.226	Delete	Negative corrected item-total correlation ($r = -.226$)
Q16	0.31	0.022	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q17	0.69	0.244	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q18	0.57	-0.012	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q19	0.62	0.347	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q20	0.95	0.256	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q21	0.67	-0.007	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q22	0.4	0.134	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q23	0.45	-0.075	Delete	Negative corrected item-total correlation ($r = -.075$)
Q24	0.76	0.044	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q25	0.93	0.144	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q26	1.0	0.000	Delete	Zero response variance ($p = 1.00$); item-total correlation = .000
Q27	0.71	0.010	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q28	1.0	0.000	Delete	Zero response variance ($p = 1.00$); item-total correlation = .000
Q29	0.57	0.105	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q30	0.74	-0.027	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q31	0.24	0.039	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q32	0.88	0.409	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q33	0.29	0.177	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q34	0.86	0.505	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q35	0.83	0.120	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q36	1.0	0.000	Delete	Zero response variance ($p = 1.00$); item-total correlation = .000
Q37	0.71	0.022	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q38	0.9	0.361	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q39	0.26	0.025	Retain	Proceeded to refined 50-item analysis (see Table 9)

Q40	0.98	0.368	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q41	0.9	0.417	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q42	0.88	0.257	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q43	0.98	0.368	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q44	0.93	-0.258	Delete	Negative corrected item-total correlation ($r = -.258$); weakest discriminator in the pool
Q45	0.95	0.331	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q46	0.93	0.165	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q47	0.93	0.416	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q48	0.98	0.263	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q49	0.69	0.304	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q50	0.98	0.263	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q51	0.64	0.073	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q52	0.9	0.380	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q53	0.95	0.106	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q54	0.98	0.229	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q55	0.88	-0.101	Delete	Negative corrected item-total correlation ($r = -.101$)
Q56	0.88	0.191	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q57	0.9	0.417	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q58	0.95	-0.065	Delete	Negative corrected item-total correlation ($r = -.065$)
Q59	0.95	0.331	Retain	Proceeded to refined 50-item analysis (see Table 9)
Q60	0.98	0.263	Retain	Proceeded to refined 50-item analysis (see Table 9)

Table 9: Item-Total Statistics – Refined 50 Items (retained items only, recalculated against the 50-item total)

Item	Corrected Item-Total Correlation (vs. refined 50-item total)	Note
Q1	0.057	Weak positive discrimination; retained for content coverage
Q2	0.309	Acceptable positive discrimination
Q3	0.225	Acceptable positive discrimination
Q5	0.125	Below-average positive discrimination; acceptable for retention
Q6	0.298	Acceptable positive discrimination
Q7	0.069	Weak positive discrimination; retained for content coverage
Q8	0.367	Acceptable positive discrimination
Q10	0.342	Acceptable positive discrimination
Q11	0.245	Acceptable positive discrimination
Q12	0.219	Acceptable positive discrimination
Q13	0.295	Acceptable positive discrimination
Q14	0.240	Acceptable positive discrimination
Q16	0.050	Weak positive discrimination; retained for content coverage
Q17	0.347	Acceptable positive discrimination
Q18	-0.006	Negative discrimination after refinement; retained for content coverage
Q19	0.361	Acceptable positive discrimination

Q20	0.248	Acceptable positive discrimination
Q21	0.014	Weak positive discrimination; retained for content coverage
Q22	0.147	Below-average positive discrimination; acceptable for retention
Q24	0.082	Weak positive discrimination; retained for content coverage
Q25	0.154	Below-average positive discrimination; acceptable for retention
Q27	0.072	Weak positive discrimination; retained for content coverage
Q29	0.115	Below-average positive discrimination; acceptable for retention
Q30	0.009	Weak positive discrimination; retained for content coverage
Q31	0.044	Weak positive discrimination; retained for content coverage
Q32	0.436	Acceptable positive discrimination
Q33	0.147	Below-average positive discrimination; acceptable for retention
Q34	0.548	Acceptable positive discrimination
Q35	0.107	Below-average positive discrimination; acceptable for retention
Q37	0.050	Weak positive discrimination; retained for content coverage
Q38	0.402	Acceptable positive discrimination
Q39	-0.011	Negative discrimination after refinement; retained for content coverage
Q40	0.401	Acceptable positive discrimination
Q41	0.473	Acceptable positive discrimination
Q42	0.230	Acceptable positive discrimination
Q43	0.401	Acceptable positive discrimination
Q45	0.295	Acceptable positive discrimination
Q46	0.154	Below-average positive discrimination; acceptable for retention
Q47	0.450	Acceptable positive discrimination
Q48	0.236	Acceptable positive discrimination
Q49	0.301	Acceptable positive discrimination
Q50	0.236	Acceptable positive discrimination
Q51	0.065	Weak positive discrimination; retained for content coverage
Q52	0.402	Acceptable positive discrimination
Q53	0.154	Below-average positive discrimination; acceptable for retention
Q54	0.236	Acceptable positive discrimination
Q56	0.183	Below-average positive discrimination; acceptable for retention
Q57	0.367	Acceptable positive discrimination
Q59	0.295	Acceptable positive discrimination
Q60	0.236	Acceptable positive discrimination

Item Difficulty Analysis

Difficulty indices (p -values) were calculated for each item. The initial version of the test showed a right-skewed distribution, marked by a high concentration of items with p -values above .80, including three items where p -values = 1.00 (Q26, Q28 and Q36). Harder items included Q31 (p = .24), Q5 (p = .26), Q39 (p = .26) and Q7 (p = .29). This high concentration of high scores was expected because the participants had already covered the inclusive education content during their B.Ed. coursework.

The final version of the test maintained a similar difficulty distribution, which fits the purpose of a curriculum mastery test. Items with moderate difficulty fell within the .30 to .70 range, such as Q22 (p = .40), Q14 (p = .55), Q12 (p = .60) and Q19 (p = .62). these items distinguished performance across different levels of candidate knowledge.

Item Retention Summary

The refinement process resulted in the removal of 10 items across two stages. This selection relied on combined evidence from the reliability, discrimination and difficulty analysis. Ultimately, the remaining 50 items demonstrated stronger psychometric performance than the initial pool. Tables 8 and 9 together report the full item-level evidence for all 60 items. Table 8 gives each item's difficulty index and its correlation against the 60-item total along with the retain/delete decision. Table 9 gives the retained items' correlations against the refined 50-item total.

DISCUSSION

Data from the pilot study show that systematic item refinement improved the overall performance metrics of the test. Raising Cronbach's alpha from α = .641 to α = .727 resulted directly from deleting items that reduced internal consistency. These deletions fell under two categories: items with negative correlation coefficients and items with zero response variance.

The final 50 item test has several advantages:

1. The reliability estimate of α = .727 meets the general threshold required for educational research tools, though it should be read as preliminary given the pilot-scale sample.
2. The refinement process preserved curriculum coverage across all five domains of the original blueprint. Items were kept if their removal would have left a content area unrepresented, even if their individual statistical indicators were marginal.
3. In terms of cognitive distribution, the questions cover multiple levels of Bloom's Revised Taxonomy.
4. The most effective items in terms of discrimination coefficients corresponded to the application and analysis levels, which matches observations in the wider literature on achievement test development.

The mastery-oriented nature of the assessment is reflected by the high success rates on many of the questions and the fact that participants had already completed the relevant coursework. Rather than implicating poor item construction, these high scores stem from the specific academic background of the cohort. Grounding the assessment directly in the Nagaland University B.Ed. syllabus and drawing from a pilot/refinement sample at the state's only M.Ed.-granting institution provides preliminary evidence of item functioning for the group rather than validation for the instrument as a whole. Applying this tool to other student populations would require separate validation studies with a larger and more representative sample.

CONCLUSION AND IMPLICATIONS

An achievement test was developed, refined and pilot-tested to assess what pre-service teachers in Nagaland know about inclusive education. The pilot study involved 42 M.Ed. students at Nagaland University. Following two rounds of item refinement, a 50-item test was developed, with an internal consistency reliability coefficient of α = .727. This candidate tool has acceptable internal consistency, distributed content representing five curriculum units and five cognitive levels based on Bloom's Revised Taxonomy. The difficulty distribution matches the requirements of a curriculum mastery test.

A contextually grounded, psychometrically evaluated knowledge test of this type offers direct utility for teacher education within Nagaland, once validated with a larger and more representative B.Ed. sample. Under the National Education Policy 2020, inclusive is a core competency for B.Ed. programs. Teacher education institutions require diagnostic tools that identify specific conceptual gaps rather than merely measuring attitudinal profiles. The developed test moves towards filling that requirement. Teacher educators can analyse score distributions to determine which areas of curriculum delivery require attention. This analysis provides an empirical rather than speculative foundation for B.Ed. program redesign. From a policy perspective, this tool provides a replicable methodology for building contextually valid evaluation tools in other underrepresented states across Northeast India.

LIMITATIONS AND FUTURE DIRECTIONS

1. The pilot sample of 42 participants was drawn from the M.Ed. cohort at Nagaland University's Kohima Campus, the state's only M.Ed.-granting institution (annual intake capacity: 50 seats). This sample served as a pilot/refinement sample for item analysis, not a validation sample, and the absolute sample size remains small. The reported psychometric properties must therefore be treated as preliminary. These initial results are sufficient to support tool refinement and progression towards the main study, but they are not yet generalizable beyond this cohort.
2. Two retained items: Q18 ($r = -.006$) and Q39 ($r = -.011$), show negative corrected item-total correlations when recalculated against the refined 50-item total (Table 9), and were kept to preserve minimum content coverage in their respective units. These items should be prioritized for revision or removal in the next round of item analysis with a larger sample.
3. The main study will involve a larger sample of B.Ed. pre-service teachers across teacher education colleges in Nagaland. This larger group will support confirmatory reliability analysis, exploratory factory structure examination and the development of score-based performance benchmarks.
4. Future investigations should also examine the relationship between knowledge scores and actual classroom practice. While the current tool helps facilitate this inquiry, it cannot establish a direct link to classroom performance on its own.

REFERENCES

1. Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's educational objectives*. Longman.
2. Antonak, R. F., & Larrivée, B. (1995). Psychometric analysis and revision of the Opinions Relative to Mainstreaming scale. *Exceptional Children*, 62(2), 139–149. <https://doi.org/10.1177/001440299506200204>
3. Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
4. DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). SAGE.
5. Findler, L., Vilchinsky, N., & Werner, S. (2007). The Multidimensional Attitudes Scale toward Persons with Disabilities (MAS): Construction and validation. *Rehabilitation Counseling Bulletin*, 50(3), 166–176. <https://doi.org/10.1177/00343552070500030401>
6. Forlin, C., Earle, C., Loreman, T., & Sharma, U. (2011). The Sentiments, Attitudes, and Concerns about Inclusive Education Revised (SACIE-R) scale for measuring pre-service teachers' perceptions about inclusion. *Exceptionality Education International*, 21(3),
7. Gültekin Talayhan, Ö., & Çetin, Ş. (2025). Measuring Teachers' Inclusive Education Literacy: A Scale Development Study. *The Universal Academic Research Journal*, 7(2), 58-75. <https://doi.org/10.55236/tuara.1639455>
8. Kielblock, S., Woodcock, S. & Ehrich, J. Examining teachers' attitudes towards inclusive education for all: development of a new scale. *BMC Psychol* 14, 503 (2026). <https://doi.org/10.1186/s40359-026-04478-6>
9. Mistry, H. S. (2018). Inclusive Education Teaching Aptitude Test [IETAT]: Construction and item analysis. *Online International Interdisciplinary Research Journal*, 8(3), (May–June 2018). ISSN 2249-9598. <http://www.oijrj.org>
10. Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

11. Pradhan, R. S., & Gochhayat, L. (2023). A Systematic Review on Assessment Practices in Inclusive Schools in India: Barriers and Opportunities. *International Journal of Current Educational Studies*, 2(2), 16–27. <https://doi.org/10.46328/ijces.79>
12. Sarkar, L., & Yadav, S. (2023). Teacher Preparation for Inclusive Education-Intervention of NEP 2020. *International Research Journal of Educational Psychology*, 7(2), 06-10. <https://www.irjep.in/index.php/IRJEP/article/view/83>
13. Sharma, U., & Sokal, L. (2016). Can teachers' self-reported efficacy, concerns, and attitudes toward inclusion scores predict their actual inclusive classroom practices? *Australasian Journal of Special Education*, 40(1), 21–38. <https://doi.org/10.1017/jse.2015.14>
14. Sharma, U., Forlin, C., & Loreman, T. (2008). Impact of training on pre-service teachers' attitudes and concerns about inclusive education and sentiments about disability. *Disability & Society*, 23(7), 773–785. <https://doi.org/10.1080/09687590802469271>
15. Sharma, U., Loreman, T. and Forlin, C. (2012), Measuring teacher efficacy to implement inclusive practices. *Journal of Research in Special Educational Needs*, 12: 12-21. <https://doi.org/10.1111/j.1471-3802.2011.01200.x>