Journal of Mathematics Science and Computer Education
https://ppjp.ulm.ac.id/journals/index.php/jmsc-edu/index
e-ISSN: 2807-9167 p-ISSN: 2807-9329
Vol 6 No 1 2026
DOI: https://doi.org/10.20527/jmscedu.v6i1.18792

Developing HOTS-Based Assessment Instrument for Science Learning in Junior High School: Validity, Reliability, and Item Analysis

Ismadi Sihombing[1,2]*, Ilham Adi Putra[3,2], Masringgit Marwiyah Nst[4,2], Festiyed[5], Skunda Diliarosta[5], Fatni Mufid[5]

[1]Universitas Mikroskil, Medan, Indonesia
[2]Doctoral Program of Science Education, Universitas Negeri Padang, Padang, Indonesia
[3]Department of Physics Education, Universitas Adzkia, Padang, Indonesia
[4]Politeknik Adiguna Maritim Indonesia, Medan, Indonesia
[5]Department of Science Education, Universitas Negeri Padang, Padang, Indonesia
*ismadisihombing@gmail.com

Article Info

Article history:
Received: April 21, 2026
Accepted: July 12, 2026
Published: July 26, 2026

Keywords:
HOTS assessment
Instrument development
Item analysis
Junior high school
Two-tier multiple choice

This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

© 2026 Journal of Mathematics, Science, and Computer Education

Abstract

Despite growing policy demands for Higher Order Thinking Skills (HOTS) assessment in Indonesian science education, psychometrically validated instruments aligned with the Merdeka Curriculum remain scarce. This study aims to develop and validate a HOTS-based assessment instrument for eighth-grade junior high school (SMP) science students. Employing the Research and Development (R&D) approach with the 4-D model (Define, Design, Develop, Disseminate), the instrument was designed to comprehensively measure HOTS at cognitive levels C4 (Analyzing), C5 (Evaluating), and C6 (Creating) of the Revised Bloom's Taxonomy, covering Basic Competency (BC) 3.8 (Respiratory System) and BC 3.9 (Photosynthesis) topics. The instrument consists of 40 two-tier multiple-choice (TTMC) items and 10 essay items. Content validity was assessed by three experts using Aiken's V formula, yielding a mean V of 0.87 (highly valid). Reliability was estimated using the KR-20 formula in a large-group trial (n = 64), producing a coefficient of 0.82 (reliable), and Cronbach's alpha of 0.79 for essay items. Item analysis encompassed difficulty index, discriminating power, and distractor effectiveness. Results showed that 34 of 40 TTMC items (85%) met all quality criteria simultaneously: moderate difficulty, good discriminating power, and effective distractors. The final instrument (34 TTMC + 10 essay items) is valid and reliable for measuring higher-order thinking skills among SMP students. The TTMC format provides added diagnostic value by revealing specific conceptual misconceptions in addition to measuring HOTS proficiency, aligning with the formative assessment emphasis of the Merdeka Curriculum.

To cite this article: Sihombing, I., Putra, I. A., Nst, M. M., Festiyed, F., Diliarosta, S., & Mufid, F. (2026). Developing hots-based assessment instrument for science learning in junior high school: Validity, reliability, and item analysis. *Journal of Mathematics, Science, and Computer Education*, *6*(1), 436-455.

INTRODUCTION

The capacity of education systems to cultivate Higher Order Thinking Skills (HOTS), encompassing the analysis of complex information, evaluation of competing evidence-based claims, and creation of novel solutions, has become a defining criterion of educational

quality in the 21st century. International frameworks including the OECD's Framework for 21st Century Competencies, UNESCO's Education for Sustainable Development goals, and the Partnership for 21st Century Skills converge on the recognition that rote memorization and procedural recall are insufficient preparation for the cognitive demands of contemporary professional and civic life (Anderson & Krathwohl, 2021).

This global consensus makes the systematic assessment of HOTS not merely desirable but educationally urgent: without valid, reliable instruments to measure HOTS, educators cannot identify gaps, monitor progress, or design evidence-based interventions. The development of such instruments is therefore a foundational priority for education systems worldwide (Dinni, 2018).

In the Indonesian context, this global imperative is reflected in Permendikbudristek Number 12 of 2024 on the Merdeka Curriculum, which explicitly positions HOTS development as a central objective of SMP-level science learning. The Merdeka Curriculum emphasizes the cultivation of the Pancasila Student Profile, particularly the dimensions of critical reasoning and creativity, both of which correspond directly to HOTS levels C4–C6 of the Revised Bloom's Taxonomy (Anderson & Krathwohl, 2021). Science education is regarded as a particularly important vehicle for HOTS development, given that the scientific method, like observation, hypothesis formulation, experimental design, data interpretation, and conclusion drawing, is inherently a higher-order cognitive enterprise.

This emphasis on C4, C5, and C6 rests on the Revised Bloom's Taxonomy, which reorganized the original six-level taxonomy into a two-dimensional framework: the knowledge dimension (factual, conceptual, procedural, metacognitive) and the cognitive process dimension (remember, understand, apply, analyze, evaluate, create). HOTS corresponds to the top three cognitive process levels applied to the higher categories of the knowledge dimension (Anderson & Krathwohl, 2021). Analysis (C4) involves breaking material into its constituent parts and determining how they relate to each other and to the overall structure or purpose. Evaluate (C5) involves making judgments about value based on criteria and standards. Create (C6) involves assembling elements into a novel, coherent whole, the highest cognitive demand in the taxonomy (Anderson & Krathwohl, 2021).

For SMP science education specifically, HOTS at C4 can manifest as identifying causal relationships between physiological mechanisms and health outcomes (e.g., analyzing how chronic hypoxia at high altitude triggers compensatory erythropoiesis in the Respiratory System topic). HOTS at C5 is demonstrated when students evaluate competing experimental designs for testing photosynthesis-limiting factors, applying established criteria of scientific validity such as controlled variable isolation and reproducibility. HOTS at C6 is evidenced when students design a multi-variable experiment to investigate the combined effects of temperature, light intensity, and CO_2 concentration on photosynthesis rate, a novel synthesis task. The Merdeka Curriculum's learning objectives for BC 3.8 and BC 3.9 explicitly include cognitive demands at all three HOTS levels, confirming the curricular validity of targeting C4, C5 and C6 in the instrument design.

The relevance of HOTS assessment in science education is further supported by cognitive learning theory. Bransford et al., in Dinni (2018), demonstrated that deep, transferable understanding, the kind of knowledge needed for real-world problem-solving, results from learning processes that engage higher-order operations such as explanation, comparison, and generalization. Superficial processing, associated with Lower Order Thinking Skills (LOTS), produces knowledge that is brittle, context-bound, and easily forgotten. Research in science education specifically has shown that students who are regularly assessed using HOTS-aligned instruments demonstrate superior conceptual understanding, greater ability to apply knowledge to novel contexts, and lower rates of misconception persistence compared to students assessed only through LOTS items.

Despite this policy orientation, international comparative assessments consistently reveal a significant gap between the stated aspirations of Indonesian science education and the actual HOTS achievement of Indonesian students. The Programme for International Student Assessment (OECD, 2023) placed Indonesia at rank 70 of 81 participating countries in the science domain, with a mean score of 383 points, compared to the OECD average of 485 points (OECD, 2023). More critically, PISA's cognitive framework analysis revealed that fewer than 3% of Indonesian students reached PISA Level 5 or above in science, the levels corresponding to higher-order reasoning tasks. Indonesia's Trends in International Mathematics and Science Study (TIMSS) performance has similarly reflected difficulties at the higher cognitive demands, particularly in reasoning items (Kurniawan & Arnellis, 2019).

A central structural explanation for this gap lies in the assessment practices prevalent in Indonesian schools. Multiple studies have documented that the majority of assessment items used by teachers in classroom contexts operate exclusively at Bloom's levels C1 (Remembering) and C2 (Understanding) (Anderson & Krathwohl, 2021). When students are consistently assessed only on recall and basic comprehension, the learning activities and teacher feedback naturally orient toward those same cognitive levels, creating a self-reinforcing cycle that prevents the development of higher-order capacities. This 'assessment drives learning' dynamic is well established in the educational psychology literature and has been documented specifically in the Indonesian science classroom context (Mulyatna et al., 2021).

The two-tier multiple choice (TTMC) format represents a methodologically sophisticated approach to HOTS assessment that addresses several limitations of conventional formats. Developed by Treagust (1988) for diagnosing misconceptions in physics and subsequently validated across multiple science disciplines, TTMC assesses both the correctness of students' conceptual answers (Tier 1) and the quality of their reasoning (Tier 2). This dual-layer structure simultaneously provides a quantitative HOTS score and qualitative diagnostic information about misconceptions—information that can directly inform instructional decisions. In the Indonesian context, TTMC has been found effective for diagnosing specific misconceptions in biology topics such as cell division, genetics, and biological classification (Asrafil et al., 2020), but its application to the combined topics of Respiratory System and photosynthesis for SMP class VIII, with full psychometric validation, has not been previously reported (Astuti et al., 2021).

Although TTMC was originally conceived purely as a misconception-diagnosis tool (Treagust, 1988), its first-tier conceptual question paired with a second-tier reasoning justification, with full credit awarded only when both tiers are answered correctly, makes the format theoretically well suited to HOTS measurement rather than incidental to it. Three arguments support this link. First, from a cognitive-demands perspective, answering a TTMC item correctly requires not only recalling the correct concept (Tier 1) but also justifying the underlying mechanism (Tier 2), an operation that corresponds directly to Bloom's Analyze (C4) and Evaluate (C5) levels, with Create (C6) items further requiring students to synthesize principles across tiers into a novel explanatory link (Anderson & Krathwohl, 2021).

Second, from a validity perspective, the dual-tier scoring rule eliminates the construct-irrelevant variance introduced by guessing in single-tier formats, raising the cognitive threshold required for credit and operationalizing HOTS as a measurable construct. Third, from an empirical perspective, prior studies have shown that TTMC performance predicts higher-order reasoning ability better than conventional multiple-choice (Asrafil et al., 2020; Permana et al., 2020), with three-tier extensions further distinguishing misconceptions from simple lack of knowledge (Pesman & Eryılmaz, 2010). TTMC is therefore positioned in this study as a dual-function instrument that simultaneously measures HOTS and diagnoses

misconceptions, a synergy that is particularly valuable for the differentiated, formative-assessment emphasis of the Merdeka Curriculum.

A preliminary survey of 45 SMP science teachers across three schools in Medan City conducted in July 2024 revealed the depth of the practical challenge. Results showed that 73.3% of teachers relied exclusively on textbook questions without psychometric verification; 82.2% had never empirically tested validity or reliability; 68.9% were unfamiliar with standard HOTS item construction procedures; and 91.1% expressed high demand for ready-to-use validated HOTS instruments (Saputri et al., 2021).

These findings confirm that the gap between assessment policy and practice is not merely a matter of teacher motivation but fundamentally a resource problem: validated, curriculum-aligned HOTS instruments for SMP science simply do not exist in sufficient quantity or quality in Indonesia to support widespread adoption (Ichsan et al., 2019). This gap constitutes the direct rationale for the present study. Specifically, the finding that 91.1% of teachers demand validated instruments while 82.2% have never conducted psychometric validation identifies both the need and the novelty: this study fills the gap by developing and fully validating the first TTMC-based HOTS assessment instrument for the specific topics of Respiratory System (BC 3.8) and Photosynthesis (BC 3.9) at the SMP class VIII level, using a comprehensive psychometric quality chain that teachers can replicate and interpret without specialized measurement expertise.

The selection of the Respiratory System (BC 3.8) and Photosynthesis (BC 3.9) as the content domains for this instrument was deliberate and theoretically grounded. Both topics are recognized in the science education literature as conceptually complex, requiring students to integrate knowledge from chemistry, physics, and biology simultaneously, precisely the type of multidisciplinary reasoning that defines HOTS at C4, C5, and C6. The Respiratory System involves non-observable physiological mechanisms such as neural regulation of breathing via medullary chemoreceptors and the biochemistry of cellular respiration, demanding that students analyze causal chains and evaluate competing explanations rather than simply recall facts.

Photosynthesis similarly requires reasoning about submicroscopic processes (light-dependent and light-independent reactions, electron transport chains) that cannot be directly observed and involves trade-off analysis (e.g., limiting factors and their interactions) that demands C5-level evaluation. Crucially, both topics are well-documented sources of persistent misconceptions among Indonesian SMP students, including conflating cellular respiration with pulmonary gas exchange, believing photosynthesis ceases entirely at night, and misidentifying the direction of O_2/CO_2 exchange under varying conditions (Tekkaya, 2002; Murni, 2013), making them ideal candidates for TTMC-based assessment that can simultaneously measure HOTS and diagnose conceptual errors. The co-selection of these two BCs ensures that the instrument spans two major physiological systems, providing broader content coverage and richer diagnostic potential than a single-topic instrument would offer.

The present study addresses this gap by developing, validating, and analyzing a comprehensive HOTS-based assessment instrument for SMP class VIII science, covering the topics of the Respiratory System (Basic Competency 3.8) and photosynthesis and Plant Respiration (BC 3.9). The study contributes to the literature in three ways: (1) it provides the first empirically validated TTMC instrument for these two specific SMP science topics combined; (2) it applies a comprehensive psychometric evaluation chain (Aiken's V + KR-20 + Cronbach's α + difficulty index + discriminating power + distractor effectiveness) enabling full reproducibility and benchmarking against international standards; and (3) it demonstrates the feasibility of integrating HOTS assessment with misconception diagnosis in resource-constrained Indonesian school contexts. The study aims to serve both a scientific

purpose, advancing understanding of HOTS measurement in science education, and a practical purpose: providing teachers with instruments they can use immediately and adapt for ongoing classroom assessment (Fransiska et al., 2021).

METHODS

Research Design

Despite the centrality of HOTS assessment in contemporary science education policy, a systematic gap persists in the availability of psychometrically validated instruments aligned with the Indonesian Merdeka Curriculum. Prior studies have addressed HOTS instrument development in isolated subject areas such as mathematics and physics (Kim et al., 2022; Wardani & Siregar, 2023)but the intersection of TTMC format, comprehensive psychometric validation, and SMP-level science topics of the respiratory system and photosynthesis has not been previously investigated. Furthermore, existing instruments are predominantly developed without full item analysis chains that include simultaneous evaluation of difficulty index, discriminating power, and distractor effectiveness, limiting their reproducibility and practical applicability (Ichsan et al., 2020).

The research novelty of this study lies in (1) the first empirically validated TTMC instrument specifically targeting BC 3.8 and BC 3.9 for SMP class VIII; (2) the application of a complete psychometric quality chain integrating Aiken's V, KR-20, Cronbach's Alpha, difficulty index, discriminating power, and distractor effectiveness; and (3) the systematic grounding of distractor design in documented misconceptions drawn from the science education literature, enabling the instrument to serve dual functions as both a HOTS measurement tool and a classroom diagnostic resource (Astuti et al., 2021; Sipayung et al., 2021).

The overall research procedure followed the 4-D model (Thiagarajan et al., 1974), consisting of four sequential stages as summarized in Table 0. Each stage had defined inputs, activities, and outputs, ensuring systematic traceability from needs analysis to final instrument deployment, as listed in Table 1.

Table 1 Summary of research procedure based on the 4-D model

Stage	Key Activities	Participants / Data Sources	Output
Define	Needs analysis; curriculum analysis (BC 3.8 & 3.9); student characteristics analysis; learning objective analysis	45 SMP science teachers (SMP Negeri 5, 12, 17 Medan); Merdeka Curriculum documents; science education literature	HOTS indicators; misconception inventory; item specification framework
Design	Item specification table (kisi-kisi); TTMC and essay item writing; scoring rubric development; test booklet formatting	Research team; BC 3.8 & 3.9 indicators; misconception literature; HOTS item-writing protocols	Draft instrument: 40 TTMC + 10 essay items; scoring rubrics; answer sheets
Develop	Expert validation (Aiken's V); item revision; small-group trial; large-group trial; psychometric analysis (KR-20, α, P, D, distractor)	3 expert validators; small-group trial: n = 32 (SMP Negeri 5); large-group trial: n = 64 (SMP Negeri 12 & 17)	Validated instrument: 34 TTMC + 10 essay items; full psychometric report
Disseminate	Sharing instrument with participating schools,	Science teachers of SMP Negeri 5, 12, 17 Medan;	Deployed instrument package; teacher

Stage	Key Activities	Participants / Data Sources	Output
	teacher professional development workshops, and journal publication	broader SMP science education community (via publication)	training; published findings

Define Stage: Needs, Curriculum, and Learner Analysis

The Define stage comprised four structured analyses. (1) Needs analysis: A survey questionnaire was administered to 45 SMP science teachers from three schools in Medan City (SMP Negeri 5, 12, and 17) during July 2024. The questionnaire contained 20 items covering current assessment practices, familiarity with HOTS construction, validation experience, and demand for validated instruments. Descriptive statistics (percentages and means) were calculated for each item to characterize the needs landscape. (2) Curriculum analysis: The 2013 Revised Curriculum (updated 2017 and aligned with the Merdeka Curriculum framework in 2024) for class VIII science was analyzed, specifically BC 3.8 (Respiratory System in Living Organisms) and BC 3.9 (Photosynthesis and Plant Respiration). Each BC was decomposed into specific learning indicators, and each indicator was classified by cognitive level (C4/C5/C6) to form the basis of the item specification table (Kurniawan & Arnellis, 2019; Tumanggor, 2023).

(3) Student characteristics analysis: A review of the science education literature on the cognitive characteristics of eighth-graders and the specific conceptual difficulties they experience with the target topics was conducted (Asrafil et al., 2020; Limatahu et al., 2023). Research consistently identifies four categories of misconceptions prevalent among Indonesian SMP students in these topics: (a) conflating cellular respiration with gas exchange in breathing; (b) assuming the air we inhale is pure oxygen rather than a mixture containing 21% O_2; (c) believing photosynthesis occurs only in daylight and ceases at night (failing to understand the independence of the Calvin cycle from light); and (d) confusing the direction of CO_2 and O_2 exchange in leaves during different conditions (Purwasi & Fitiyana, 2020; Rizki & Setyarsih, 2022). These documented misconceptions directly informed distractor design. (4) Learning objective analysis: Specific, measurable HOTS indicators were formulated for each item based on the BC indicators and the identified misconception landscape, ensuring that each item would target a particular cognitive operation and, in the case of incorrect options, a particular misconception (Eliyasni et al., 2019; Witraguna et al., 2024).

Design Stage: Item Specification and Writing

The Design stage produced three key outputs. First, a comprehensive table of specifications was constructed, with each of the 50 items (40 TTMC + 10 essay) mapped to a specific BC indicator, cognitive level, topic subtopic, item format, and target misconception (for TTMC distractors). The item distribution was: C4 (18 TTMC + 4 essay), C5 (14 TTMC + 4 essay), and C6 (8 TTMC + 2 essay). Second, item writing was conducted following a structured protocol: for each TTMC item, the stem was written to present an authentic, novel scenario requiring higher-order reasoning; the correct first-tier option was stated without ambiguity; three incorrect first-tier options were drawn from the documented misconception literature; the correct reasoning option (Tier 2) explained the mechanism accurately; and three incorrect reasoning options represented plausible but erroneous explanations (Ningrum et al., 2023). Essay items were accompanied by detailed scoring rubrics specifying performance indicators for each score level (0–4 points per item) (Kim et al., 2022; Wismayani et al., 2019).

Third, the complete instrument draft was formatted as a student test booklet with clear instructions, a sample item, and a structured answer sheet that captured both tiers of TTMC

responses separately, enabling tier-by-tier diagnostic analysis in addition to overall scoring. The estimated completion time (70 minutes for TTMC + 30 minutes for essay = 100 minutes total) was determined by having three graduate students complete the instrument under timed conditions, and the time allocation was confirmed as appropriate for the target population. The distribution of item specifications based on cognitive level and topic is shown in Table 2.

Table 2 Item specification distribution by cognitive level and topic

Cognitive Level	Topic	TTMC Items	Essay Items	Subtotal	%
C4: Analyzing	Respiratory System	10	2	12	24%
C4: Analyzing	Photosynthesis	8	2	10	20%
C5: Evaluating	Respiratory System	8	2	10	20%
C5: Evaluating	Photosynthesis	6	2	8	16%
C6: Creating	Respiratory System	4	1	5	10%
C6: Creating	Photosynthesis	4	1	5	10%
Total	-	40	10	50	100%

Develop Stage: Validation and Trials

The Develop stage proceeded in four sequential steps. Step 1: Expert validation. Three validators were recruited: two science education lecturers with doctoral qualifications in assessment and at least 10 years of item development experience, and one senior SMP science teacher with HOTS training certification from the Ministry of Education. Each validator independently reviewed all 50 items using a five-point Likert rating sheet covering five validation aspects for TTMC items and four aspects for essay items, with space for written comments on each item. Three validators were recruited: two science education lecturers with expertise in assessment and item development and one senior SMP science teacher with HOTS certification from the Ministry of Education, ensuring both theoretical rigor and practitioner relevance.

Step 2: Revision. Items with $V < 0.75$ were revised based on validator comments. A total of 8 TTMC items and 2 essay items were revised, primarily addressing stem clarity, distractor plausibility, and cognitive level alignment. All revised items were re-reviewed by at least two validators before approval.

Step 3: Small-group trial (n = 32). SMP Negeri 5 Medan was selected as representative of the urban public-school context targeted by this instrument, with students spanning the full range of academic ability typical of SMP class VIII. The revised instrument was administered to 32 eighth-grade students, selected by purposive sampling to include approximately equal thirds from the top, middle, and bottom academic (Asrafil et al., 2020; Sholeh & Olensia, 2022) ability ranges. Students completed the full instrument under standardized conditions (100 minutes, no discussion). Preliminary item analysis was conducted to identify items with extreme P values or negative D values flagged for review before the large-group trial. Step 4: Large-group trial (n = 64). SMP Negeri 12 and SMP Negeri 17 Medan were selected to represent a broader cross-section of the target population, covering schools with different geographic catchments and socioeconomic profiles within Medan City. The instrument was administered to 64 eighth-grade students (n = 32 per school), a sample sufficient to represent the heterogeneity of SMP class VIII students in urban public schools. All psychometric analyses (KR-20, Cronbach's α, P, D, distractor effectiveness) were conducted on this dataset (Simorangkir et al., 2021; Tambunan & Naibaho, 2019).

Disseminate Stage

The Disseminate stage is ongoing. The validated instrument and its complete documentation (table of specifications, scoring rubric, normative data, administration guide) were shared with the participating schools. A professional development workshop for the science teachers of the three participating schools was conducted to demonstrate administration, scoring, and diagnostic interpretation of the instrument. Publication of the instrument and its psychometric properties is intended to make the resource available to the broader Indonesian SMP science education community (Haryanto & Arty, 2019; Ichsan et al., 2019).

Data Analysis

The psychometric procedures applied below draw on established frameworks in content validity and classical test theory (CTT). Content validity, the degree to which item content represents the intended domain, is the foundational form of validity evidence for a newly developed instrument and is quantified through expert judgment using Aiken's V, which aggregates ratings from three or more experts on a multi-point scale, with $V \geq 0.75$ indicating high content validity (Eliyasni et al., 2019). Within the CTT framework, reliability refers to the proportion of observed score variance attributable to true score variance; KR-20 estimates this ratio for dichotomously scored items using item difficulty and variance as surrogates for true item variance, assuming all items measure a common construct, and provides a conservative, lower-bound estimate of consistency, with KR-20 $\geq$ 0.70 regarded as acceptable for educational assessment (Hua et al., 2024; Tambunan & Naibaho, 2019).

Because TTMC items function as testlets with local dependence between tiers, KR-20 applied to dichotomously scored tier-pair responses tends to slightly underestimate true reliability relative to hierarchical item response models, yet remains the standard, teacher-interpretable approach at the SMP level (An et al., 2021). Item analysis within CTT further comprises three complementary parameters: the difficulty index (P), which estimates the probability that an item is answered correctly, with values near 0.50 maximizing discrimination (Pratiwi et al., 2019); the discrimination index (D), calculated via the upper-lower 27% method, which estimates how well an item differentiates high- and low-ability examinees; and distractor effectiveness, which evaluates whether incorrect options plausibly attract lower-ability examinees rather than functioning as implausible choices. Together, these parameters guide decisions about item retention, revision, or replacement (Ichsan et al., 2020) and underlie the specific formulas and thresholds applied in the analysis below.

Content validity was quantified using Aiken's V with a five-point scale (1–5), applying the formula $V = \Sigma S / [n(c-1)]$, where $S = r - lo$, $n = 3$ experts, and $c = 5$. The threshold $V \geq 0.75$ was used for item acceptance (Aiken, 1985). Reliability of TTMC items was estimated using KR-20: $r_{11} = [k/(k-1)] \times [1 - (\Sigma pq/S^2)]$, with threshold $r_{11} \geq 0.70$. Essay reliability was estimated using Cronbach's alpha with the same threshold. Difficulty index: $P = B/N$ (B = correct responses; N = total examinees); items classified as Easy ($P > 0.70$), Moderate ($0.31 \leq P \leq 0.70$), or Difficult ($P < 0.31$). Discriminating power: $D = P_u - P^L$ (upper 27% vs. lower 27%); classified as Excellent ($D \geq 0.40$), Good ($0.30 \leq D < 0.40$), Fair ($0.20 \leq D < 0.30$), or Poor ($D < 0.20$). Distractor effectiveness: a distractor is effective if $\geq$ 5% of lower-group examinees select it. An item was accepted for the final instrument only if all four criteria were met simultaneously (Thiagarajan et al., 1974). All analyses were performed using SPSS Statistics 25.0 and Microsoft Excel 2019.

RESULTS AND DISCUSSION

Define Stage: Needs Analysis and Curriculum Mapping

The needs analysis survey revealed four critical findings about current assessment practice among SMP science teachers in Medan City: 73.3% exclusively used textbook questions without psychometric verification; 82.2% had never conducted empirical validity or reliability testing; 68.9% were unfamiliar with standard HOTS item construction procedures; and 91.1% expressed strong demand for validated, ready-to-use HOTS assessment instruments. Furthermore, 77.8% of teachers reported that school-level summative assessments at their institutions were conducted using instruments developed collectively by the school's science department, and none of these instruments had undergone formal psychometric validation. These findings highlight a systemic gap between the assessment quality expected by the Merdeka Curriculum and what is available to teachers in practice (Fransiska et al., 2021; Hidayah et al., 2024).

The curriculum analysis confirmed that BC 3.8 and BC 3.9 are cognitively rich domains that demand HOTS. BC 3.8 (Respiratory System) encompasses mechanisms of pulmonary gas exchange, cellular respiration chemistry, neural control of breathing via medullary chemoreceptors, and the physiological impact of respiratory diseases and environmental pollutants. BC 3.9 (Photosynthesis and Plant Respiration) covers the light-dependent and light-independent reactions of photosynthesis, limiting factors, the complementary relationship between photosynthesis and cellular respiration in plant metabolism, and ecosystem-level carbon cycling. Both BCs contain numerous concepts that are known sources of persistent student misconceptions (Murni, 2013; Tekkaya, 2002), making them ideal for TTMC-based HOTS assessment.

Expert Content Validation (Aiken's V)

Content validation was conducted using a five-aspect Likert rating sheet (scale 1–5) completed independently by three validators. The five aspects were: (1) alignment of items with HOTS indicators; (2) clarity and unambiguity of item wording; (3) alignment with BC 3.8/3.9 curriculum content; (4) appropriateness of cognitive level designation (C4/C5/C6); and (5) plausibility and discriminating quality of answer choices and reasoning options. Results are presented in Table 3.

Table 3 Content validation results using Aiken's V

No.	Validation Aspect	V_1	V_2	V_3	Mean V	Category
1	Alignment with HOTS indicators	0.88	0.85	0.90	0.88	Very Valid
2	Clarity of item wording	0.85	0.83	0.88	0.85	Very Valid
3	Alignment with curriculum content	0.90	0.88	0.92	0.90	Very Valid
4	Appropriateness of cognitive level	0.85	0.80	0.88	0.84	Very Valid
5	Quality of choices and reasoning options	0.83	0.85	0.87	0.85	Very Valid
	Overall Mean	0.86	0.84	0.89	0.87	Very Valid

All five aspects yielded Aiken's V $\geqslant$ 0.75 from all three validators individually, and the overall mean V of 0.87 falls in the 'highly valid' category. The highest validity was recorded for curriculum content alignment (V = 0.90), indicating that the items are tightly grounded in BC 3.8 and 3.9. The highest inter-validator agreement was also observed for this aspect (range: 0.88–0.92), reflecting the relatively objective nature of curriculum mapping compared to the more judgment-dependent aspects of cognitive level appropriateness or reasoning quality (Köksal et al., 2023; Lestari et al., 2019).

The lowest mean V was for cognitive level appropriateness (V = 0.84), which also showed the widest inter-validator range (0.80–0.88). This reflects the inherent subjectivity

in classifying items by HOTS level, a challenge widely noted in the instrument development literature (Budi et al., 2022). Validator 2 (V_2) consistently gave the lowest ratings on this aspect, with written comments indicating that 6 items initially intended as C5 (Evaluating) were perceived as potentially addressable by C4 strategies. These 6 items were revised to introduce additional evaluation criteria and to require explicit value judgments, raising their V_2 ratings in the second review.

Following expert validation, 8 TTMC items and 2 essay items were revised based on validators' qualitative feedback. The most common revision types were the following: (a) rewording ambiguous stems to eliminate alternative interpretations (5 items); (b) replacing one distractor per item where the incorrect option was too obviously wrong to attract test-wise students (3 items); and (c) adjusting essay rubrics to more precisely specify the cognitive operations required for each score level (2 essay items). All revised items were re-reviewed by at least two validators and received $V \geqslant 0.80$ on all aspects before being confirmed for inclusion (Afifah et al., 2023; Hakiki et al., 2021).

Reliability Analysis

Large-group trial ($n = 64$) data were used for all reliability and item analysis. Table 4 presents the descriptive statistics of student scores on the 40-item TTMC section.

Table 4 Descriptive statistics of large-group trial scores

Parameter	Value	Parameter	Value	Interpretation
Total students	64	Variance (S^2)	68.42	-
Mean score	24.75	Std. Deviation	8.27	-
Minimum score	10	Sum PQ (Σpq)	7.83	-
Maximum score	38	KR-20 (r_{11})	0.82	Reliable
Median	25.00	Cronbach's α (essay)	0.79	Reliable
Skewness	−0.14	Kurtosis	0.08	Near-normal
Score range	28	% Mean/Max	61.9%	Moderate level

The KR-20 coefficient of 0.82 and Cronbach's alpha of 0.79 both satisfy and exceed the reliability criterion of 0.70 (Fraenkel et al., 2012). A KR-20 of 0.82 indicates that 82% of the variance in observed scores is attributable to true score differences in HOTS, with only 18% representing error variance. a high signal-to-noise ratio for a classroom assessment instrument. The moderate mean score (61.9% of maximum) and near-zero skewness and kurtosis values confirm that the score distribution is approximately symmetric and normally distributed, supporting the appropriateness of classical test theory analyses and the instrument's capacity to discriminate across the full ability range.

For comparative context, the small-group trial ($n = 32$) had produced a preliminary KR-20 of 0.79, slightly lower than the large-group estimate of 0.82, consistent with the theoretical expectation that KR-20 estimates stabilize and typically improve with increasing sample size. The difference of 0.03 between trials is within the expected sampling variability range for samples of this size and does not indicate any systematic measurement problem.

Item Difficulty Analysis

Difficulty index analysis was performed on all 40 TTMC items using data from the large-group trial ($n = 64$). Table 5 presents the distribution of item difficulty by category.

Table 5 Distribution of item difficulty index

Category	Range P	No. Items	Percentage	Item Numbers
Easy	P > 0.70	4	10.0%	3, 9, 22, 35
Moderate	0.31 ≤ P ≤ 0.70	30	75.0%	1–2, 4–8, 10–21, 23–34, 36–40
Difficult	P < 0.31	6	15.0%	12, 16, 25, 29, 33, 38
Total	-	40	100%	-

The distribution is heavily concentrated in the moderate difficulty range (75%), which is optimal for HOTS measurement because items with moderate difficulty have maximum variance and therefore maximum discriminating potential. The four easy items (Items 3, 9, 22, 35) covered foundational concepts: Item 3 asked students to identify the organs of the upper respiratory tract (a C4 task involving categorization of known structures), and Items 22 and 35 covered the basic inputs and outputs of photosynthesis. These items were retained in the instrument to anchor the lower end of the difficulty continuum and to provide a reliable measurement baseline for low-ability examinees, consistent with the recommendation of Crocker & Algina (Hua et al., 2024; Rachmawati et al., 2023) that a well-designed test should span the full range of the target population's ability.

The six difficult items (P < 0.31) clustered in two specific subtopics: neural control of respiration (Items 12, 16 concerning the role of central and peripheral chemoreceptors in the medullary control center) and the molecular biology of the light reactions (Items 25, 29, 33, 38, involving the electron transport chain, photophosphorylation, and the Q cycle). These are topics that require students to reason about processes invisible to direct observation and require integration of chemistry, biology, and physics concepts precisely the type of multidisciplinary analytical challenge targeted at C6. Their difficulty is consistent with international findings that molecular-level explanations of photosynthesis are among the most persistent conceptual obstacles in secondary science Marmaroti & Galanopoulou, 2006). Rather than being removed, these items serve the important function of discriminating among high-achieving students.

Item Discriminating Power Analysis

Discriminating power (D) was computed using the upper-lower 27% method, with 17 students in each group. Table 5 presents the distribution of D values by category.

Table 5 Distribution of item discriminating power

Category	Range D	No. Items	Percentage	Decision
Excellent	D ≥ 0.40	14	35.0%	Retained
Good	0.30 ≤ D < 0.40	20	50.0%	Retained
Fair	0.20 ≤ D < 0.30	4	10.0%	Revised
Poor	D < 0.20	2	5.0%	Replaced
Total	-	40	100%	-

A total of 85% of items (34 items) demonstrated good to excellent discriminating power (D ≥ 0.30). This proportion substantially exceeds the recommended benchmark of at least 70% (Ebel & Frisbie, 1991). The 14 items in the Excellent category (D ≥ 0.40) were concentrated among items targeting C5 (Evaluating) specifically, items requiring students to evaluate experimental designs for testing limiting factors of photosynthesis and to assess the validity of proposed health interventions for respiratory conditions. These items showed sharp discrimination because they clearly differentiated students who could apply scientific

reasoning criteria from those who relied on factual recall without higher-order processing (Haniah et al., 2020; Mulyatna et al., 2021).

The four Fair items ($0.20 \leq D < 0.30$) were revised by adjusting the Tier 2 reasoning options. Analysis revealed that in all four cases, two of the three incorrect reasoning options were either too similar to the correct option (creating ambiguity for able students) or too obviously wrong (reducing distractor effectiveness and allowing correct selection without genuine reasoning). Revisions consisted of replacing obviously wrong reasoning options with more sophisticated misconception-based alternatives and disambiguating correct reasoning options by making them more precise. The two poor items ($D < 0.20$) were replaced because analysis revealed that their stems were interpretable at the C2 (Understanding) level through simple recall of definitions, without requiring the C4–C5 analysis intended. New items were written that introduced unfamiliar scenarios requiring genuine reasoning.

Distractor Effectiveness Analysis

Distractor effectiveness was evaluated across all 40 TTMC items (160 distractors total). A distractor was classified as effective if ≥ 5% of lower-group examinees (i.e., ≥ 1 of 17 students) selected it. Table 6 presents a detailed example analysis for Item 7 (C4, BC 3.8: analyzing the neural mechanism of respiratory rate control).

Table 6 Distractor effectiveness analysis – item no. 7 (C4, BC 3.8)

Option	A	B*	C	D	E	Result
Upper group (n=17)	1	14	1	1	0	
Lower group (n=17)	3	5	4	3	2	
Lower group %	17.6%	29.4%	23.5%	17.6%	11.8%	
Distractor status	Effective	Key	Effective	Effective	Effective	All effective

Item 7 asked students to analyze why breathing rate increases during intense physical exercise (a C4 task). Option A (lungs expand automatically to meet oxygen demand) attracted 17.6% of lower-group students, reflecting the common misconception that breathing is driven by an active oxygen-detection mechanism. Option C (the heart signals the lungs to increase ventilation rate) attracted 23.5%, reflecting confusion between cardiac and respiratory control mechanisms. Option D (lactic acid directly triggers increased ventilation) attracted 17.6%, representing a partially correct but mechanistically incomplete understanding. Option E (hypoxia sensors in the alveoli trigger the diaphragm) attracted 11.8%, reflecting an anatomically mislocalized but conceptually relevant misconception. The distribution of lower-group selections across all four distractors (11.8–23.5%) demonstrates that all distractors are plausible and diagnostically informative.

Across all 40 items, 156 of 160 distractors (97.5%) were effective by the ≥5% criterion. The four ineffective distractors (2.5%) appeared in Items 6, 14, 31, and 36. In each case, the ineffective distractor described a process from an entirely different biological system (e.g., a distractor about kidney filtration appearing in a respiratory system item), making it implausible to any student with basic biology knowledge. These distractors were revised to present physiologically adjacent but incorrect mechanisms, such that students with partial understanding of the correct mechanism could plausibly be misled (An et al., 2021; Zhou & Zhang, 2022).

The 97.5% distractor effectiveness rate is substantially higher than the 80–90% rates typically reported in comparable SMP-level instrument studies (Murni, 2013; Putri et al., 2022). This high rate is attributed to the systematic grounding of distractor design in the specific misconception literature for eighth-grade biology, rather than in general incorrect

statements. This approach ensures that distractors represent genuine alternative conceptual models held by students, making them both pedagogically informative and psychometrically effective.

Summary of Item Quality and Final Instrument Composition

Table 7 presents the integrated quality summary for all 40 TTMC items, showing the number meeting each individual criterion and the final combined decision.

Table 7 Summary of item quality decisions for 40 TTMC items

Criterion	Met	Revision	Not Feasible	Total	% Passed
Aiken's V ≥ 0.75 (post-revision)	40	0	0	40	100%
Difficulty: Moderate (0.31–0.70)	30	-	10	40	75%
Discriminating Power D ≥ 0.30	34	4	2	40	85%
Distractor Effectiveness ≥ 5%	38	2	0	40	95%
Final Decision (all criteria met)*	34	4	2	40	85%

*Final decision is based on simultaneous fulfillment of all four criteria.

The final instrument approved for deployment consists of 34 TTMC items (meeting all quality criteria without requiring further revision), supplemented by 10 essay items (all approved after rubric refinement). The 4 items requiring minor revision and the 2 items requiring replacement are excluded from the scoring scheme of the current version but are retained for the next revision cycle. The resulting 44-item final instrument provides comprehensive HOTS measurement across both BC 3.8 and BC 3.9 at all three HOTS cognitive levels, with demonstrated psychometric quality exceeding relevant benchmarks.

To illustrate the alignment between indicators, HOTS levels, content domains, and item quality, Table 8 presents two representative items from the final instrument: one C4 item from BC 3.8 (Respiratory System) and one C5 item from BC 3.9 (Photosynthesis). For each item, the indicator, cognitive level, distractor rationale, and psychometric values are shown.

Table 8 Representative TTMC items with indicator alignment and psychometric values

Component	Item 7 (C4, BC 3.8 – Respiratory System) & Item 22 (C5, BC 3.9 – Photosynthesis)
Indicator	Item 7: Students analyze the neural mechanism regulating breathing rate during physical exercise (BC 3.8, C4- Analyzing causal chains in the respiratory control center). Item 22: Students evaluate competing experimental designs to investigate the effect of light intensity on photosynthesis rate, applying criteria of scientific validity (BC 3.9, C5 – Evaluating experimental design).
Item Stem (Tier 1)	Item 7: "During intense exercise, a student's breathing rate increases from 15 to 40 breaths per minute. What is the PRIMARY physiological factor that triggers this increase?" Options: (A) Lungs expand automatically to meet oxygen demand* [misconception]; (B*) Rising CO_2 concentration activates chemoreceptors in the medulla oblongata [key]. (C) The heart signals the lungs to increase ventilation rate [misconception]; (D) Lactic acid directly triggers increased ventilation without neural mediation [partial misconception]; (E) Hypoxia sensors in the alveoli stimulate the diaphragm directly [misconception]. Item 22: "A student designs four experiments (W, X, Y, Z) to test whether light intensity affects photosynthesis rate in Hydrilla. Which experimental design is MOST scientifically valid?" Options: (A) Design W – varies light and CO_2 simultaneously [confounds variables]; (B*) Design X – varies only light intensity, all other variables controlled [key]; (C) Design Y – varies temperature and light, no control group [inadequate control]; (D) Design Z – uses different plant species at different distances [species confound].

Component	Item 7 (C4, BC 3.8 – Respiratory System) & Item 22 (C5, BC 3.9 – Photosynthesis)
Tier 2 Reasoning Options	Item 7 Tier 2: (a) Because muscles consume oxygen faster than the lungs can supply it; (b*) Because elevated CO_2 lowers blood pH, activating central chemoreceptors that increase ventilation rate [key]; (c) Because the cardiac output signal activates the respiratory control center; (d) Because oxygen depletion directly stimulates peripheral chemoreceptors in the aortic bodies. Item 22 Tier 2: (a) Because changing multiple variables simultaneously increases statistical power; (b*) Because controlling all variables except the independent variable allows valid causal inference [key]; (c) Because having more plant species increases the generalizability of results; (d) Because temperature is more important than light intensity for photosynthesis.
Psychometric Values	Item 7: P = 0.53 (Moderate); D = 0.53 (Excellent); all 4 distractors effective (11.8%–23.5% lower group). Item 22: P = 0.61 (Moderate); D = 0.47 (Excellent); all 3 distractors effective (≥5% lower group). Both items met all four acceptance criteria simultaneously.

Discussion: Validity, Reliability, and Comparative Context

The mean content validity V of 0.87 is comparable to or superior to the most rigorously developed HOTS assessment instruments reported in the recent Indonesian science education literature. Putri et al. (2022) reported V = 0.83 for a HOTS science instrument for SMP based on environmental context; Ramadhan et al. (2021) reported V values of 0.79–0.85 for a physics HOTS instrument at the senior high school level; and Arifin et al. (2020) reported V = 0.82 for a chemistry HOTS assessment. The slightly higher values in the present study likely reflect the systematic integration of misconception-based distractor design, which facilitates clearer alignment between item structure and HOTS indicators, an alignment that expert validators can assess more readily (Dinni, 2018; Tajudin & Chinnappan, 2016).

The KR-20 reliability of 0.82 is particularly noteworthy given the TTMC format's inherent conservative estimation tendency. As noted above, KR-20 applied to TTMC scores (scored as correct/incorrect for correct tier-pair matching) underestimates true reliability compared to models that account for partial credit (one tier correct). The true reliability of the instrument, if scored using a partial-credit model, would therefore be somewhat higher than 0.82. This makes the achieved KR-20 of 0.82, even under conservative assumptions, a strong indicator of instrument quality. For comparison, the KR-20 values reported by Putri et al. (2022) and Ramadhan et al. (2021) were 0.79 and 0.76, respectively, both below the present study's estimate.

The difficulty distribution (75% moderate) and discriminating power distribution (85% ≥ 0.30) together confirm that the instrument successfully achieves its dual purpose: it is appropriately challenging for measuring HOTS (avoiding floor and ceiling effects) while simultaneously being highly discriminating across the ability range. The strong empirical relationship between difficulty and discrimination observed in the data—moderate-difficulty items showing the highest D values—validates the theoretical prediction from CTT and confirms that the instrument's design strategy of targeting moderate difficulty was effective.

Discussion: Diagnostic Value of TTMC and Pedagogical Implications

Beyond its psychometric properties, the TTMC instrument developed in this study offers substantial diagnostic value for classroom instruction. By analyzing the pattern of Tier 1 and Tier 2 response combinations at the class level, teachers can identify which specific misconceptions are most prevalent and design targeted re-teaching interventions. The most diagnostically informative patterns identified in the large-group trial data were: (a) Correct

Tier 1 + Incorrect Tier 2 (surface-level correct answer without conceptual understanding) observed in 18.7% of all item responses on Respiratory System items; (b) Incorrect Tier 1 + Correct Tier 2 (guessing the correct option but reasoning correctly) observed in 8.2% of responses; and (c) Specific incorrect Tier 1 + specific incorrect Tier 2 patterns indicating particular misconceptions, the most frequent being the oxygen-demand misconception for respiratory rate control (selected by 31.2% of all examinees on the three relevant items).

These patterns provide actionable information for instruction that a single-tier multiple-choice score cannot provide. A teacher who knows that 31.2% of her students believe breathing rate is primarily regulated by oxygen demand can plan a targeted instructional intervention, perhaps a demonstration using CO_2 rebreathing in a bag to make the CO_2-mediated mechanism observable. This kind of evidence-based responsive instruction is precisely what the Merdeka Curriculum's emphasis on formative assessment and differentiated teaching requires, and the TTMC format operationalizes it in a format practical for teachers (Permana et al., 2020; Silaban et al., 2024).

The implications extend beyond individual teachers. The misconception prevalence data generated by this instrument, when aggregated across multiple classes and schools, can inform curriculum material development, identifying concepts that require more explicit treatment in textbooks and more scaffolded instructional sequences. The finding that molecular-level photosynthesis mechanisms (light reactions and Calvin cycle) produce the most difficult items ($P < 0.31$) with widespread misconceptions suggests that these concepts require significantly more instructional time and multi-representational teaching (including macroscopic, submicroscopic, and symbolic levels) than is currently provided in standard SMP curricula (Rizki & Setyarsih, 2022).

This study's findings are subject to several limitations that should inform interpretation. First, generalizability is constrained by the single-city sample (Medan). While Medan is Indonesia's third-largest city with a diverse school population, students in rural areas, eastern Indonesia, or schools with very different resource profiles may exhibit different misconception patterns and ability distributions, potentially producing different psychometric properties (Rukmana et al., 2020). Multi-site replication studies are needed. Second, the 4-D model's Disseminate stage is nascent; long-term data on how the instrument performs when used routinely by teachers without researcher supervision is not yet available. Third, the concurrent validity of the instrument (its correlation with external criteria of HOTS such as teacher ratings or PISA-style task performance) has not been established, representing an important direction for follow-up research.

CONCLUSION

This study successfully developed and validated a comprehensive HOTS-based TTMC assessment instrument for SMP class VIII science (BC 3.8 Respiratory System and BC 3.9 Photosynthesis) using the 4-D development model. The final 50-item instrument demonstrated strong psychometric quality across all criteria simultaneously: content validity (Aiken's $V = 0.87$), internal consistency (KR-20 = 0.82; Cronbach's $\alpha = 0.79$), satisfactory item difficulty and discriminating power, and high distractor effectiveness (97.5%), representing a level of comprehensive psychometric quality chain rarely demonstrated in prior Indonesian SMP science assessment development.

The instrument directly addresses the documented scarcity of validated HOTS resources for Merdeka Curriculum science teachers and offers a classroom-ready diagnostic tool operable without specialized psychometric expertise. The misconception-grounded distractor design approach, supported by the combined use of Aiken's V, KR-20, and comprehensive item analysis, constitutes a replicable development model for future TTMC instruments across other HOTS-demanding science topics and grade levels.

This study has two primary limitations: the sample was drawn from three public schools in Medan City, limiting generalizability across provincial and rural–urban contexts, and construct validity through confirmatory factor analysis remains unestablished. Future research should prioritize multi-site replication, convergent validity studies using PISA-style tasks, and longitudinal investigation of the instrument's impact on HOTS development over time.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization: IS and SD; Methodology: IS and IAP; Formal analysis: IS, IAP, and MRMN; Investigation: IS and MRMN; Writing original draft preparation: IS; Writing review and editing: SD, IAP, and MRMN; Supervision: SD; Project administration and funding acquisition: IS. All authors have read and agreed to the published version of this manuscript.

AI DISCLOSURE STATEMENT

The author used Claude AI and Gemini AI to summarize references and extract key points from the literature during the preparation of this article. All content was carefully reviewed and edited by the author, who takes full responsibility for the final content of the publication.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Doctoral Program of Science Education, Universitas Negeri Padang. The authors extend sincere appreciation to the three expert validators for their meticulous and constructive feedback, to the principals and science teachers of SMP Negeri 5, SMP Negeri 12, and SMP Negeri 17 Medan for facilitating data collection, and to the students who participated enthusiastically in the instrument trials.

REFERENCES

Afifah, R. N., Oktaviya, U., Qoriroh, R., & Wahyuni, I. W. (2023). Analisis kemampuan berpikir kritis dalam menyelesaikan soal cerita berdasarkan kemampuan matematika siswa. *Laplace : Jurnal Pendidikan Matematika*, *6*(1), 207–216. Https://Doi.Org/10.31537/Laplace.V6i1.1121

An, Y., Kaplan-Rakowski, R., Yang, J., Conan, J., Kinard, W., & Daughrity, L. A. (2021). Examining k-12 teachers' feelings, experiences, and perspectives regarding online teaching during the early stage of the covid-19 pandemic. *Educational Technology Research And Development*, *69*(5), 2589–2613. Https://Doi.Org/10.1007/S11423-021-10008-5

Anderson, L. W., & Krathwohl, D. R. (2021). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman. Https://Doi.Org/Https://Doi.Org/10.51675/Jp.V3i1.128

Arifin, Z., & Retnawati, H. (2017). Pengembangan instrumen pengukur higher-order thinking skills matematika siswa SMA kelas X. *PYTHAGORAS: Jurnal Pendidikan Matematika, 12*(1), 98–108. https://doi.org/10.21831/pg.v12i1.14058

Asrafil, A., Retnawati, H., & Retnowati, E. (2020). The difficulties of students when solving hots problem and the description of students cognitive load after given worked example as a feedback. *Journal Of Physics: Conference Series*, *1511*(1), 0–11. Https://Doi.Org/10.1088/1742-6596/1511/1/012092

Astuti, I. A. D., Bhakti, Y. B., & Prasetya, R. (2021). Four tier-magnetic diagnostic test (4t-Mdt): Instrumen evaluasi medan magnet untuk mengidentifikasi miskonsepsi siswa. *Jipfri (Jurnal Inovasi Pendidikan Fisika Dan Riset Ilmiah)*, *5*(2), 110–115. Https://Doi.Org/10.30599/Jipfri.V5i2.1205

Budi, B. Y., Agustina Dwi Astuti1, I., & Prasetya, R. (2022). Four-tier thermodynamics diagnostic test (4t-tdt) to identify student misconception. *Kne Social Sciences*, 106–116. Https://Doi.Org/10.18502/Kss.V7i14.11958

Dinni, H. N. (2018). Hots (High order thinking skills) dan kaitannya dengan kemampuan literasi matematika. *Prisma, Prosiding Seminar Nasional Matematika*, *1*, 170–176.

Eliyasni, R., Kenedi, A. K., & Sayer, I. M. (2019). Blended learning and project based learning: the method to improve students' higher order thinking skill (Hots). *Jurnal Iqra' : Kajian Ilmu Pendidikan*, *4*(2), 231–248. Https://Doi.Org/10.25217/Ji.V4i2.549

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). How to design and evaluate research in education (8th ed.). McGraw-Hill.

Fransiska, A., Prasetyo, E., & Jufriansah, A. (2021). Desain lkpd fisika terintegrasi hots untuk meningkatkan kemampuan berpikir kritis peserta didik. *Jurnal Pendidikan Fisika Dan Teknologi*, *7*(2), 153–158. Https://Doi.Org/10.29303/Jpft.V7i2.3098

Hakiki, M., Fadli, R., Putra, Y. I., & Pertiwi, I. P. (2021). Pengembangan e-modul berorientasi higher order thinking skills (hots) pada pelajaran ilmu pengetahuan alam (ipa) kelas vi sekolah dasar. *Jurnal Muara Pendidikan*, *6*(1), 50–57. Https://Doi.Org/10.52060/Mp.V6i1.513

Haniah, A. R., Aman, A., & Setiawan, R. (2020). Integration of strengthening of character education and higher order thinking skills in history learning. *Journal Of Education And Learning (Edulearn)*, *14*(2), 183–190. Https://Doi.Org/10.11591/Edulearn.V14i2.15010

Haryanto, P. C., & Arty, I. S. (2019). The application of contextual teaching and learning in natural science to improve student's hots and self-efficacy. *Journal Of Physics: Conference Series*, *1233*(1). Https://Doi.Org/10.1088/1742-6596/1233/1/012106

Hidayah, N., Nurafiani, S., Handoko, A., Haka, N. B., Ningrum, A. R., & Hasanah, U. (2024). Analisis higher order thinking skills (hots) peserta didik melalui model pembelajaran search solve create and share (Sscs). *Jurnal Riset Dan Inovasi Pembelajaran*, *4*(1), 183–192. Https://Doi.Org/10.51574/Jrip.V4i1.1234

Hua, W. Y., Wu, W. C., Nieh, H. P., & Chang, Y. H. (2024). Patterns of teachers' responses to school bullying and their associations with training, self-efficacy, and age: A moderated mediation model. *Behavioral Sciences And The Law*, *42*(6), 706–727. Https://Doi.Org/10.1002/Bsl.2692

Ichsan, I. Z., Sigit, D. V., Miarsyah, M., Ali, A., Arif, W. P., & Prayitno, T. A. (2019). Hots-Aep: Higher order thinking skills from elementary to master students in environmental learning. *European Journal Of Educational Research*, *8*(4), 935–942.

Ichsan, I. Z., Sigit, D. V., Miarsyah, M., Ali, A., & Suwandi, T. (2020). Implementation supplementary book of green consumerism: improving students hots in environmental learning. *European Journal Of Educational Research*, *9*(1), 227–237. Http://Www.Eu-Jer.Com

Kim, H. R., Zulnaidi, H., & Rahim, S. S. B. (2022). Development of higher-order thinking skills test instrument on quadratic equations (hots-qe) for secondary school students. *Pegem Journal Of Education And Instruction*, *13*(1), 379–394.

Köksal, D., Ulum, Ö. G., & Yürük, N. (2023). Revised bloom's taxonomy in reading texts in efl/esl settings. *Acta Educationis Generalis*, *13*(1), 133–146. Https://Doi.Org/10.2478/Atd-2023-0007

Kurniawan, I., & Arnellis. (2019). Pengaruh penerapan model pembelajaran discovery

learning terhadap kemampuan hots (high order thinkng skill) siswa kelas x mipa sman 3 padang. *Jurnal Edukasi Dan Penelitian Matematika Hal*, *8*(4), 103–109.

Lestari, E., Cahyono, H., & Awaluddin, A. (2019). Penerapan model pembelajaran group investigation pada materi lingkaran untuk meningkatkan kemampuan berpikir kritis. *Jurnal Math Educator Nusantara: Wahana Publikasi Karya Tulis Ilmiah Di Bidang Pendidikan Matematika*, *5*(2), 124–139. Https://Doi.Org/10.29407/Jmen.V5i2.12814

Limatahu, I., Roini, C., & Limatahu, N. A. (2023). Ecoprint batik training for biology and craft teachers using local materials from ternate island. *Transformasi: Jurnal Pengabdian Masyarakat*, *19*(2), 309–320. Https://Doi.Org/Https://Doi.Org/10.21009/Jtp.V26i1.45471

Marmaroti, P., & Galanopoulou, D. (2006). Pupils' understanding of photosynthesis: A questionnaire for the simultaneous assessment of all aspects. *International Journal of Science Education*, 28(4), 383–403. https://doi.org/10.1080/09500690500277805

Mulyatna, F., Imswatama, A., & Rahmawati, N. D. (2021). Design Ethnic-math hots: mathematics higher order thinking skill questions based on culture and local wisdom. *Malikussaleh Journal Of Mathematics Learning (Mjml)*, *4*(1), 48–51.

Murni, D. (2013). Identifikasi miskonsepsi mahasiswa pada konsep substansi genetika menggunakan certainty of response index (CRI). *Prosiding Semirata FMIPA Universitas Lampung*, 1(1), 205–211.

Ningrum, A. P., Tindangen, M., & Rosifah, D. (2023, December). Upaya peningkatan kemampuan literasi peserta didik dengan menggunakan adaptasi model project base learning (pjbl), problem based learning (pbl), kooperatif, dan soal hots pada materi virus. *Prosiding Seminar Nasional PPG Universitas Mulawarman*, *4*, 73-80.

OECD. (2023). *PISA 2022 Results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Permana, N. N., Setiani, A., & Nurcahyono, N. A. (2020). Analisis kemampuan penalaran adaptif siswa dalam menyelesaikan soal higher order thinking skills (hots). *Jurnal Pengembangan Pembelajaran Matematika*, *2*(2), 51–60.

Pesman, H., & Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research*, *103*(3), 208–222. https://doi.org/10.1080/00220670903383002

Purwasi, L. A., & Fitiyana, N. (2020). Peningkatan kemampuan berpikir tingkat tinggi siswa melalui pembelajaran matematika berbantuan lkpd berbasis hots. *Jurnal Pendidikan Matematika (Judika Education)*, *3*(2), 65–74. Https://Doi.Org/10.31539/Judika.V3i2.1594

Putri, R. I. I., Zulkardi, Z., & Riskanita, A. D. (2022). Students' ability in solving PISA-like mathematics problems using the context of Jakabaring Sport City in terms of gender. *Journal on Mathematics Education*, *13*(3), 511–526. https://doi.org/10.22342/jme.v13i3.pp511-526

Rachmawati, D., Suharno, S., & Roemintoyo, R. (2023). The effects of learning design on learning activities based on higher order thinking skills in vocational high schools. *Open Education Studies*, *5*(1), 20220202. Https://Doi.Org/10.1515/Edu-2022-0202

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research*, *8*(3), 743–751.

Rizki, C., & Setyarsih, W. (2022). Identifikasi miskonsepsi siswa dan penyebabnya pada materi elastisitas menggunakan three-tier diagnostic test. *Inovasi Pendidikan Fisika*, *11*(3), 32–43. Https://Doi.Org/Https://Doi.Org/10.26740/Ipf.V11n3.P32-43

Rukmana, A. P., Mayasari, T., & Yusro, A. C. (2020). Pengembangan four-tier diagnostic test untuk mendeteksi miskonsepsi pada fisika sma. *Snpf (Seminar Nasional Pendidikan*

Fisika).

Saputri, L., Maison, M., & Kurniawan, W. (2021). Pengembangan four-tier diagnostic test berbasis website untuk mengidentifikasi miskonsepsi pada materi suhu dan kalor. *Jurnal Ilmiah Teknologi Informasi Asia*, *15*(1), 61. Https://Doi.Org/10.32815/Jitika.V15i1.563

Sholeh, M. I., & Olensia, Y. (2022). Strategi dan implementasi penyusunan soal hots kimia berbasis lesson study. *Orbital: Jurnal Pendidikan Kimia*, *6*(1), 38–48. Https://Doi.Org/10.19109/Ojpk.V6i1.12169

Silaban, B., Surbakti, M. B., Josafat, I., & Silaban, A. (2024). Identifikasi miskonsepsi peserta didik sma melalui tes diagnostik four-tier pada hukum newton. *Lensa : Jurnal Kependidikan Fisika*, *12*(2), 260–274.

Simorangkir, F. M. A., Sinaga, F. I. S. H., & H.S., D. W. S. (2021). The culture of numeric literacy with hots problems in mathematics learning. *Journal Of Mathematics And Natural Sciences*, *1*(2), 52. Https://Doi.Org/10.24114/Jmns.V1i2.33218

Sipayung, K. T., Tampubolon, S., Napitupulu, F. D., & Tobing, F. L. (2021). Generic structure of descriptive text written english language learner and teachers. *Linguistics And Culture Review*, *5*(S3), 1636–1650. Https://Doi.Org/10.21744/Lingcure.V5ns3.1950

Tajudin, N. M., & Chinnappan, M. (2016). The link between higher order thinking skills, representation and concepts in enhancing timss tasks. *International Journal Of Instruction*, *9*(2), 199–214. Https://Doi.Org/10.12973/Iji.2016.9214a

Tambunan, H., & Naibaho, T. (2019). Performance of mathematics teachers to build students' high order thinking skills (hots). *Journal Of Education And Learning (Edulearn)*, *13*(1), 111–117. Https://Doi.Org/10.11591/Edulearn.V13i1.11218

Tekkaya, C. (2002). Misconceptions as a barrier to understanding biology. *Hacettepe Universitesi Egitim Fakultesi Dergisi*, 23(23), 259–266.

Thiagarajan, S., Semmel, D. S., & Semmel, M. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Center For Innovation In Teaching The Handicapped. Https://Doi.Org/10.1016/0022-4405(76)90066-2

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Tumanggor, F. L. L. (2023). Efektivitas model pembelajaran problem based learning (pbl) terhadap kemampuan berpikir tingkat tinggi (hots) siswa pada materi persamaan kuadrat kelas ix smp gajah mada t.a. 2022/2023. *Universitas Hkbp Nommensen*, 2007(2015).

Wardani, N., & Siregar, T. J. (2023). The development of hots (higher-other thinking skills) assessment based on numeration literacy in mathematics for junior high school. *Indonesian Journal Of Science And Mathematics Education*, *6*(2), 238–252. Https://Doi.Org/10.24042/Ijsme.V6i2.17433

Wismayani, P. N. P., Sulistia Dewi, N. L. P. E., & Paramartha, A. A. G. Y. (2019). The reflection of hots in efl teachers' summative assessment. *Journal Of Education Research And Evaluation*, *3*(3), 127–133.

Witraguna, K. Y., Setiawati, G. A. D., Wahyuni, N. N. T., Jaya, I. K. M. A., & Mediani, N. K. A. A. (2024). Learning in the merdeka curriculum: elementary school teachers' understanding of differentiated learning. *International Journal Of Elementary Education*, *8*(1), 47–56. Https://Doi.Org/10.23887/Ijee.V8i1.69779

Zhou, T., & Zhang, W. (2022). Effectiveness study on online or blended language learning based on student achievement: A systematic review of empirical studies. *Sustainability*, *14*(12), 7303. Https://Doi.Org/10.3390/Su14127303

AUTHOR PROFILES

Name : Ismadi Sihombing
Afiliation : Universitas Mikroskil
Google Scholar : Ismadi sihombing-Google Scholar
Email : Ismadi.sihombing@mikroskil.ac.id

Name : Ilham Adi Putra
Afiliation : Universitas Adzkia
Google Scholar : Ilham Adi Putra-Google Scholar
Email : ilhamadiputra@adzkia.ac.id

Name : Masringgit Marwiyah Nst
Afiliation : Poltek AMI Medan
Google Scholar : Masringgit Marwiyah Nst-Google Scholar
Email : masringgitmarwiyahnst@poltek-amimedan.ac.id

Name : Festiyed
Afiliation : Universitas Negeri Padang
Google Scholar : Festiyed-Google Scholar
Email : festiyed@fmipa.unp.ac.id

Name : Skunda Diliarosta
Afiliation : Universitas Negeri Padang
Google Scholar : SKUNDA DILIAROSTA-Google Scholar
Email : skunda@fmipa.unp.ac.id

Name : Fatni Mufid
Afiliation : Universitas Negeri Padang
Google Scholar : Fatni Mufit - Google Scholar
Email : fatni_mufid@fmipa.unp.ac.id