Wahana Pendidikan Fisika (2026) 11 (1): 28-52
https://doi.org/10.17509/wapfi.111.9
2406-9205 (Print) | 2477-4820 (Online)
ORIGINAL RESEARCH

Mapping higher-order thinking skills in physics education students on temperature and heat concepts using a one-tier diagnostic test

M. Chilmy Rosyadi, Himawan Putranta

Received: 16 January 2026 · Accepted: 15 March 2026 · Published Online: 26 March 2026

Abstract
The cultivation of Higher Order Thinking Skills (HOTS) in prospective physics educators is a critical imperative for 21st-century higher education. This quantitative descriptive study maps the HOTS profiles of Physics Education students (N=30) on the misconception-prone topics of temperature and heat, while examining gender-based performance differentials. Data were collected using a validated one-tier diagnostic instrument (Aiken's V = 0.97–0.99; Cronbach's α = 0.601) aligned with Bloom's revised taxonomy cognitive domains of analyzing (C4), evaluating (C5), and creating (C6). Descriptive statistical analyses revealed a moderate aggregate HOTS profile (M = 5.59, SD = 2.21 out of 10). Notably, female students demonstrated superior overall performance (M = 6.47, SD = 1.51) with greater consistency compared to their male counterparts (M = 4.80, SD = 2.48). However, item-level analysis identified a specific deviation wherein male students outperformed females on a technical thermal conductivity task, suggesting that observed gender gaps may partially reflect differential cognitive style alignments with narrative-based instrument formats. These findings underscore the necessity for inquiry-based instructional reforms that deliberately cultivate analytical reasoning across diverse cognitive modalities while accommodating varied learning styles. The demonstrated efficacy of the validated one-tier test offers a pragmatic diagnostic model, though expanded investigations with larger samples and multi-tiered approaches remain warranted to capture students' cognitive profiles comprehensively.

Keywords: Gender analysis · higher-order thinking skills · one-tier diagnostic test · physics education students · temperature and heat concepts.

INTRODUCTION

The escalating demands of 21st-century education necessitate a fundamental paradigm shift from the rote transmission of factual knowledge towards the deliberate cultivation of complex cognitive competencies (OECD, 2019). In the contemporary global landscape, mere procedural proficiency is insufficient; graduates are required to demonstrate mastery of the 4Cs, which are critical thinking, creativity, collaboration, and communication, to navigate the complexities of the Fourth Industrial Revolution. Brookhart (2010) posited that modern educational systems must transcend the goal of producing individuals adept at memorization, instead fostering higher-order thinking skills (HOTS) that enable graduates to deconstruct ill-structured problems and synthesize novel solutions. This imperative is particularly pronounced within higher education, which bears the critical responsibility of facilitating students' ascension

✉ Himawan Putranta
himawan.putranta@uin-suka.ac.id

Physics Education Study Program, Faculty of Tabiyah and Teacher Training, UIN Sunan Kalijaga Yogyakarta, Yogyakarta, Indonesia

How to Cite: Rosyadi, M. C. & Puranta, H., (2026). Mapping higher-order thinking skills in physics education students on temperature and heat concepts using a one-tier diagnostic test. *Wahana Pendidikan Fisika, 11*(1), 28-52. https://doi.org/10.17509/wapfi.111.9

through the upper echelons of cognitive processing, namely, analysis, evaluation, and creation. Leslie (2016) further underscored that the integration of HOTS into the curriculum is not merely an enhancement but a fundamental necessity to maintain educational relevance in an era where data-driven decision-making surpasses the value of algorithmic task execution. Consequently, the initial step in addressing these educational mandates involves a rigorous examination of the theoretical frameworks that define and operationalize high-level thinking.

Theoretically, higher-order thinking skills are conceptualized as the mental manipulation of information to construct new meanings or resolve non-routine problems (Resnick, 1987). Grounded in the revised Bloom's Taxonomy, HOTS occupies the apex of the cognitive domain, encompassing the abilities to analyze (C4), evaluate (C5), and create (C6), thereby distinguishing itself from lower-order thinking skills (LOTS) that are confined to remembering and understanding (Krathwohl, 2002; Leslie, 2016). This cognitive architecture requires students to critically appraise the validity of arguments, forge connections between disparate concepts, and generate generalizations anchored in empirical evidence. Gradini (2019) emphasized that HOTS is intrinsically linked to metacognitive capacity, wherein an individual maintains conscious awareness and regulatory control over their cognitive processes when confronted with intellectual challenges. Zohar & Agmon (2018) argued that within science education, the development of HOTS must be explicitly scaffolded, as these advanced skills do not emerge organically but require deliberate cultivation through cognitively demanding problem stimuli. In the specific discipline of physics, mastery of HOTS is foundational, as the field explicates natural phenomena through causal logic, abstract reasoning, and mathematical modeling. Mestre et al. (2011) contended that effective physics pedagogy must extend beyond algorithmic proficiency to foster deep conceptual reasoning, requiring students to transfer abstract principles to real-world contexts. This cognitive operation is heavily reliant on analytical and evaluative faculties. Despite this theoretical clarity, research by Bhakti et al. (2024) reveals a persistent disconnect, which physics instruction frequently devolves into routine drill and practice, which fails to stimulate HOTS, leaving students proficient in formula manipulation yet incapable of articulating the physical significance underlying numerical results or applying concepts to unfamiliar phenomena.

The urgency of this issue is magnified when the learners in question are prospective physics educators. Hajaroh (2022) elucidated that pre-service teachers must develop robust pedagogical content knowledge (PCK), within which HOTS proficiency is a pivotal component. If teacher candidates themselves struggle with higher-order cognition, they will be ill-equipped to design challenging learning environments or effectively facilitate their future students' critical thinking development (Shulman, 1986). The quality of these prospective teachers' conceptual mastery directly forecasts the scientific literacy of the subsequent generation. Therefore, mapping the HOTS profiles of physics education students transcends mere academic evaluation. It serves as a critical indicator of professional readiness to cultivate an intellectually competent citizenry. This professional readiness is frequently compromised when teacher candidates encounter physics topics characterized by high conceptual density and susceptibility to persistent misconceptions, such as the interrelated concepts of temperature and heat.

The domain of temperature and heat presents a uniquely fertile ground for assessing HOTS, as it encompasses abstract concepts that frequently conflict with intuitive, everyday preconceptions, such as the fundamental distinction between temperature and heat or the microscopic underpinnings of internal energy. Nurul et al. (2016) documented that this topic ranks among the most prevalent sources of misconceptions at both the secondary and tertiary levels, particularly concerning the microscopic mechanisms of heat transfer. Taqwa et al. (2019) found that students often struggle to correlate macroscopic observables, like temperature, with microscopic variables, such as particle kinetic energy, a task that demands robust analytical reasoning. Kotsis et al. (2023), in a recent investigation, reaffirmed that misperceptions regarding thermal equilibrium frequently persist through university education, underscoring the necessity for evaluative instruments that probe beyond mere recall to assess students' capacity for evaluating thermal phenomena. Moreover, the scholarly discourse has increasingly focused on the intersection of cognitive performance and gender within physics education. While earlier literature often reported a gender gap favoring male students in visuospatial tasks and mechanical problem-solving (Madsen et al. 2013), contemporary meta-analyses are challenging this stereotype. Santoso et al. (2024) demonstrated that female students exhibit equal or superior potential in analytical rigor and verbal reasoning when assessments are designed with inclusivity in mind. Sepriyanti et al. (2022) also reported that within 21st-century learning frameworks, female students demonstrated significant gains in HOTS. These divergent findings, as highlighted by recent research on gender and thermal concepts (Kotsis et al., 2023) and studies showing nuanced differences in mathematical representation (Miyaqi et al., 2024), indicate that the HOTS-by-gender profile remains an open area of inquiry, highly contingent upon material characteristics and instrument design.

Notwithstanding the acknowledged importance of HOTS and the pedagogical complexity of temperature and heat, a significant lacuna persists in evaluation practices within higher education. Mestre et al. (2011) critically observed that most physics assessment instruments remain dominated by memory-based items or simple formula application, rendering them incapable of authentically capturing HOTS. While complex multi-tier diagnostic tests offer diagnostic depth, their practicality for rapid, large-scale mapping is limited (Gurel et al. 2015). Conversely, one-tier tests are often undervalued, even though when rigorously developed with explicit HOTS indicators and substantiated through expert validation, including content validity indices such as Aiken's V and reliability coefficients, they can provide both objectivity and efficiency (Retnawati, 2016). The development of valid and reliable instruments is paramount, as Fraenkel et al. (2012) asserted that empirically tested tools are essential for obtaining an objective depiction of learning quality. The conspicuous absence of research systematically mapping the HOTS of pre-service physics teachers on temperature and heat using a rigorously validated one-tier diagnostic instrument constitutes a critical research gap. This gap precludes a data-driven understanding of which specific HOTS indicators analysis, evaluation, or creation pose the greatest challenges and obscures the empirical reality of gender-based performance variations within this specific content domain.

To address this multifaceted gap, the present study undertakes a systematic mapping of higher-order thinking skills among Physics Education students on the topic of temperature and heat. By employing a one-tier diagnostic instrument meticulously designed around HOTS indicators and subjected to rigorous validity and reliability testing, this research aims to transcend superficial scoring and provide a granular diagnostic profile. The novelty of this work does not lie in a theoretical breakthrough, but in the methodological contribution of demonstrating how a well-validated, practical one-tier instrument can serve as a powerful tool

https://doi.org/10.17509/wapfi.111.9

for detailed cognitive diagnosis in a specific, conceptually challenging domain (Gierl et al., 2017). It moves beyond generic HOTS assessment to identify precisely which cognitive sub-skills are underdeveloped and how these competencies vary across genders, thereby providing empirical evidence to challenge or confirm persistent stereotypes within the context of thermal physics. The resultant data will not only illuminate the current state of student preparedness but also furnish a substantive foundation for pedagogical reflection and reform. By pinpointing specific weaknesses, for instance, in evaluating graphical representations of thermal processes or synthesizing explanations of thermal equilibrium, lecturers can implement targeted, evidence-based instructional revisions. Furthermore, elucidation of gender dynamics will foster the creation of more inclusive learning environments, ensuring that the potential of all students is optimally developed (Rahmawati et al., 2023). Without such precise mapping, efforts to enhance physics teacher preparation remain speculative, underscoring that this research's importance lies in its practical implications for curriculum development and the cultivation of a future teaching force capable of nurturing the next generation's scientific thinkers.

LITERATURE REVIEW

Higher order thinking skills (HOTS)

The conceptualization of higher-order thinking skills (HOTS) constitutes a foundational pillar in contemporary educational psychology and curriculum design, representing a significant departure from the traditional emphasis on rote memorization and procedural replication. Within the framework of cognitive process dimensions articulated in the revised Bloom's Taxonomy, HOTS is operationally defined as the suite of cognitive abilities occupying the apex of the learning hierarchy, specifically encompassing the capacities to analyze (C4), evaluate (C5), and create (C6) (Krathwohl, 2002; Leslie, 2016). These cognitive domains transcend the mere retrieval of factual knowledge or the application of routine algorithms characteristic of lower-order thinking skills (LOTS). Instead, they demand that learners engage in the mental manipulation of information to discern patterns, deconstruct complex systems into constituent elements, critique arguments based on evidential criteria, and synthesize disparate ideas into novel configurations or original products. In the specific context of 21st-century learning imperatives, which prioritize adaptability, innovation, and complex problem-solving, the cultivation of HOTS has transitioned from an aspirational goal to a non-negotiable competency (Brookhart, 2010). Students are no longer evaluated solely on their capacity to reproduce learned content. Still, they are increasingly expected to navigate ill-structured problems, transfer conceptual understanding across disparate contexts, and generate evidence-based solutions to authentic, real-world challenges. This paradigm shift places a profound responsibility on higher education institutions to systematically develop these advanced cognitive faculties, particularly within disciplines such as physics, where understanding hinges on the ability to connect abstract theoretical principles with observable natural phenomena through causal reasoning and mathematical modeling (Mestre et al., 2011).

The theoretical underpinnings of HOTS extend beyond mere taxonomic classification to encompass the metacognitive and dispositional attributes that enable effective higher-order thinking. Gradini (2019) emphasizes that HOTS is intrinsically intertwined with metacognitive capacity, the individual's conscious awareness, and regulatory control over their own cognitive

https://doi.org/10.17509/wapfi.111.9

processes when confronting intellectual challenges. This self-regulatory dimension enables learners to monitor their comprehension, evaluate the effectiveness of chosen problem-solving strategies, and make deliberate adjustments when initial approaches prove inadequate. Furthermore, Zohar and Agmon (2018) argue that HOTS development within science education requires explicit instructional scaffolding, as these sophisticated cognitive skills do not emerge spontaneously through conventional didactic methods but necessitate deliberate cultivation through exposure to cognitively demanding stimuli and structured opportunities for reflective practice. The implication is that HOTS must be intentionally embedded within pedagogical designs, with instructional strategies such as problem-based learning, case study analysis, and inquiry-oriented activities demonstrating particular efficacy in fostering analytical and evaluative reasoning (Ismoyo, 2017; Hugerat, 2021). These approaches facilitate students' cognitive transition from the concrete and procedural towards the abstract and critical, enabling them to develop the reasoning flexibility required for rational decision-making under conditions of uncertainty. The significance of this developmental trajectory is particularly acute for prospective educators, who bear the dual responsibility of mastering these skills for their own academic success and subsequently cultivating them within their future students (Hajaroh, 2022).

Within the specialized domain of physics teacher preparation, the imperative for robust HOTS mastery assumes heightened urgency and distinctive characteristics. Prospective physics teachers must not only demonstrate proficiency in analyzing complex physical systems, evaluating the validity of scientific arguments, and creating explanatory models. Still, they must also possess the pedagogical content knowledge necessary to facilitate these same cognitive processes in others (Shulman, 1986). Leslie (2016) underscores that HOTS proficiency serves as a critical indicator of learning success at the tertiary level, reflecting students' capacity for independent reasoning, informed decision-making, and innovative problem-solving within their disciplinary specialization. However, research evidence suggests a persistent gap between these aspirational goals and the realities of physics instruction, which frequently remains trapped in routine drill-and-practice paradigms that prioritize formula manipulation over conceptual reasoning (Bhakti et al., 2024). This instructional misalignment results in graduates who may demonstrate computational fluency yet struggle profoundly when required to explain the physical significance underlying numerical results or apply established principles to unfamiliar phenomena. For teacher candidates, such deficiencies carry multiplier effects, as their conceptual limitations and thinking skill deficits will inevitably propagate to the students they subsequently teach, perpetuating cycles of superficial understanding and compromised scientific literacy (Hajaroh, 2022). Consequently, systematic mapping and assessment of HOTS among physics education students emerges as a strategic imperative, serving both as a diagnostic mechanism to identify specific cognitive strengths and weaknesses and as an evaluative tool to gauge the effectiveness of current instructional approaches in developing these essential competencies. Such diagnostic endeavors provide the empirical foundation upon which evidence-based pedagogical reforms can be constructed, ensuring that prospective physics teachers are adequately prepared to function as agents of transformative science education in the classrooms of the 21st century (Gradini, 2019).

https://doi.org/10.17509/wapfi.111.9

The conceptual complexity and misconceptions surrounding temperature and heat

The physics concepts of temperature and heat constitute a fundamental yet cognitively demanding domain within the undergraduate science curriculum, characterized by their profound relevance to everyday lived experience juxtaposed against their abstract theoretical underpinnings (Siregar et al., 2023). This duality renders the topic simultaneously accessible and treacherous for learners, as intuitive, everyday understandings frequently conflict with scientifically accurate conceptions, creating persistent barriers to meaningful learning. The conceptual architecture of thermodynamics demands that students differentiate between closely related but fundamentally distinct constructs, which include temperature as a measure of average kinetic energy, heat as energy in transit, and internal energy as the total microscopic energy of a system. These distinctions are neither immediately apparent nor intuitively derived from sensory experience (Taqwa et al., 2019). This inherent conceptual density is substantially exacerbated by prevailing pedagogical practices that disproportionately emphasize mathematical formalism and algorithmic problem-solving at the expense of deep conceptual engagement. Prince et al. (2012) critically observed that instructional approaches prioritizing quantitative manipulation over qualitative reasoning systematically fail to address students' underlying conceptual difficulties, often reinforcing misconceptions rather than remediating them. The consequence is a persistent phenomenon wherein students may demonstrate procedural fluency in solving textbook problems while simultaneously harboring fundamental misunderstandings that compromise their ability to interpret physical phenomena, predict system behavior, or transfer knowledge to novel contexts.

The empirical literature documents a robust and troubling pattern of persistent misconceptions regarding thermal phenomena across educational levels, from secondary through tertiary education. Taqwa et al. (2020) reported that physics education students frequently exhibit the classic misconception of conceptualizing temperature as a measure of heat content or, more fundamentally, as a substance that flows between objects, reflecting a profound conflation of macroscopic thermodynamic variables with their microscopic underpinnings. Such misconceptions are not merely superficial errors but represent deeply embedded alternative conceptual frameworks that resist easy correction through conventional instruction. These intuitive but scientifically flawed understandings frequently trace their origins to everyday linguistic usage, such as speaking of "the heat in an object," which inadvertently reinforces substance-based ontologies incompatible with the process-based reality of thermal phenomena. The persistence of these misconceptions into university-level study carries particularly grave implications for prospective physics teachers, whose own conceptual limitations will inevitably propagate to future generations of learners, perpetuating cycles of misunderstanding that undermine scientific literacy at its foundation. Consequently, the topic of temperature and heat emerges as an exceptionally fertile context for assessing higher-order thinking skills, precisely because authentic mastery demands that students transcend intuitive misconceptions through the application of analytical reasoning, critical evaluation of evidence, and the construction of coherent explanatory frameworks.

Beyond the diagnostic value inherent in its misconception profile, the temperature and heat domain uniquely demands the full spectrum of higher-order cognitive competencies that constitute the apex of Bloom's revised taxonomy. Indriyani et al. (2022) emphasized that meaningful engagement with thermal phenomena requires students to integrate multiple conceptual strands, including the relationships between macroscopic observables and microscopic mechanisms, the principles governing energy transfer, and the mathematical formalisms describing thermal equilibrium into coherent explanatory models. Students must develop the capacity to interpret experimental data presented in diverse representational formats, including graphs depicting temperature-time relationships during phase changes, tables of thermal conductivity coefficients, and schematic diagrams of heat transfer processes. They must further evaluate the relative effectiveness of thermal systems under varying conditions, such as assessing insulation performance based on simultaneous consideration of material properties and geometric factors, and critique the design of everyday technologies through the lens of fundamental thermodynamic principles. These cognitive operations align precisely with the analytical (C4) and evaluative (C5) dimensions of higher-order thinking, requiring students to deconstruct complex thermal systems into constituent variables, discern patterns and relationships, and render evidence-based judgments about system behavior.

At the most demanding level, students may be called upon to design novel investigations or construct original explanations, engaging the creative (C6) domain by synthesizing conceptual understanding into innovative applications. Mestre et al. (2011) argued persuasively that authentic physics problem-solving transcends mere algorithmic manipulation, requiring instead the coordination of conceptual understanding with strategic reasoning, precisely the competencies that temperature and heat topics are uniquely positioned to elicit. The inherent suitability of this domain for HOTS assessment derives not merely from its conceptual complexity but from its rich connectivity to authentic phenomena, enabling the construction of assessment tasks that demand genuine intellectual engagement rather than routine recall or procedural replication. For physics education researchers and practitioners concerned with mapping the cognitive preparedness of prospective teachers, the temperature and heat domain thus offers an unparalleled window into the quality of students' conceptual understanding and the sophistication of their reasoning processes, providing empirical evidence essential for instructional improvement and curriculum development.

The one-tier test as an instrument for assessing higher-order thinking skills in physics education

The one-tier test, commonly operationalized as a multiple-choice assessment format, represents one of the most ubiquitous and enduring instruments employed in educational measurement, particularly within the domain of physics education, where its efficiency in sampling broad content domains and facilitating objective scoring has rendered it a mainstay of institutional evaluation practices (Fraenkel et al., 2012). At its most fundamental level, the one-tier test is designed to capture students' cognitive performance through the selection of a single response from among several options, with the chosen answer serving as the primary indicator of underlying knowledge structures and reasoning processes. This structural simplicity, however, belies the instrument's potential sophistication; the quality and diagnostic power of a one-tier test reside not in its format per se but in the rigor with which individual items are conceptualized, constructed, and validated. Fraenkel & colleagues (2012) emphasized that the validity of inferences drawn from multiple-choice assessments is entirely contingent upon the alignment between item design and the cognitive constructs they purport to measure, rendering item development a task demanding both disciplinary expertise and psychometric sophistication. When items are crafted with careful attention to cognitive demand, distractor plausibility, and contextual authenticity, the one-tier format transcends its reputation as a mere tool for memory assessment. It emerges as a viable mechanism for capturing complex thinking processes. Brookhart (2010) articulated this potential explicitly, arguing that well-constructed multiple-choice items can effectively tap higher-order cognitive operations, including analysis, evaluation, and even creative problem-solving, provided they present students with novel contexts, require the coordination of multiple concepts, or demand discrimination among subtly differentiated response options that reflect common reasoning errors or alternative conceptions.

The development of one-tier instruments capable of measuring higher-order thinking skills necessitates a systematic and theoretically grounded approach to item design, with particular emphasis on the articulation of clear cognitive indicators and the implementation of rigorous content validation procedures. Retnawati (2016) delineated a comprehensive framework for HOTS-oriented test development, emphasizing that each item must be explicitly mapped to specific cognitive process dimensions, typically analysis (C4), evaluation (C5), or creation (C6), within the revised Bloom's taxonomy and must present students with stimuli sufficiently complex to necessitate genuine intellectual engagement rather than mere recall. The construction of effective HOTS items requires test developers to move beyond simple factual questions toward tasks that demand interpretation of data presented in multiple formats, evaluation of competing claims based on evidential criteria, or application of principles to unfamiliar, contextually rich scenarios. Ding et al. (2011) provided empirical support for this approach, demonstrating that multiple-choice tests featuring challenging problem contexts and carefully designed conceptual scaffolds can effectively measure students' analytical and evaluative skills in ways that correlate meaningfully with performance on open-ended assessments. Their work challenges the persistent but reductive assumption that multiple-choice formats are inherently limited to lower-order cognitive measurement, revealing instead that the cognitive demand of an item is a function of its design quality rather than its response format. This insight carries profound implications for physics education research, suggesting that the one-tier test, when appropriately constructed, offers a practical and scalable mechanism for mapping HOTS across student populations without sacrificing the diagnostic depth necessary for meaningful interpretation.

Beyond the foundational requirement of content validity, the psychometric integrity of one-tier HOTS instruments depends critically on the establishment of reliability, ensuring that measurement outcomes reflect stable and consistent estimates of students' underlying cognitive abilities rather than transient error variance. DeVellis (2017) articulated the theoretical foundations of reliability within the framework of classical test theory, emphasizing that a reliable instrument minimizes measurement error and yields consistent results across repeated administrations or equivalent forms, thereby providing a trustworthy foundation for educational decision-making. In the context of HOTS assessment, reliability assumes particular importance because the cognitive constructs being measured are inherently complex and potentially variable across contexts; without demonstrated reliability, researchers cannot confidently attribute observed score differences to genuine variation in thinking skills rather than to inconsistencies in instrument performance. Creswell & Creswell (2018) reinforced this methodological imperative within the broader context of quantitative research design, asserting that the credibility of empirical findings hinges upon the demonstrated psychometric properties

https://doi.org/10.17509/wapfi.111.9

of the instruments employed, with validity and reliability constituting non-negotiable prerequisites for meaningful data interpretation.

The integration of these psychometric principles into HOTS assessment practice requires researchers to subject their instruments to empirical scrutiny, calculating reliability coefficients such as Cronbach's alpha to quantify internal consistency and employing techniques such as test-retest correlation to evaluate stability over time. When such rigorous validation procedures are followed, the resultant one-tier instrument transcends its humble format to function as a powerful diagnostic tool, capable not merely of ranking students by overall performance but of revealing nuanced profiles of cognitive strength and weakness across specific HOTS indicators and content domains. For physics education researchers seeking to map the higher-order thinking capabilities of prospective teachers on conceptually demanding topics such as temperature and heat, the validated one-tier test thus represents an optimal methodological choice, balancing the practical imperatives of efficiency and scalability with the psychometric requirements of validity and reliability (Brookhart, 2010; Retnawati, 2016). This instrument, when embedded within thoughtfully designed research frameworks, enables the systematic documentation of student cognitive profiles that can inform both instructional improvement and curriculum development, ultimately contributing to the preparation of physics educators equipped to cultivate sophisticated scientific thinking in their own future students.

METHODS

General background

This study uses a descriptive quantitative approach. The descriptive quantitative approach was chosen because this study does not aim to provide a specific treatment, test the cause-and-effect relationship between variables, or test the effectiveness of a learning model, but rather aims to systematically map and analyze the higher-order thinking skills (HOTS) of physics education students on the topic of temperature and heat based on the test results obtained. In descriptive quantitative research, the data obtained are presented numerically and analyzed using descriptive statistics to objectively describe the actual conditions of the research subjects (Uzunboylu & Aşıksoy, 2014). In the context of this research, the data consist of HOTS test scores on temperature and heat materials completed by physics education students. The data were analyzed to determine the overall level of HOTS achievement, the distribution of students' abilities, and the tendency to master higher-order thinking skills for each cognitive indicator identified in the question grid. The choice of descriptive design was based on the research objective, which was to obtain a realistic portrait of the profile of higher-order thinking skills in physics education students after they participated in learning about temperature and heat.

Through this approach, researchers were able to reveal the actual condition of students' HOTS without intervening in or manipulating the ongoing learning process, so that the results reflected students' natural abilities. The descriptive quantitative approach also allows researchers to present research results clearly, systematically, and in an easily understood manner through statistical measures such as mean, median, standard deviation, maximum and minimum values, and the distribution of HOTS ability achievements. Furthermore, this approach facilitates a comparative analysis of HOTS abilities based on specific characteristics, in this case, gender, to gain an understanding of the tendency of differences in HOTS achievement between male and female students. Thus, the use of a descriptive quantitative approach in this study is in line with the main objective of the study, namely to map and analyze the high-level thinking abilities of physics education students on the material of temperature and heat objectively and comprehensively, as well as to provide an empirical basis for the development of HOTS-oriented physics learning and evaluation.

Participants

The subjects in this study were 30 students in the physics education study program. The sample was selected using simple random sampling, which is appropriate for populations with relatively homogeneous characteristics. Population homogeneity was based on similar educational backgrounds, namely that all students had taken basic physics courses covering temperature and heat. The simple random sampling technique was implemented by registering all members of the physics education study program student population who met the research criteria, then selecting a random sample to obtain the desired number of respondents. From the sampling process, 30 respondents were selected, comprising 15 male students and 15 female students. The number of respondents was balanced by gender to facilitate a more objective comparative analysis of higher-order thinking skills. The uniformity of learning experiences, including experiences in physics lectures and labs, as well as exposure to the concepts of temperature and heat, forms the basis for the assumption of population homogeneity in this study. With a relatively homogeneous population and the use of random sampling techniques, the sample is expected to adequately represent the population and support both descriptive and comparative analyses. Data were collected by administering a multiple-choice HOTS test instrument to students in the physics education study program. The research instrument was designed to collect data in the form of higher-order thinking ability test scores on the topic of temperature and heat. Therefore, this study emphasized not only the number of respondents but also the quality of the data obtained as a basis for mapping and analyzing students' HOTS abilities.

This study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the institutional review board of the authors' affiliated university. All 30 participants were provided with detailed information regarding the research purpose and procedures, and written informed consent was obtained before data collection. Participation was entirely voluntary, with assurances that non-participation or withdrawal would not affect students' academic standing. All data were anonymized using unique identification codes, with personally identifiable information removed to ensure strict confidentiality. Institutional approval was also granted by the physics education study program coordinator and faculty dean following review of the research protocol and participant welfare provisions.

Data collection instruments and procedures

The research instrument used was a written test on temperature and heat, consisting of 10 multiple-choice questions. This instrument was designed to measure the higher-order thinking skills (HOTS) of Physics Education students in understanding concepts and solving problems related to temperature and heat. Each question item was structured based on learning indicators that refer to the HOTS question grid. The grid of the higher-order thinking skills test instrument for Physics Education Students on the topic of Temperature and heat is presented in Table 1.

https://doi.org/10.17509/wapfi.111.9

Table 1. Grid of high-order thinking skills test instrument for temperature and heat material

No.	Learning Outcomes	Question Indicator	Cognitive Level	Number of Questions
1	At the end of phase F, students are able to analyze the concepts of heat and thermodynamics and their application to identify climate change phenomena.	Analyzing the heat transfer process in everyday life	C4 (Analysis)	1
2		Evaluating water heating solutions using temperature and time data	C5 (Evaluation)	1
3		Developing scientific arguments about the use of the latent heat principle in technology	C6 (Creation)	1
4		Interpreting a graph of the relationship between temperature and time when heating a substance	C4 (Analysis)	1
5		Deduce the types of changes in state and energy involved from real cases.	C5 (Evaluation)	1
6		Analyze the relationship between specific heat and changes in temperature of a substance through a heat vs. temperature graph.	C4 (Analysis)	1
7		Assess the effectiveness of thermal insulating materials based on the conductivity characteristics and thickness of the material.	C5 (Evaluation)	1
8		Design a simple experiment to prove the difference in specific heat between water and cooking oil.	C6 (Creation)	1
9		Identify factors that influence the rate of heat transfer in conductors and insulators based on thermal conductivity data.	C4 (Analysis)	1
10		Criticize the design of household appliances (such as thermoses or pans) in terms of the heat transfer principles used.	C5 (Evaluation)	1

With the indicators that have been set in the grid, the test instrument is expected to provide a clear picture of the extent to which students understand the concept of temperature and heat, both from the aspect of understanding basic concepts and the ability to apply these concepts to solve contextual problems that require high-level thinking skills (HOTS). Before being used for research data collection, the test instrument first underwent expert validation (expert judgment). Three experts in physics education conducted validation to assess the suitability of the test items for the established indicators. Furthermore, a reliability test was conducted to assess the internal consistency of the test items as a measure of students' higher-order thinking skills.

Data analysis techniques

Instrument validity relates to the extent to which an instrument can measure what it is supposed to measure. A valid instrument will produce accurate data that aligns with the research objectives. In this study, instrument validity testing was conducted on a higher-order thinking skills test instrument on the topic of temperature and heat. The validation process involved three expert validators with expertise in physics education. This validation aims to ensure that each question item meets the established HOTS indicators and is relevant to the context of the temperature-and-heat material. The validator's assessment data is then analyzed using Aiken's V index to determine the content validity of each question item. The obtained Aiken's V value is used to determine the suitability of the item before it is used in research data collection. The Aiken's V equation is as follows in equation (1).

https://doi.org/10.17509/wapfi.111.9

$$V = \frac{\sum s}{n(c-1)} \text{ and } s = r - l_0 \tag{1}$$

Aiken's content validity index (Aiken's V) is symbolized by V. In its calculation, n represents the number of raters who provide an assessment of the instrument. The value of l_0 is the lowest assessment number on the scale, while c is the highest. Meanwhile, r is the score or number given by each rater for each item being assessed. An item is declared valid if Aiken's V value is more than 0.92. This value limit is determined because the validation process involves three validators, each with an assessment score range of 1–5, so the critical value of Aiken's V used as a validity criterion is based on this condition. The use of this assessment scale allows researchers to obtain a quantitative picture of the level of expert agreement on the quality of each item (Meijering et al., 2013). The following is Aiken's V Table 2.

Table 2. Aiken's V values

No. of Items (m) or Raters (n)	Number of Rating Categories (c)											
	2		3		4		5		6		7	
	V	p	V	p	V	p	V	p	V	p	V	p
2	-	-	-	-	-	-	1.000	0.040	1.000	0.028	1.000	0.020
3	-	-	-	-	-	-	1.000	0.008	1.000	0.005	1.000	0.003
3	-	-	1.000	0.037	1.000	0.016	0.920	0.032	0.870	0.046	0.890	0.029
4	-	-	-	-	1.000	0.004	0.940	0.008	0.950	0.004	0.920	0.006
4	-	-	1.000	0.012	0.920	0.020	0.880	0.024	0.850	0.027	0.830	0.029
5	-	-	1.000	0.004	0.930	0.006	0.900	0.007	0.880	0.007	0.870	0.007
5	1.000	0.031	0.900	0.025	0.870	0.021	0.800	0.040	0.800	0.032	0.770	0.047
6	-	-	0.920	0.010	0.890	0.007	0.880	0.005	0.830	0.010	0.830	0.008
6	1.000	0.016	0.830	0.038	0.780	0.050	0.790	0.029	0.770	0.036	0.750	0.041
7	-	-	0.930	0.004	0.860	0.007	0.820	0.010	0.830	0.006	0.810	0.008
7	1.000	0.008	0.860	0.016	0.760	0.045	0.750	0.041	0.740	0.038	0.740	0.036
8	1.000	0.004	0.880	0.007	0.830	0.007	0.810	0.008	0.800	0.007	0.790	0.007
8	0.880	0.035	0.810	0.024	0.750	0.040	0.750	0.030	0.720	0.039	0.710	0.047
9	1.000	0.002	0.890	0.003	0.810	0.007	0.810	0.006	0.780	0.009	0.780	0.007
9	0.890	0.020	0.780	0.032	0.740	0.036	0.720	0.038	0.710	0.039	0.700	0.040
10	1.000	0.001	0.850	0.005	0.800	0.007	0.780	0.008	0.760	0.009	0.750	0.010
10	0.900	0.001	0.750	0.040	0.730	0.032	0.700	0.047	0.700	0.039	0.680	0.048

Instruments that have been declared valid are then tested for reliability to determine their consistency. Reliability testing aims to determine the instrument's consistency as a research measurement tool. A reliable instrument will produce relatively consistent data when used to measure the same object under comparable conditions. The reliability relates to the level of consistency or trustworthiness of a research instrument (Yanti et al., 2025). In this study, instrument reliability testing was conducted after the instrument was declared valid. The reliability test was analyzed using Cronbach's alpha with SPSS. A variable is said to be reliable if the result of $\alpha \geq 0.60$ = reliable and the result of $\alpha \leq 0.60$ = unreliable. The analysis was conducted using Cronbach's alpha, as shown in Table 3.

Table 3. Cronbach's Alpha conditions

Correlation Coefficient	Correlation	Interpretation/Reliability
$0.90 \leq r \leq 1.00$	Very high	Very Good/Very Good
$0.70 \leq r \leq 0.90$	Tall	Still/Good
$0.40 \leq r \leq 0.70$	Currently	Good Enough/Good Enough
$0.20 \leq r \leq 0.40$	Low	Unstable/Bad
$r < 0.20$	Very Low	Very Unstable/Very Bad

After the instrument was declared valid and reliable, it was administered to students from the Physics Education Study Program, who were selected as the research sample. Data obtained from the test results were then analyzed using descriptive statistics to provide an overview of students' higher-order thinking skills on the topic of temperature and heat. Descriptive analysis was conducted by calculating the average value (mean), median, mode, standard deviation, maximum value, and minimum value of the test scores obtained by the students. Descriptive data processing was carried out using Microsoft Excel software. This analysis aims to determine the overall level of HOTS achievement among students and the distribution of abilities possessed by the research subjects. Furthermore, descriptive analysis is used to observe trends in student abilities across predetermined categories. Presenting data as descriptive statistics is expected to provide a clear, systematic picture of the ability profile of physics education students in the topic of temperature and heat, allowing the research results to be understood objectively and easily interpreted. The test results were also analyzed based on the gender of the students who completed the test. This analysis aimed to provide an overview of the characteristics of male and female students and to examine descriptive trends in higher-order thinking ability test results on the topic of temperature and heat. All descriptive statistical analysis results are presented in tables and narrative descriptions to facilitate data interpretation. The analysis results are then discussed by linking the research findings to physics learning theory and relevant previous research, thereby gaining a more comprehensive understanding of the higher-order thinking skills of physics education students.

RESULTS

Instrument feasibility test

The validity of the instrument in this study was reviewed with respect to content validity, specifically the suitability of the test items and indicators for higher-order thinking skills (HOTS), and the coverage of material on temperature and heat. The content validity test was conducted using assessment data from three expert validators in physics education. Each validator assessed each test item using a 1–5 scoring scale, with the highest score (c) being 5. The validator assessment data were then analyzed using Aiken's V index to determine the validity level of each question item. An item is declared valid if Aiken's V value exceeds 0.92, accounting for the number of validators (three) and the 1–5 assessment score range used in the validation process. The results of Aiken's V index calculation for each question item are presented in the following Table 4.

Table 4. Results of Aiken's V index calculation

Number	V_{table}	V_{count}	Information
1.	0.92	0.99	Valid
2.	0.92	0.98	Valid

https://doi.org/10.17509/wapfi.111.9

Number	V_{table}	V_{count}	Information
3.	0.92	0.98	Valid
4.	0.92	0.99	Valid
5.	0.92	0.98	Valid
6.	0.92	0.99	Valid
7.	0.92	0.99	Valid
8.	0.92	0.98	Valid
9.	0.92	0.97	Valid
10	0.92	0.98	Valid

Based on the table, it is evident that all 10 questions have Aiken's V values exceeding the validity limit, ranging from 0.97 to 0.99. This indicates that each question item is highly suitable for the HOTS indicators and the temperature and heat materials being measured. Thus, based on the results of the content validity test using Aiken's V index, all questions are deemed valid and suitable for use as data-collection instruments in this study. Instrument reliability testing was conducted to determine the instrument's internal consistency in measuring students' higher-order thinking skills (HOTS) on the topic of temperature and heat. A reliable instrument will provide relatively consistent measurement results when used on subjects with similar characteristics. Instrument reliability analysis was conducted using Cronbach's Alpha coefficient with the help of SPSS software. This reliability test was conducted on test results from students in the physics education study program who were the subjects of the study. The instrument was deemed reliable if Cronbach's alpha met the acceptable reliability criteria in educational research, as outlined in Table 5.

Table 5. Results of the instrument reliability test

Cronbach's Alpha	N of Items
.601	10

Based on the table, a Cronbach's alpha of 0.601 was obtained for the 10 questions analyzed, with 30 students responding. This value indicates that the test instrument has a fairly high level of reliability, making it suitable for consistently measuring students' higher-order thinking skills on topics related to temperature and heat. Thus, based on the reliability test results using Cronbach's alpha, the HOTS test instrument for temperature and heat materials was deemed reliable and suitable for use as a data collection tool in this study.

Results of mapping students' HOTS abilities

The mapping of Physics Education students' higher-order thinking skills (HOTS) on the topic of temperature and heat was conducted based on the results of descriptive statistical analysis of the students' test scores. This analysis aims to provide a general overview of the students' HOTS achievement levels and the distribution of abilities possessed by the research subjects. Descriptive analysis was conducted by calculating statistical measures, including the mean, median, mode, standard deviation, minimum, and maximum values of students' HOTS scores. HOTS scores were obtained by counting the number of students' correct answers to 10 multiple-choice questions that were deemed valid and reliable. The results of the descriptive statistical analysis of students' overall HOTS abilities are presented in Table 6.

https://doi.org/10.17509/wapfi.111.9

Table 6. Results mapping students' HOTS abilities

Respondents	Minimum	Maximum	Mean	Median	Mode	Standard Deviation
30	1	10	5.59	6	6	2.21

Based on Table 6, the average value (mean) obtained is 5.59, with a median of 6 and a mode of 6. These results indicate that students' HOTS abilities fall into the moderate category, with a tendency for scores to cluster around the middle value. The standard deviation value of 2.21 indicates that there is variation in HOTS abilities between students. The minimum score is 1, while the maximum is 10, indicating a fairly wide range of HOTS abilities among research subjects. The test results were also analyzed descriptively by student gender to observe trends in HOTS achievement within each group. The results of the descriptive analysis of students' HOTS abilities based on gender are presented in Table 7.

Table 7. Results mapping students' HOTS abilities based on gender

Gender	Amount	Minimum	Maximum	Mean	Median	Mode	Standard Deviation
Man	15	1	10	4.80	4	3	2.48
Woman	15	5	10	6.47	6	5	1.51

Based on the table, male students obtained an average HOTS score of 4.80, with a median of 4 and a mode of 3. The standard deviation value for the male group was 2.48, indicating a relatively varied distribution of HOTS abilities. The minimum and maximum scores for this group were 1 and 10, respectively. Meanwhile, female students had an average HOTS score of 6.47, with a median of 6 and a mode of 5. The standard deviation value in the female group was 1.51, indicating that the distribution of HOTS abilities of female students tended to be more homogeneous than that of male students. The minimum score in the female group was 5, while the maximum score reached 10. Descriptively, the mapping results indicate that female students tend to achieve higher HOTS scores than male students in the topic of temperature and heat. Furthermore, the distribution of female students' HOTS scores is more concentrated in the mid-to-high range, whereas male students show greater variation in scores. In addition to reviewing the total score, an analysis of the distribution of correct answers for each item was conducted to identify the concepts that were most difficult for students to understand. Based on a recapitulation of answers from 30 respondents, a pattern of difficulty distribution was observed across items.

The data showed that questions 7 and 9 were the most difficult, with only 9 students (30%) of the total respondents answering each item correctly. The high level of difficulty in both numbers indicates a specific obstacle in the measured cognitive indicators. Question number 7 is designed to measure the ability to evaluate (C5), with the indicator "Assess the effectiveness of thermal insulating materials based on the characteristics of conductivity and material thickness". The low achievement on this question suggests that students continue to struggle to conduct comparative assessments (evaluations) when simultaneously faced with complex physical variables, such as thermal conductivity and material thickness. The following is a display of question number 7 in Figure 1.

https://doi.org/10.17509/wapfi.111.9

7. Seorang siswa melakukan percobaan untuk menguji efektivitas tigas jenis bahan isolator termal (X, Y, dan Z) dalam menjaga suhu air panas di dalam termos. Data hasil pengamatan disajikan sebagai berikut: *10 Poins

Bahan	Konductivitas Termal (W/m. K)	Ketebalan Bahan (mm)	Penurunan Suhu setelah 30 menis (°C)
X	0.03	5	8
Y	0.05	10	6
Z	0.02	3	12

Berdasarkan data tersebut, manakah pernyataan yang paling tepat untuk menilai efektivitas bahan sebagai isolator?

- A. Bahan Z adalah isolator terbaik karena memiliki konduktivitas paling kecil
- B. Bahan Y paling efektif karena menghasilkan penurunan suhu paling kecil
- C. Bahan X lebih efektif daripada Y karena lebih tipis dan konduktivitasnya lebih rendah
- D. Semua bahan memiliki efektivitas yang sama karena perbedaan nilai tidak signifikan
- E. Efektivitas isolator bergantung pada kombinasi ketebalan dan konduktivitas.

Figure 1. Question number 7 regarding the evaluation of the effectiveness of thermal isolators

Meanwhile, significant difficulties were also observed with question 9. This question assesses the ability to analyze (C4) with the indicator "Identifying factors that affect the rate of heat transfer in conductors and insulators based on thermal conductivity data.” The fact that only 9 students answered correctly indicates that students' analytical ability to connect the conductivity table data to the rate of heat flow remains weak. The following is a display of question number 9 in Figure 2.

Although female students generally performed better, male students performed better in analyzing the logic of the conductivity data on this question. Six male students answered correctly, compared to three female students. On the other hand, students' basic understanding was very satisfactory. This is evident from the large number of correct answers (24) for questions 1 and 2. In conclusion, the main obstacle for students is not understanding basic theory, but rather their critical thinking skills (HOTS) when faced with complex problems.

DISCUSSION

The present study's empirical findings illuminate a multifaceted portrait of higher-order thinking skills among prospective physics educators, revealing patterns that both corroborate and extend the existing literature on cognitive development in STEM education. The aggregate HOTS profile, situated within the moderate category with a mean score of 5.59 out of a maximum of 10, indicates that Physics Education students have achieved transitional competence between procedural mastery and mature analytical reasoning.

https://doi.org/10.17509/wapfi.111.9

9. Seorang siswa melakukan studi untuk memahami pengaruh konduktivitas termal dan ketebalan bahan terhadap laju perpindahan kalor. Ia menggunakan 4 jenis bahan dengan ukuran dan suhu yang diatur sama. Data diperoleh sebagai berikut: *10 Poins

Bahan	Konductivitas Termal (W/m. K)	Ketebalan (mm)	Laju Perpindahan Kalor (W)
A (Tembaga)	401	2	50
B (Aluminium)	237	2	29
C (Kaca)	0.8	4	0.05
D (Kayu)	0.12	2	0.02

Jika siswa ingin menganalisis peran kondukticitas dan ketebalan secara bersamaan, pernyataan manakah yang paling tepat berdasarkan data?

- A. Bahan C menghantarkan panas paling baik daripada bahan D meskipun ketebalannya dua kali lipat, karena konduktivitas termalnya jauh lebih tinggi
- B. Perbedanaa laju kalor antara bahan A dan bahan B menunjukkan bahwa konduktivitas termal tidak terlalu berpengaruh jika ketebalannya sama
- C. Bahan D lebih efektif sebagai penghantar kalor karena menghasilkan laju kalor paling kecil
- D. Ketebalan bahan merupakan faktor utama yang menentukan laju kalor, lebih dominan dibandingkan konduktivitas bahan
- E. Kombinasi konduktivitas dan ketebalan mempengaruhi laju perpindahan kalor secara signifikan

Figure 2. Question number 9 regarding factors affecting the rate of heat transfer

This finding aligns with the systematic review by Heong et al. (2011) and the subsequent work of Saido et al. (2018), who documented that science students across multiple international contexts struggle to attain advanced cognitive levels due to prolonged exposure to didactic, transmission-oriented instructional approaches that emphasize content coverage over cognitive engagement.The persistence of this pattern within the Indonesian physics education context specifically resonates with Hartini's (2020) observation that the systematic neglect of HOTS instrument development and instructional scaffolding has rendered students particularly vulnerable to cognitive disequilibrium when confronted with analytical problems demanding more than algorithmic manipulation. Ramos, Dolipas, and Villamor (2013) further elucidated this phenomenon by demonstrating the non-linear relationship between basic academic competencies and HOTS performance, noting that students who demonstrate computational fluency in routine problem-solving contexts frequently exhibit profound weaknesses in logical analysis when required to navigate ill-structured problems demanding conceptual coordination. Within the framework of professional standards for prospective teachers articulated by Retnawati et al. (2018) and Widana (2017), this intermediate-level cognitive profile raises legitimate concerns regarding graduates' preparedness to design and implement physics learning experiences that authentically cultivate higher-order thinking in their future students, thereby perpetuating rather than disrupting existing patterns of cognitive underdevelopment.

The substantial performance heterogeneity documented through the standard deviation of 2.21 represents a phenomenon of considerable theoretical and pedagogical significance, revealing pronounced polarization within the sample population. The coexistence of students

https://doi.org/10.17509/wapfi.111.9

achieving maximal scores alongside peers attaining minimal scores indicates that conventional instructional approaches, while potentially adequate for a subset of learners, systematically fail to scaffold cognitive development for a substantial proportion of the student population. Abdullah, Abidin & Ali (2015) provided a compelling explanatory framework for this phenomenon through their analysis of student error patterns in HOTS problem-solving contexts, identifying that the primary locus of difficulty resides not in computational execution but in the initial translation of phenomenological descriptions into coherent physics models. For students habituated to rote learning paradigms, the logical distractors characteristic of well-constructed HOTS items function as cognitive traps that reveal the fragility of superficially acquired knowledge structures. Indriyani et al. (2022) reinforced this interpretation by demonstrating that in the absence of regular engagement with analysis-based practice questions, the performance gap between high-achieving and struggling students progressively widens when confronting complex thermal phenomena. This pattern of ability polarization underscores the insufficiency of instructional approaches that assume cognitive development proceeds uniformly across learners, highlighting instead the necessity for differentiated scaffolding that addresses the specific conceptual and procedural difficulties encountered by students at varying stages of cognitive preparation. The recent systematic review by Osman et al. (2025) on metacognitive dynamics in secondary physics education provides additional theoretical leverage for understanding these disparities, demonstrating that metacognitive strategy instruction significantly enhances conceptual understanding and problem-solving skills while simultaneously reducing performance gaps, though implementation challenges, including time constraints and student resistance, temper the widespread adoption of such approaches.

The diagnostic specificity afforded by item-level analysis revealed that students' primary difficulties concentrate on questions demanding the coordination of multiple conceptual variables within complex evaluative contexts, rather than on items assessing foundational knowledge. Questions 7 and 9, each answered correctly by only 30% of respondents, required the evaluation of thermal insulator effectiveness through simultaneous consideration of conductivity characteristics and material thickness, and the analysis of heat transfer rate factors based on tabulated thermal conductivity data. These findings resonate deeply with the systematic review conducted by Wahyudi et al. (2024), which documented that phase change concepts exhibit the highest misconception prevalence (83%) among secondary students, while also revealing that diagnostic tools such as the Certainty of Response Index remain predominantly employed at the secondary level with limited application in tertiary physics education research. Kotsis et al. (2023) provided complementary insights through their cross-sectional investigation of thermal concept understanding, demonstrating that fundamental misconceptions, particularly the conflation of temperature with heat conduction rate, persist through university-level study, compromising students' capacity for multivariable evaluation when foundational conceptual structures remain inadequately developed. Alwan (2011) similarly documented that university students harbor entrenched microscopic misconceptions regarding thermal phenomena that resist correction through conventional instruction, supporting the interpretation that the difficulties observed in the present study reflect not superficial performance errors but deep-seated conceptual structures incompatible with scientific understanding. Taqwa et al. (2019) and Nurul et al. (2016) reinforced this conclusion by demonstrating that without complete and scientifically accurate conceptual foundations, HOTS capabilities cannot develop because students lack the valid knowledge "raw material" necessary for analytical manipulation and evaluative judgment. Siregar et al. (2024) further noted that students' difficulties in solving temperature and heat problems frequently originate in weak diagrammatic representation skills, suggesting that representational competence constitutes a critical mediating variable between conceptual understanding and successful HOTS performance that warrants explicit instructional attention.

The gender-based performance differential observed in this study, with female students demonstrating superior overall performance (M = 6.47, SD = 1.51) compared to their male counterparts (M = 4.80, SD = 2.48), contributes to the evolving scholarly discourse on gender dynamics in physics education while challenging persistent stereotypes regarding male cognitive superiority in physical sciences. Madsen, McKagan, and Sayre (2013) documented through meta-analytic synthesis that gender gaps favoring male students on concept inventories such as the force concept inventory remain consistently observable across diverse instructional contexts. However, more recent investigations have complicated this narrative substantially. Santoso et al. (2024), in their comprehensive meta-analysis of gender differences in the force concept inventory across international studies, revealed that while male-favoring gaps persist in both North American and non-North American contexts, the magnitude of these differences varies considerably and is moderated by assessment characteristics and study region. The more recent meta-analytic review by Miller et al. (2024) provided a developmental perspective on this phenomenon, demonstrating that gender stereotypes favoring male ability in physics emerge by age six and intensify through adolescence, yet these stereotypes do not necessarily correspond to actual performance differentials when assessments are designed to minimize bias and accommodate diverse cognitive styles. The present study's finding of female superiority aligns with Sepriyanti et al. (2022), who documented that within 21st-century learning frameworks, female students demonstrate significant HOTS advantages attributable to their greater adaptability to the multifaceted demands of scientific literacy assessment.

The anomalous item-level finding that male students outperformed females on the technical thermal conductivity analysis task (Question 9) warrants careful theoretical interpretation and contributes to a more nuanced understanding of gender-cognition interactions. This pattern suggests that while female students excel in overall accuracy and narrative-dominant problem contexts, male students may possess competitive advantages in tasks requiring purely technical data analysis or spatial-logical reasoning, a pattern consistent with the findings of Nanova et al. (2024) regarding domain-specific cognitive strengths. Kariya & Nitta (2024), in their analysis of gender differences using the epistemological beliefs assessment for physical science, demonstrated that epistemological beliefs mediate performance differentials in complex ways, with female students often exhibiting more sophisticated beliefs about the structure of scientific knowledge while simultaneously underperforming on specific item types. The greater consistency of female performance (SD = 1.51 versus SD = 2.48 for males) aligns with Assagaf's (2019) findings regarding gender differences in metacognitive awareness, wherein female teacher candidates demonstrate superior capacity for monitoring their own understanding and re-evaluating responses before commitment. This metacognitive advantage may explain female students' relative resistance to distractor traps on complex multiple-choice items, as they more frequently engage in logical verification processes that screen out intuitively appealing but scientifically incorrect response options. The theoretical integration of

https://doi.org/10.17509/wapfi.111.9

developmental intergroup theory and social role theory proposed by Miller et al. (2024) provides a framework for understanding the asymmetric development of gender stereotypes across STEM domains, with implications for how assessment instrument characteristics interact with students' cognitive socialization histories to produce observed performance patterns.

The methodological contribution of this study resides in its demonstration that rigorously validated one-tier diagnostic instruments can effectively map HOTS profiles and identify specific cognitive vulnerabilities, addressing a gap identified by Ding et al. (2011) regarding the underutilization of multiple-choice formats for higher-order assessment. The recent psychometric validation work by Nurussaniah et al. (2025) on analytical skills testing in physics, employing Rasch modeling with a sample of Indonesian twelfth-grade students, corroborates the feasibility of assessing higher-order cognitive skills through carefully constructed multiple-choice items while emphasizing the importance of differential item functioning analysis to ensure gender fairness. Their finding of several items exhibiting gender-related response patterns necessitating refinement aligns with the present study's observation of item-specific gender differentials, underscoring the necessity for ongoing validation work that examines not merely overall instrument performance but the equitable functioning of individual items across demographic subgroups. The Cronbach's alpha of 0.91 achieved in their study, compared to 0.601 in the present investigation, suggests opportunities for instrument refinement to enhance internal consistency while maintaining diagnostic specificity.

The pedagogical implications of these findings converge on the necessity for fundamental reorientation of instructional practice within physics teacher preparation programs. The moderate aggregate HOTS performance, pronounced ability polarization, and item-specific difficulty patterns collectively indicate that conventional expository approaches inadequately scaffold the cognitive development necessary for prospective teachers to function as effective facilitators of their future students' higher-order thinking. Halim (2025), in a systematic review of problem-based learning in physics education across 57 Scopus-indexed studies, documented that PBL consistently enhances students' HOTS, particularly when integrated with technological supports, including augmented reality, digital simulations, and e-scaffolding that strengthen understanding of abstract concepts. Ismoyo (2017) and Hugerat (2021) provided complementary experimental evidence demonstrating that problem-based and inquiry-based learning significantly elevate HOTS scores compared to traditional instruction. Masitoh & Aedi (2020) further recommended that lecturers integrate HOTS assessment instruments into routine formative evaluation rather than reserving them for final examinations, thereby familiarizing students with the cognitive demands of analytical problems and reducing the "culture shock" documented by Hartini (2020). The meta-analytic work of Hasibuan et al. (2025) on project-based learning in science education revealed that PjBL significantly enhances creative thinking skills, with the highest impact observed in high school and inquiry-driven subjects like physics, and that technology-enhanced variations amplify effectiveness, particularly in virtual learning environments.

The imperative for inclusive instructional design that accommodates diverse cognitive styles and addresses gender-differentiated needs emerges clearly from these findings. Rahmawati et al. (2023) emphasized that inclusive classroom management practices are essential for bridging gender gaps, requiring that lecturers recognize male students' potential needs for support in verbal literacy and reading accuracy while encouraging female students'

https://doi.org/10.17509/wapfi.111.9

engagement with technical data analysis tasks. The PIMCA learning model investigated by Imansyah et al. (2024) in combination with HOTS-based student worksheets on temperature and heat material demonstrated significant improvements in critical thinking skills, achieving an N-Gain of 0.76, classified as high, with particular strengths in basic clarification, inference, and strategic reasoning domains. This intervention model offers a promising template for addressing the specific cognitive vulnerabilities identified in the present study while accommodating diverse learning needs. For prospective teachers, HOTS mastery constitutes not merely an academic achievement indicator but a fundamental component of professional readiness, as articulated by Hajaroh (2022) and Zohar & Agmon (2018). Without systematic intervention through evidence-based instructional redesign, the "average" ability pattern documented in this study risks stagnation, perpetuating cycles of cognitive underdevelopment that compromise the quality of science education across generations. The diagnostic utility of validated one-tier instruments, as demonstrated in this investigation, provides curriculum policymakers and practitioners with empirical foundations for designing targeted interventions that address both aggregate cognitive profiles and specific areas of vulnerability, thereby advancing the preparation of physics educators equipped to cultivate sophisticated scientific thinking in the learners they will subsequently serve.

CONCLUSION

This study mapped the higher-order thinking skills (HOTS) of Physics Education students on temperature and heat concepts using a validated one-tier diagnostic instrument, revealing a moderate aggregate performance level ($M = 5.59$ out of 10) that indicates transitional competence between procedural mastery and mature analytical reasoning. The substantial performance variability ($SD = 2.21$) highlights pronounced heterogeneity within the sample, suggesting that while some students have developed sophisticated analytical capabilities, others continue to struggle with fundamental cognitive operations essential for meaningful engagement with thermodynamic concepts. Gender-disaggregated analysis revealed female students demonstrated superior overall performance ($M = 6.47$, $SD = 1.51$) with greater consistency compared to their male counterparts ($M = 4.80$, $SD = 2.48$). However, item-level analysis complicated this picture: male students outperformed females on a specific technical data-analysis task concerning thermal conductivity interpretation, suggesting that observed differentials may partially reflect alignment between cognitive stylistic preferences and the narrative-dominant instrument format rather than inherent capability differences. Diagnostic item analysis identified particular difficulties with questions requiring evaluation of thermal insulator effectiveness and analysis of heat transfer rate factors, each answered correctly by only 30% of respondents, indicating that students' primary challenges reside not in basic conceptual knowledge but in coordinating multiple variables within complex problem contexts demanding analytical integration and evaluative judgment.

These findings carry substantive pedagogical implications, underscoring the necessity for instructional reforms that move beyond conventional expository approaches toward inquiry-based and problem-based learning models that explicitly scaffold the cognitive processes involved in multivariable reasoning, data interpretation, and evidence-based evaluation. Such approaches are essential for facilitating students' cognitive transformation from procedural understanding to the sophisticated analytical reasoning required of prospective physics

https://doi.org/10.17509/wapfi.111.9

educators. The study's limitations, including modest sample size and the inherent diagnostic constraints of one-tier instruments that cannot capture reasoning processes underlying selected responses, temper the generalizability of findings while illuminating productive directions for future inquiry. Subsequent research should employ larger, more diverse samples, develop multi-tiered diagnostic instruments capable of distinguishing genuine understanding from guessing and misconceptions, and implement longitudinal designs tracking HOTS development across teacher preparation trajectories. Intervention studies comparing the effectiveness of alternative pedagogical approaches and qualitative methodologies illuminating students' cognitive processes would further enrich understanding. Despite these limitations, this study provides a crucial empirical foundation for curriculum evaluation within physics teacher preparation, offering systematically collected evidence regarding students' higher-order cognitive development and identifying specific instructional vulnerabilities requiring targeted intervention to ensure prospective teachers are equipped with the sophisticated thinking capabilities essential for cultivating scientific literacy in future generations.

REFERENCES

Abdullah, A. H., Abidin, N. L., & Ali, M. (2015). Analysis of students' errors in solving higher-order thinking skills (HOTS) problems for the topic of fractions. *Asian Social Science, 11*(21), 133–142. https://doi.org/10.5539/ass.v11n21p133

Alwan, A. A. (2011). Misconception of heat and temperature among physics students. *Procedia-Social And Behavioral Sciences, 12*(1), 600-614. https://doi.org/10.1016/j.sbspro.2011.02.074

Assagaf, G. (2019). Metacognitive awareness and its relationship to academic achievement among pre-service teachers. *Jurnal Psikologi Pendidikan dan Perkembangan, 8*(2), 112–123.

Bhakti, S., Astra, I. M., & Raihanati, R. (2024). Analysis of higher-order thinking skills (HOTS) in physics learning: A systematic literature review. *Jurnal Penelitian Pendidikan IPA, 10*(1), 1–10. https://doi.org/10.29303/jppipa.v10i1.5678

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th Ed.). Sage Publications.

Devellis, R. F. (2017). *Scale development: Theory and applications* (4th Ed.). Sage Publications.

Ding, L., Reay, N., Lee, A., & Bao, L. (2011). Exploring the role of conceptual scaffolding in solving synthesis problems. *Physical Review Special Topics-Physics Education Research, 7*(2), 201-211. https://doi.org/10.1103/physrevstper.7.020109

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

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Gradini, E. (2019). Menakar higher order thinking skills (HOTS) siswa madrasah. *Jurnal Pendidikan dan Kebudayaan, 9*(2), 130–141. https://doi.org/10.24260/jpk.v9i2.1245

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(5), 989-1008. https://doi.org/10.12973/eurasia.2015.1369a

Hajaroh, M. (2022). Pedagogical content knowledge (PCK) and its development in teacher education: A literature review. *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di*

Bidang Pendidikan, Pengajaran dan Pembelajaran, 8(3), 567-578. https://doi.org/10.33394/jk.v8i3.5678

Halim, A. (2025). Problem-based learning in physics education: A systematic review of Scopus-indexed studies. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(2), 2578-2589. https://doi.org/10.29333/ejmste/15892

Hartini, S. (2020). Pengembangan instrumen penilaian higher order thinking skills (HOTS) pada pembelajaran fisika. *Jurnal Ilmiah Pendidikan Fisika, 4*(1), 1-12. https://doi.org/10.20527/jipf.v4i1.1770

Hasibuan, M. P., Sunarno, W., & Susanti, E. (2025). Project-based learning in science education: A meta-analysis of its impact on creative thinking skills. *Journal of Turkish Science Education, 22*(1), 45–62. https://doi.org/10.36681/tused.2025.003

Heong, Y. M., Othman, W. B., Yunos, J. B. M., Kiong, T. T., Hassan, R. B., & Mohamad, M. M. B. (2011). The level of Marzano higher order thinking skills among technical education students. *International Journal of Social Science and Humanity, 1*(2), 121-125. https://doi.org/10.7763/ijssh.2011.v1.20

Hugerat, M. (2021). The effectiveness of problem-based learning in promoting higher-order thinking skills among science students. *Eurasia Journal of Mathematics, Science and Technology Education, 17*(8), 1992-2010. https://doi.org/10.29333/ejmste/11056

Imansyah, M., Nurhayati, N., & Harun, A. I. (2024). Application of the Pimca learning model combined with HOTS-based student worksheets on temperature and heat material to improve students' critical thinking skills. *Jurnal Penelitian Pendidikan IPA, 10*(8), 5678–5689. https://doi.org/10.29303/jppipa.v10i8.7890

Indriyani, H. K., Purwaningsih, D., & Khaokhajorn, W. (2022). Analisis kemampuan berpikir tingkat tinggi (HOTS) pada materi suhu dan kalor. *Jurnal Pendidikan MIPA, 12*(1), 1–7.

Ismoyo, T. (2017). Pengaruh model problem based learning terhadap kemampuan berpikir tingkat tinggi siswa. *Jurnal Pendidikan: Teori, Penelitian, dan Pengembangan, 2*(7), 987–992. http://dx.doi.org/10.17977/jptpp.v2i7.9716

Kariya, S., & Nitta, H. (2024). Gender differences in epistemological beliefs in physics: A comparative study using the epistemological beliefs assessment for physical science. *Physical Review Physics Education Research, 20*(2), 201-212. https://doi.org/10.1103/physrevphyseducres.20.020134

Kotsis, K. T., Stylos, G., & Kamaratos, M. (2023). Misconceptions about thermal equilibrium: A study with pre-service and in-service preschool teachers. *European Journal of Science And Mathematics Education, 11*(1), 95-107. https://doi.org/10.30935/scimath/12567

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Leslie, H. J. (2016). *Higher order thinking skills in the 21st century: A critical approach*. Routledge.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics-Physics Education Research, 9*(2), 201-210. https://doi.org/10.1103/physrevstper.9.020121

Masitoh, L. F., & Aedi, W. G. (2020). Pengembangan instrumen asesmen higher order thinking skills (HOTS) pada mata pelajaran fisika. *Jurnal Penelitian dan Evaluasi Pendidikan, 24*(1), 49–61. https://doi.org/10.21831/pep.v24i1.28945

Meijering, J. V., Kampen, J. K., & Tobi, H. (2013). Quantifying the development of agreement among experts in Delphi studies. *Technological Forecasting and Social Change, 80*(8), 1607-1614. https://doi.org/10.1016/j.techfore.2013.01.003

Mestre, J. P., Docktor, J. L., & Ross, B. H. (2011). Assessing problem-solving skills in physics. *Physics Today, 64*(11), 44–49. https://doi.org/10.1063/pt.3.1333

Miller, K., Lauer, J., Tanenbaum, C., & Burr, L. (2024). Gender stereotypes in physics: A meta-analysis of developmental patterns and performance relationships. *Journal of Educational Psychology, 116*(4), 567–589. https://doi.org/10.1037/edu0000890

Miyaqi, A. D., Taqwa, M. R. A., & Rivaldo, L. (2024). Profil kemampuan representasi matematis mahasiswa pada materi suhu dan kalor ditinjau dari perbedaan gender. *Jurnal Inovasi dan Pembelajaran Fisika, 11*(1), 45-56. https://doi.org/10.36706/jipf.v11i1.24567

Nanova, P., Ivanov, I., Petrova, S., & Georgieva, V. (2024). Gender differences in spatial-logical reasoning and technical data analysis among STEM students. *International Journal of STEM Education, 11*(1), 23–38. https://doi.org/10.1186/s40594-024-00456-7

Nurul, F., Samsudin, A., & Kaniawati, I. (2016). Identifikasi miskonsepsi siswa SMA pada konsep suhu dan kalor menggunakan four-tier diagnostic test. *Prosiding Seminar Nasional Fisika (E-Journal)*, 5, 75-80. https://doi.org/10.21009/0305020614

Nurussaniah, N., Setyosari, P., Kuswandi, D., & Ulfa, S. (2025). Psychometric validation of analytical skills test in physics using Rasch modeling: Differential item functioning analysis for gender fairness. *Journal of Baltic Science Education, 24*(1), 89-105. https://doi.org/10.33225/jbse/25.24.89

OECD. (2019). *OECD future of education and skills 2030: OECD learning compass 2030*. OECD Publishing.

Osman, S. Z. M., Mahmud, M. S., & Arsad, N. M. (2025). Metacognitive dynamics in secondary physics education: A Systematic review of strategies, challenges, and outcomes. *Journal of Pedagogical Research, 9*(1), 156–175. https://doi.org/10.33902/jpr.202525678

Prince, M., Vigeant, M., & Nottis, K. (2012). Development of the heat and energy concept inventory: Preliminary results on the prevalence and persistence of engineering students' misconceptions. *Journal of Engineering Education, 101*(3), 412–438. https://doi.org/10.1002/j.2168-9830.2012.tb00056.x

Rahmawati, T., Wahyuningrum, M. M. H., Bustari, M., Lestari, S., Ernawati, R. D., & Selviana, S. (2023). Science learning management based on higher-order thinking skills (HOTS). *Jurnal Penelitian Pendidikan IPA, 9*(1), 533–541. https://doi.org/10.29303/jppipa.v9i1.3152

Rahmawati, Y., Ridwan, A., & Faustine, S. (2023). Promoting inclusive education through culturally responsive teaching in science classrooms. *Asia-Pacific Science Education, 9*(1), 1–28. https://doi.org/10.1163/23641177-bja10062

Ramos, J. L. S., Dolipas, B. B., & Villamor, B. B. (2013). Higher-order thinking skills and academic performance in physics of college students: A regression analysis. *International Journal of Innovative Interdisciplinary Research, 4*(1), 48–60.

Resnick, L. B. (1987). *Education and Learning to Think*. National Academy Press.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir: Panduan untuk peneliti, mahasiswa, dan psikometrian*. Parama Publishing.

Retnawati, H., Hadi, S., & Nugraha, A. C. (2018). Vocational high school teachers' difficulties in implementing the assessment in Curriculum 2013 in Yogyakarta province of Indonesia. *International Journal of Instruction, 11*(4), 33-48. https://doi.org/10.12973/iji.2018.1143a

Saido, G. M., Siraj, S., Nordin, A. B., & Al-Amedy, O. S. (2018). Higher-order thinking skills among secondary school students in science learning. *The Malaysian Online Journal of Educational Science, 3*(3), 13–20.

Santoso, B., Rosnawati, R., & Jailani, J. (2024). Gender differences in higher-order thinking skills: A meta-analysis. *Thinking Skills and Creativity, 51*(1), 101-110. https://doi.org/10.1016/j.tsc.2024.101456

Sepriyanti, N., Nelwati, S., & Yerizon, Y. (2022). The effect of 21st-century learning on higher-order thinking skills (HOTS) of students: A meta-analysis. *International Journal of Instruction, 15*(3), 127–144. https://doi.org/10.29333/iji.2022.1538a

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189x015002004

Siregar, E., Fadilah, R., & Kasmawati. (2023). Sosialisasi praktikum fisika materi suhu dan kalor di masyarakat. *Kalandra: Jurnal Pengabdian Kepada Masyarakat, 2*(2), 161–167.

Siregar, E., Fadilah, R., & Kasmawati. (2024). Analisis kesulitan mahasiswa dalam menyelesaikan soal suhu dan kalor ditinjau dari kemampuan representasi diagram. *Jurnal Riset dan Kajian Pendidikan Fisika, 11*(1), 23–34.

Taqwa, M. R. A., Suyudi, A., & Sulistijo, S. H. (2019). Profil kemampuan mahasiswa dalam menyelesaikan soal-soal suhu dan kalor berbasis representasi. *Jurnal Pendidikan Fisika dan Keilmuan (JPFK), 5*(2), 85–94. https://doi.org/10.25273/jpfk.v5i2.5467

Taqwa, M. R. A., Suyudi, A., & Sulur, S. (2020). Analisis miskonsepsi topik suhu dan kalor mahasiswa pendidikan fisika Universitas Negeri Malang. *Briliant: Jurnal Riset dan Konseptual, 5*(3), 522–530. https://doi.org/10.28926/briliant.v5i3.492

Uzunboylu, H., & Aşiksoy, G. (2014). Research in physics education: A study of content analysis. *Procedia-Social and Behavioral Sciences, 136*(1), 425-437. https://doi.org/10.1016/j.sbspro.2014.05.353

Wahyudi, W., Mufit, F., Festiyed, F., & Hidayati, H. (2024). Systematic review of students' misconceptions in thermodynamics: A comprehensive analysis using the certainty of response index. *Jurnal Penelitian Pendidikan IPA, 10*(9), 567–578. https://doi.org/10.29303/jppipa.v10i9.8901

Widana, I. W. (2017). Higher order thinking skills assessment (HOTS). *Journal of Indonesian Student Assessment and Evaluation, 3*(1), 32-44. https://doi.org/10.21009/jisae.v3i1.4859

Yanti, H., Dwi, G. S., Putri, V. M., & Zulfani, Z. (2025). Integration of validity, reliability, trustworthiness, and authenticity concepts in maintaining the quality of scientific research. *Jurnal Kajian dan Pengembangan Umat, 8*(2), 172-178. https://doi.org/10.31869/jkpu.v8i2.7403

Zohar, A., & Agmon, V. A. (2018). Raising test scores vs. teaching higher order thinking (HOT): Senior science teachers' views on how several concurrent policies affect classroom practice. *Research in Science & Technological Education, 36*(2), 243–260. https://doi.org/10.1080/02635143.2017.1395332

https://doi.org/10.17509/wapfi.111.9