Englisia
Journal of Language, Education and Humanities
ISSN:
2089-6239 (Print)
2549-0648 (Online)

HOTS and LOTS-based assessment: The challenges faced by high school EFL teachers in assessing the students' summative performance

ABSTRACT - Despite the Indonesian curriculum's explicit emphasis on Higher-Order Thinking Skills (HOTS), senior high school EFL teachers continue to experience difficulties in achieving a balanced integration of HOTS and Lower-Order Thinking Skills (LOTS) within summative assessments. This study investigates the challenges encountered by senior high school EFL teachers in assessing both higher-order and lower-order thinking skills in such assessments. Employing a qualitative descriptive design, data were collected from five EFL teachers in Magelang, Central Java, through semi-structured interviews and document analysis of teacher-made tests. The findings reveal a structural imbalance in which multiple-choice items—predominantly measuring LOTS—prevail, whereas HOTS-focused tasks remain limited in scope and frequency. This imbalance is attributed to institutional policies, students' linguistic proficiency, pandemic-induced learning gaps, and insufficient teacher training in assessment design. The study concludes that meaningful integration of HOTS is contingent upon alignment among curricular objectives, assessment formats, teacher preparedness, and student proficiency levels. Such alignment necessitates enhanced assessment literacy among teachers and the provision of comprehensive instructional scaffolding within the Indonesian EFL context.

Ivana Arlene Wellington[1*]
Ouda Teda Ena[2]
[1,2]Sanata Dharma University, Indonesia
[*]*Corresponding email:* ivanaarlene.168@gmail.com

Article History
Submitted: 6 November 2025
Accepted: 18 February 2026
Published: 31 May 2026

Keywords
HOTS; LOTS; Summative assessment; EFL teachers

Recommended APA Citation
Wellington, I. A., & Ena, O. T., Z. (2026). HOTS and LOTS-based assessment: The challenges faced by high school EFL teachers in assessing the students' summative performance. *Englisia: Journal of Language, Education, and Humanities, 13*(2), 238-252. https://doi.org/10.22373/englisia.132

1. Introduction

Nowadays, modern educational practices have exceeded their predecessors. Success and the efficient learning process are significant topics that keep the emphasis on the education sector. As part of this process, educators, especially those who are teaching young learners, must be creative and capable of preparing a wide range of learning resources in the hopes that their students will be inspired to study and grow, ultimately leading to betterment on a personal, institutional, and global scale. The importance of teaching primary school pupils English cannot be underestimated. Experts agree that age has a significant role in language acquisition, claimed Long (1990). Mastery of a language depends on when it is taught, yet anybody may learn a new language. Beginners studying a second or foreign language before puberty (around thirteen) have a better chance of reaching native-like proficiency levels.

Assessment plays a vital role in the process of teaching a language. As Jalongo states, assessment is the process of determining the worth of something (1992). The positive evaluations teachers make of the teaching-learning process indicate the students' high achievement. The preparedness of teachers, especially in determining how to assess students' learning, ensures the smoothness of the learning process. To enable the learning process to achieve its desired outcome, it is important to ensure that teachers have the appropriate assessment tool. This is especially important in the construction of the tool. There is no other way to evaluate a person's ability to think beyond the evident. This is important because contemporary education has a goal that is beyond teaching students how to solve problems and think critically (McCormick et al., 2015). This objective is to develop the ability to think at higher levels. In a world where information is rapidly changing, this ability is very important for students. According to Bloom's taxonomy, Anderson and Kraetwohl (2001) placed Higher-Order Thinking Skills (HOTS) at the top three levels: analyzing, evaluation, and production.

On the other hand, more conventional forms of assessment emphasize lower-order thinking skills (LOTS), which of course are also crucial in stimulating the higher-order thinking processes. There are possibilities as well as challenges in respect of HOTS exams for teachers of EFL. It is going to necessitate a radical change in teaching, assessment, and instructional strategies and methods but that is how educational institutions are going to achieve their objectives for the 21st century. Kusuma, et al. (2017) stated that the absence of professional growth, scarcity of materials, and deficiency of proficiency in English are among the barriers to the application of HOTS by teachers in EFL examination.

Higher-Order Thinking Skills (HOTS) has been a crucial focus in Indonesia for nearly 10 years, emphasizing its integration into English language assessment. Each point in time has had the same policies and a very similar curriculum which always advises us that learners should be; analyzing, evaluating & creating -skills abode with 21st century competences. Yet, despite all of the attention these decades-old ideas have received, findings drawn from recent research indicate that teachers are still struggling to construct assessments capable accurately gauging such higher-order skills. Pratiwi et al. (2019) argued that although senior high school teachers generally recognize the significance of HOTS but they are not able to make assessment items which could accommodate their skills effectively and therefore has resulted in summative tests

with less emphasis on lower-level cognitive processes. It was the same with Johansson (2020), Abkary and Purnawarman (2020), Wiyaka et al. (2020) elucidated the prevailing issues regarding assessment design, learner heterogeneity and comprehension of HOTS based evaluation.

While these studies have provided important contributions, they are limited in their own right and together suggest a significant gap in the literature. Some studies have focused too narrowly on Higher-Order Thinking Skills (HOTS), ignoring how the use Lower-Order Thinking Skills (LOTS) might influence assessment design. For beginners, and even some intermediates, LOTS are still an important base to build on and a good formative assessment should be about getting the balance right between them both kind of thinking ideally. Second, there is little understanding regarding how teachers' beliefs, interpretations, and emotional responses impact on their assessment decisions (which in turn greatly affects whether they are able or willing to use HOTS/ LOTS meaningfully). Thirdly, existing studies mostly attend to procedural problems or student classroom behavior and rarely analyze the actual testing corpus to provide insight into emergent item construction distributional patterns of cognitive levels as well as alignment with Bloom's Revised Taxonomy.

In response to study by Ginting and Kuswandono (2020), concerning teacher difficulty developing HOTS-based tasks in East Indonesia, an even greater need for this study arises. Yet, the continued existence of such problems in 2025 Java, a region renowned for their trade and education infrastructure simply shows that this is not an issue isolated to any one particular place. Preliminary observations from senior high schools in Magelang suggest that it remains a challenge to design and implement summative assessments enabling valid assessment of both HOTS as well as LOTS. These results suggest that teacher difficulties are not only related to the resources available, but can also be due to more deep-rooted issues such as misconceptions and mismatches of curriculum content with high stakes examinations which might result in unintentional continuation of LOTS practices.

With these gaps being still unresolved, this study is warranted to provide a new and culture-specific image on the problem of what senior high school EFL teachers truly face in relation with assessing students' higher- and lower-order thinking when it comes into summative tests. It acts as well on how teachers interpret the HOTS and LOTS, their lived experience, beliefs, and emotions regarding to the specific issues which inform assessment design, an area very less discussed in previous research. Moreover, this study offers specific examples of teachers' test materials to help illustrate how HOTS and LOTS are currently observable in classroom assessments. Collectively these insights are intended to provide specific recommendations for instruction so that together teachers, school leaders, and policy makers can better support the requisite skills students require in a 21st century globalized world by enhancing assessment literacy and HOTS- & LOTS-based evaluation practices within Indonesian EFL contexts. Therefore, this study tries to answer this research question: What challenges do Indonesian senior high school EFL teachers face in assessing students' higher- and lower-order thinking skills through summative assessments?

2. Literature review

2.1. High order thinking skills (HOTS) and lower order thinking skills (LOTS)

A teacher has helped their students learn more effectively when that teacher realizes that teaching is about helping them comprehend, not just transmitting knowledge and information (Muhammadiah et al., 2022b). It is prevalent for teachers to consider how they might include higher-order thinking skills in their lesson plans and assessments. There are two main types of critical thinking skills: LOTS (lower-order thinking skills) and HOTS (higher-order thinking skills). These statements, which come from Bloom's Taxonomy, a way of organizing brain processes in a hierarchy, show how thought processes progress from the most basic to the most complex. Students must acquire LOTS before progressing to higher-order thinking skills (HOTS). Memory and understanding are the hallmarks of lower-order thinking skills (LOTS) (Jansen & Möller, 2022). Among the several LOTS abilities are the following: remembering (C1), understanding (C2), and applying (C3). On the other hand, HOTS provides a structure for how students' critical thinking abilities develop through school (Mubarok & Anggraini, 2020). Capabilities in analyzing (C4), evaluating (C5), and creating (C6) are all part of the Higher-Order Thinking Skills group (HOTS).

In order to begin thinking, one must first remember (Lau et al., 2018). Next, students are asked to describe, explain, inform, or identify a particular issue. Then, comprehending is the following stage. It happens when students fully grasp the material they have read. Students should focus on retelling, inferring, interpreting, explaining, forecasting, and outlining knowledge at this stage. The third stage is applying. It is the responsibility of the students to find creative ways to put what they have learned into practice. Performing analyses is the following task. Examining anything entails discovering its individual pieces and how they integrate to form a more apparent whole. Differentiating, arranging, and identifying are just a few of the many terms employed. Evaluation is the next stage. This decision-making process involves considering several factors before settling on a final option. Creating something new is the last phase. When students create something, they combine parts to make an outline that makes coherence and work. Problem-making, planning, and production are the words that describe this stage (Anderson & Kraetwohl, 2001).

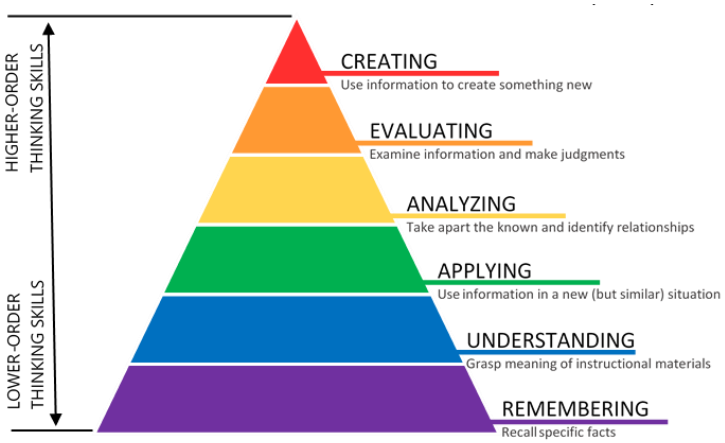

Figure 1. Bloom's taxonomy – Cognitive domain (2001)

Table 1
Cognitive levels (Bloom's taxonomy revised by Anderson and Krathwohl).

Cognitive level	Cognitive process	Action words	Criteria
Recalling (C1)	Identifying, retrieving	Naming, repeating, copying, imitating, recognizing, stating	LOTS
Understanding (C2)	Explaining, demonstrating, organizing, summarizing	Clarifying, interpreting, illustrating, describing, defining	
Applying (C3)	Carrying out, putting into action	Using, showing, applying, executing, performing, demonstrating	
Analyzing (C4)	Breaking down, distinguishing, categorizing	Comparing, differentiating, examining, critiquing, assessing	HOTS
Evaluating (C5)	Judging, concluding, justifying	Checking, evaluating, deciding, supporting, defending, reviewing	
Creating (C6)	Designing, generating, constructing	Building, producing, formulating, writing, planning, arranging	

2.2. Summative assessment

Summative assessment is a very important part of all school systems worldwide. It evaluates a student's learning at the end of a teaching unit by comparing it to a standard or norm. Summative assessment refers to the process of evaluating student learning that takes place at the completion of a unit of learning. It usually takes the form of exams, final projects, or standardized tests. People often see formative testing, which is meant to see how kids grow as they learn, as something that goes against this. Meanwhile, in summative assessment, projects, regular tests, and final exams are all important ways to see how well students are fulfilling the learning goals. Formal tests do more than just measure how well students are doing; they also greatly impact teachers' work and the program. After reviewing 23 studies, Harlen (2005) concluded that final tests that are part of normal classroom activities like projects and portfolios can motivate students to learn and do better when properly incorporated.

On the other hand, high-stakes tests can make students more test-anxious and may cause teachers to "teach to the test," which could hurt larger educational goals. As a result, the capability of educators is an essential component of this scenario, as it is necessary to develop an effective HOTS-based product for their students. In addition, it is fundamental for teachers to have an extensive understanding of the brain processes that are involved. That they are able to differentiate between various cognitive processes is very necessary for the development of the exam (Pratiwi, Dewi, & Paramartha, 2019).

2.3. Assessing senior high school learners

Evaluating proficiency in senior high school literacy requires an understanding of adolescent cognitive and developmental characteristics that affect what they know about language. Students at the secondary level are required higher-order thinking tasks, larger-scale extended texts and more abstract reasoning than earlier grades. According to Anderson and Krathwohl (2001), as students get older, they need HOTS tasks since these are the sorts of cognitive processes which assesses their ability to analyze, evaluate and create- essential for academic success in preparation for higher education including real-world problems. However, good assessment should also recognize the motivational and proficiency variations, and diversity of starting points that can exist among test takers in EFL contexts such as these (Ahmadi & Nasiri, 2025). Researchers emphasize the need to provide different types of assessment methods when assessing high school EFL learners. Such methods as performance tasks, argumentative writing, problem-solving activities and project-based assessments provide LOTS in authentic contexts that include elements of HOTS (Respati, 2023). Formative feedback, additionally remains to be necessary. Black and Wiliam (1998, p. 139) explains that feedback should be focused: prompt for senior high school students to refine their ideas; regulate their learning and skilled transfer across subjects. However, studies suggest that many high school teachers are still struggling to develop HOTS-based assessment.

The problems that teachers ever faced concerning hampered students' vocabularies, not efficient reading comprehension, and unprecedented types of open-ended questioning (Abkary & Purnawarman, 2020). These limitations have the effect of leading to LOTS-based assessments such as multiple choice which are quicker to score but do little in terms of how well students are thinking and analyzing. In addition, systemic pressures such as standardized testing and school-wide test score goals discourage the regular integration of HOTS (Widana, 2020). One best practice of scholars in the field was to scaffold assessment practices that would build up students' cognitive readiness as a way to mitigate these concerns. For example, this might involve model answers or metacognition teaching as part of the assessment structure and clear structures in place for learners to understand what they are expected to produce (Zou et al., 2024). Vygotsky's (1978) socio-cultural perspective is also still applicable at the secondary level. Where appropriate, assessment should measure not just what students can accomplish on their own but also what they are able to achieve with guidance and scaffolding, peer collaboration, or classroom discourse. This kind of methodological solution simultaneously promotes fairness and equitability in the assessment of EFL learners' communicative competence and higher-order skills.

3. Method

The purpose of this study is to examine the challenges senior high school EFL teachers face when measuring HOTS and LOTS, and to provide feedback on the process of summative assessments conceived through a qualitative descriptive lens. The researcher intended to collect qualitative information to demonstrate the multi-faceted nature of teachers' experiences and daily assessment. Assessment practices emerge from contextual forces, such as the institution,

the preparedness of the students, the teachers' reflective professional judgement, and the unique dynamics of the classroom. For the purpose of understanding the interplay of context through the lens of an interpretivist approach, it is the researcher's aim to examine the assessment practices within the framework of Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001), to define and classify the cognitive processes targeted by the teacher's assessment tasks.

Therefore, this study focused on five senior high school English teachers in the city of Magelang, located in Central Java. The participants were selected through purposive sampling considering it focused on participants who met the research aims. All participants were teaching at senior high school level, were involved in the designing and marking of summative assessments, and had a minimum of three years of teaching experience. These criteria guaranteed that participants had enough experience with curriculum and assessment paper construction. The small number of participants is justified. The study had an in-depth focus on a selected context and therefore, the findings were meant to be pedagogically useful for a specific context and not meant to be generalised to all Indonesian EFL contexts.

The two primary methods for gathering data involved semi-structured interviews and document analyses. Each participant was interviewed one-on-one for 45-60 minutes. This format was beneficial for the researcher to manage the discussion on important themes (such as teachers' knowledge of HOTS vs. LOTS, experiences designing summative assessments, challenges faced, and views about institutional training), while also providing the participant with the latitude to provide details for issues they deemed to be significant. All interviews were also conducted with the participant's consent and recorded.

To supplement interview data and add credibility to the findings, document analysis was performed on selected summative exam papers developed by the teachers from the previous year. These documents are analysed for the cognitive level of the assessment items. Instead of focusing on operational verbs, each item was considered for the level of reasoning needed from the students. Using Bloom's Revised Taxonomy, items that elicited remembering (C1), understanding (C2) and applying (C3) were classified as LOTS, while those that needed analysing (C4), evaluating (C5) and creating (C6) were considered HOTS. This methodology enabled the research to go beyond the expressed beliefs of the teachers and examine most summative assessments for the cognitive levels. The document analysis combined with interviews gives a better perspective of the gap between the described practice and the patterned assessments.

As per Braun and Clarke's (2006) thematic analysis guideline, the data analysis started off with the reading of interview transcripts with the goal of understanding the data, and especially the recurring patterns. The segments that had significance were coded concerning assessment design, institutional limitations, student readiness, and teacher professional development. Coding (or assessment of data) was done personally, as the aim was to stay as engaged as possible with the language of the participants and the contextual meanings of the transcripts. Related codes to the initial codes were grouped as codes that were broader (or more inclusive) to the repeated concerns of the participants. The researcher, in following the iterative process, continually reviewed and refined the coherence of the themes to avoid overlap. It was

in the constant movement of the researcher between the emerging meanings and the raw data that was most essential to sustaining in the accounts of the participants. The analysis outcomes of the documents were also thematic, with the goal of assessing the consistency of the explanation by teachers and the cognitive levels in their tests at the intervals of the analysis to assess construct consistency of the intended learning outcomes and the assessment tasks.

The evidence was checked, and teacher responses were triangulated with assessment materials through interviews and documents. Moreover, interview summaries were checked for accuracy and misunderstandings. Also, peer debriefing was conducted with a colleague of mine who studies qualitative research, so bias in coding and theme identification was mitigated. The ethical components of the research were very important to the researcher. All participants were made aware of the research purpose, as well as their right to withdraw from the study at any point without consequence. All data was kept confidential and securely stored. Pseudonyms were used in the research findings (Teacher A to Teacher E). Importantly, all data was kept confidential and securely stored. Pseudonyms were used in the research findings (Teacher A to Teacher E) and all data were kept for research purposes.

4. Findings

Three general themes emerged from the data: assessment design, implementation obstacles, and support for and preparation of teachers. To show how the dataset is converging and diverging, each theme uses all three data sources, which are interviews with EFL instructors, survey results, and examination of PAT exam documents.

4.1. Designing assessment

Teachers emphasized the strong connection between their summative assessments and institutional requirements. The curriculum part usually includes a plan showing how many questions should cover both HOTS and LOTS. One teacher explained: *"*The blueprint is made by the curriculum section. We just follow it: a certain percentage for HOTS, a certain percentage for LOTS.*"* (Interview, Teacher A). When asked to describe the difference between the two, teachers commonly framed LOTS as focusing on recall and understanding, often tested through multiple-choice items. As one teacher put it: *"*LOTS questions are those that can be learned and memorized. Usually there are answer choices. HOTS is different. It requires students to analyze, for example using a table, or sometimes through a short dialogue.*"* (Interview, Teacher B). This data shows that teachers are aware of the theoretical distinction between LOTS and HOTS, but in practice, HOTS is usually realized only in specific task types such as dialogue analysis or essay-style questions.

One practical example of how assessment design appears in practice is the PAT exam that the school provides. For a total of forty questions, thirty were multiple-choice (representing 75%) and ten were essay/constructed responses (representing 25%). The main goals of the multiple-choice portion were to test memorization, understanding, and practical application. The goal of the text, meaning inferred from context, and grammar rules were some of the examples given to the students. The majority of these crucial abilities are covered by LOTS (C1-C3). In

contrast, students were asked to write brief texts for the essay component, such as hortatory expositions or phrases combined with conjunctions. Since these activities involve reasoning, assessment, and production, they are better suited to HOTS (C4-C6). However, LOTS is clearly the head of the assessment system, since just 10 essay questions are weighted, while 30 items are multiple-choice.

Additional evidence of this trend can be found in survey data. Nearly two-thirds of the teachers surveyed felt they understood the distinction between LOTS and HOTS, and 100% were confident in their capacity to formulate HOTS questions. Fascinatingly, though, nobody picked "strongly agree," which would indicate some reluctance. Another area where things seemed to be lacking was training; just 33% of educators were satisfied with the amount of preparation they had received, while the rest were somewhere in the middle. Despite systemic pressures to stick to multiple-choice formats, most nevertheless admitted to using projects, presentations, and open-ended questions as part of their evaluations. When these results are considered collectively, they reveal a trend of limited striving. Teachers strive to include more analytical and creative items, and they theoretically commit to balancing HOTS and LOTS.

4.2. Implementation challenges

Teachers were quite open about the challenges they have while using HOTS-based evaluations in the classroom. The most significant obstacles were found to be vocabulary and understanding. One teacher considered: "Children nowadays lack vocabulary. When asked to comprehend, they surrender quickly. Grammar is still difficult for them, and even simple questions they cannot answer." (Interview, Teacher C)

There was also an emphasis on the long-term consequences of learning loss during the pandemic. Teachers felt that students were not prepared for high school because of the ineffectiveness of remote learning, which was most noticeable at the lower secondary level: "The pandemic caused a lot of learning loss. Online classes were not effective. Now in high school, the students are still struggling." (Interview, Teacher A)

In addition, the difficulty that students had when faced with questions that did not have multiple-choice answers also repeated. The structure of multiple-choice questions made individuals feel more comfortable guessing or eliminating, whereas the unknown of open-ended questions made them nervous: "When the questions are not multiple choice, students get confused. They don't know how to answer." (Interview, Teacher D)

These challenges are reflected in the survey data. Assessment of students HOTS in various classes was considered difficult by 66.7 % and relatively easy with only 33.3 %. The fact no one dissented indicates everyone knows the task. This apparent challenge opposes their previous belief of being able to produce HOTS goods It means that the issue is not with questions but student's ability to answer them correctly. When asked whether they were comfortable providing comments on students' HOTS work, responses indicated the following: Strongly Agree (33.3%), Agree (66.7%) and Unsure (6.7%). This suggests that while teachers continue to struggle with student preparedness in respect of feed-forward, they now at least see it as practical proposition.

Together, these challenges point to a systemic issue: the difficulties in HOTS assessment are not simply a matter of teacher incompetence. Instead, they reflect long-standing pedagogical traditions that continue to drain classroom time and limit students' opportunities to build the vocabulary, skills, and open-ended thinking required for higher-order tasks and the issues worsened by lingering pandemic learning loss. Many teachers can write HOTS items, but the survey shows that students are not yet prepared to respond to them meaningfully because their learning experiences have not developed the cognitive habits needed for such thinking. This mismatch exposes a deeper problem: teachers may know how to create HOTS questions, but without understanding how to teach and assess higher-order thinking, the questions themselves lose their value. Even the best-designed items are ineffective if students are not developmentally ready to engage with them. It is that students have not been systematically taught how to think through those cognitive strategies and linguistic resources essential to answer them. This thus highlights the necessity of progressive scaffolding, long-term frequent exposure to non-multiple-choice items as well as explicit vocabulary development before HOTS instruction can take place.

4.3. Teacher preparation and support

In every case, teachers cited continuing education opportunities provided by their schools. This was done through twice-monthly study days that were supervised by education authorities and featured guest lecturers. As a teacher expressed it: *"*The institution provides training. Twice a month we have Teacher Training Days. There are guest speakers, and sometimes the supervisors accompany us.*"* (Interview, Teacher E). Nevertheless, differences between generations were also noticeable. It was noted that older teachers were more hesitant to implement strategies centered around HOTS: *"*Older teachers are difficult to invite into using HOTS. They still rely on old test banks.*"* (Interview, Teacher B). Teachers also expressed a strong desire for more practical resources. They highlighted the need for sample HOTS items, digital files, and clear technical guidelines for test construction: *"*We need tools. Guidelines for how to write HOTS items, examples, and soft files. Without that, teachers prefer to just use LOTS, because at least students get high scores.*"* (Interview, Teacher A)

Responses to a survey were inconsistent with these narratives. The majority of teachers (66.7%) were neutral in response to whether they had received enough training, while 33.3% agreed with the statement. That means they have access to some training, but it is not really meeting their needs.

When asked if they used varied assessment methods (open-ended questions, projects, presentations), 66.7% strongly agreed, suggesting a willingness to experiment. Yet the interviews reveal that such variety is difficult to sustain without stronger institutional backing and concrete examples.

The interviews and surveys show that although schools provide some training, it does not fully address the real problems teachers face when trying to use HOTS in their classroom assessments. The training sessions happen regularly, but they focus more on giving information than on helping teachers practise, receive feedback, or work together to create better HOTS

items. There is also a clear difference between generations. Many senior teachers still rely on old test banks because these materials feel safer and easier to use, especially when schools place strong pressure on good student grades. Teachers also explained that what they truly need are practical tools, examples, and digital item banks that show clearly how HOTS questions should look across different topics and levels. The survey results support this. Teachers are willing to try different types of assessment, but they are not fully satisfied with the training they receive. This means they have the motivation but not enough support. Altogether, the findings suggest a need for more hands-on training, stronger mentoring, and school policies that encourage teachers to try new approaches without fear of failing or lowering grades.

5. Discussion

This study illustrates a dilemma concerning assessment in high school EFL: though teachers grasp the difference in concepts between HOTS and LOTS and the value of higher-order thinking, out of the totality of summative assessments, almost all the structural components are weighted towards lower-order thinking. This dilemma is more than a question of teacher skill; it encompasses the blend of curriculum constraints, school demands, student preparedness, and legacy practices in assessment.

5.1. Designing assessments: Between policy and practice

Teachers often create their assessments based on school or institutional guidelines, especially rules about how many higher-order thinking skills (HOTS) and lower-order thinking skills (LOTS) questions they should include. Although teachers show some understanding of higher-order thinking, the way they include it in their assessments is often influenced by the need to follow different formats and to prove that they are meeting official requirements. Instead of thinking about how to incorporate HOTS throughout the assessment, it seems that higher-order thinking tasks were relegated to a single area of the assessment while lower-order thinking tasks were overrepresented in dominant positions in the assessment.

This construct under-representation exemplifies a deeper, systemic concern that has been noted in the literature. When curriculum and assessment are supposed to focus on higher-order thinking skills, like analysing, evaluating, and creating (Anderson & Krathwohl, 2001), but instead mainly emphasize memorizing and recognizing information, then what we truly want to measure is not properly assessed. In other words, higher-order thinking becomes overlooked. Research on HOTS education also suggests that one reason for this problem is that some teachers may struggle to design assessments that effectively include higher-order thinking skills (Pratiwi et al., 2019; Johansson, 2020). In the case of the findings presented in the current research, the problem is greater than the perceived technical issue. It exhibits the design of end of semester assessments where the need for standards, ease of scoring, and institutional conformity dominate the mental processes they represent.

Furthermore, it seems that the decisions made by teachers are more about pragmatic and ethical motivations, not about being resistant to reform. The thoughtful balancing of cognitive demand and concerns about student performance, the grading burden, and the fairness concerns

reflect the human side of assessment design. As the alignment research indicates (Khan et al., 2025), purposeful curriculum-assessment coherence goes beyond the balanced cognitive level. It also entails the alignment of learning intentions, the structure of the task, and the level of cognitive demand. The coherence in this study is still partial and not complete.

5.2. Implementation challenges: The legacy of learning loss

A second major issue deals with the interplay of linguistic readiness and cognitive complexity. Teachers uniformly mentioned students' lack of vocabulary and grammatical control which hindered their ability to perform more advanced tasks. This finding signals a central issue in EFL testing: the expression of higher-order thinking is required in a foreign language. As a result, the cognitive task is fused with the language task.

This situation gives a question whether the assessment is really measuring what it is supposed to measure. If students have difficulty answering HOTS questions mainly because of their limited English skills, and not because they cannot think critically, then the test may be measuring their language ability more than their thinking skills. Other studies in Indonesian EFL classrooms have also found similar problems (Abkary & Purnawarman, 2020; Kusuma et al., 2017). However, this study adds something new by looking at the issue after the pandemic. Teachers shared that during remote learning, students' basic skills became weaker. Because of this, moving directly to more challenging, higher-level thinking tasks was not realistic for many students.

Moreover, students' prolonged exposure to certain formats is thought to modify their assessment approach. When clear reasoning is needed for close-ended tasks, open-ended tasks can create confusion. This confusion is not due to the absence of thought but is due to a lack of exposure to that particular format. This reinforces the notion that a shift to HOTS questioning requires significant and more extensive changes to pedagogy, and is not simply about placing more questions at a higher order of thinking (Wiyaka et al., 2020). A lack of gradual scaffolding suggests that students may be ill-prepared for the posed expectations. This raises questions about the balance of equity and the integrity of the assessment.

5.3. Teacher preparation and support: Gaps between training and application

Some teachers in this study reported using AI-assisted tools to generate or revise assessment items. While such tools may increase efficiency, their pedagogical value depends on how critically teachers align them with cognitive objectives and curricular goals. Furthermore, institutional training was seen by teachers as far too theoretical. There was a disconnect between the theory and the practice. While theory creates a good framework, practice creates a good foundation. Theoretical frameworks are like the skeletons of the body. To run a classroom well, every important part needs proper attention and support. If even one key part is ignored, the classroom may feel incomplete and lifeless, lacking energy, meaning, and connection. Teachers will not be able to create a classroom that is both operational and successful. As a result, teachers will not be able to move beyond a superficial classroom that only includes a simplistic structure of HOTS.

Systemic pressure is the other influence on teachers' assessment choices. In a situation where there is pressure arising from the situation, teachers have to choose between the littoral and the deep. In other words, they have to choose between making the assessment deep or shallow. Teachers are working with the prescribed structure that is determined by the organisation of external pressures on the teachers. It is within these confines that the teachers have to function.

In other words, teachers are working in an environment where the structure of external pressures is not determined by the teachers themselves. It is these external structures that have to be put in order before teachers can successfully use a greater range of HOTS. The external structures mustn't be perceived as an end in themselves as they will need to be accompanied by a greater range of HOTS. The end must be combined with the external structures to create an end that the teachers can then have a greater range of assessment that can then enable teachers to have a greater range of HOTS.

6. Conclusion

This study further reveals that, despite teachers' recognition of the value of Higher-Order Thinking Skills (HOTS) in fostering critical thinking, most assessments remain dominated by Lower-Order Thinking Skills (LOTS), with HOTS tasks largely confined to essay questions or special assignments. Teachers encounter significant challenges in effectively integrating HOTS into summative assessments, attributable to factors such as students' limited linguistic proficiency, learning loss resulting from the COVID-19 pandemic, and a lack of practical training in assessment design.

Notably, the study also finds that teachers are actively seeking strategies to enhance their assessment practices. Several participants reported initial explorations of digital and AI-assisted tools to generate test items or refine question formats. However, their reflective accounts indicate that such tools are most beneficial when applied critically and in alignment with curricular objectives. In the absence of robust assessment literacy, AI-generated tasks risk perpetuating surface-level question types rather than substantively supporting higher-order thinking. Consequently, while technology may serve as a supplementary resource, it cannot supplant teachers' professional judgment in ensuring construct alignment and cognitive coherence.

Accordingly, policymakers should implement more adaptable assessment frameworks, practical and context-specific training, and sustained professional development opportunities that accommodate teachers across different generational cohorts. To support these efforts, schools may consider developing HOTS item banks, promoting collaborative test design, and prioritizing students' long-term cognitive development over immediate test performance. Collectively, such initiatives can facilitate a more balanced integration of HOTS and LOTS in Indonesian EFL classrooms, thereby enhancing students' critical thinking abilities and their performance on assessments.

The present study is not without limitations. Its restricted sample size—comprising only five high school teachers from Magelang, Central Java—may limit the generalizability of the findings to the broader diversity of EFL classroom contexts across Indonesia. Future research

should address these limitations by comparing urban and rural settings, examining the experiences of novice versus experienced teachers, or employing larger and more heterogeneous samples drawn from multiple geographic regions. Notwithstanding its scope, this study offers valuable insights into the practical and pedagogical challenges of balancing HOTS and LOTS in summative assessment. With enhanced institutional support, flexible policies, and further empirical inquiry, Indonesian EFL education can move toward assessment practices that not only measure knowledge acquisition but also cultivate the higher-order thinking skills essential for success in the 21st century.

Declaration on the use of AI

During the writing process of this paper, the author used Grammarly for grammar and language checking. In addition, ChatGPT was used to assist in improving the flow, clarity, and overall organization of the writing. However, all research ideas, analyses, interpretations, and conclusions presented in this paper are entirely the author's own work. Furthermore, all AI-generated suggestions were carefully reviewed, revised, and edited before being incorporated into the manuscript. Therefore, the author takes full responsibility for the accuracy and integrity of the final content. Additionally, no confidential, personal, or sensitive information was shared with any AI tool during the writing process.

References

Abkary, N. S., & Purnawarman, P. (2020). Indonesian EFL teachers' challenges in assessing students' higher-order thinking skills (HOTS). *Proceedings of the 4th International Conference on Language, Literature, Culture, and Education (ICOLLITE 2020)*. https://doi.org/10.2991/assehr.k.201215.076

Anderson, L. W., & Kraetwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessment*. Longman.

Astrid, A., Hasanah, A., & Syafryadin, S. (2022a). Integrating higher order thinking skills (HOTS) into English language teaching for elementary school students: Teachers' perspectives and challenges. *3L The Southeast Asian Journal of English Language Studies*, *28*(3), 217–230. https://doi.org/10.17576/3l-2022-2803-14

Black, P., & Wiliam, D. (1998). *Assessment and classroom learning*. Assessment in education: Principles, policy & practice, 5(1), 7-74.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Daud, A. (2017). Challenges of using portfolio assessment as an alternative assessment method for teaching English in Indonesian schools. *International Journal of Educational Best Practices. 1*. https://doi.org/10-114. 10.31258/ijebp.v1n2.p106-114.

Ghasemi, B., & Hashemi, M. (2011). Foreign language learning during childhood. *Procedia - Social and Behavioral Sciences*, *28*, 872–876. https://doi.org/10.1016/j.sbspro.2011.11.160

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Jalongo, M.R. (1992). *Early childhood language art.* Allyn and Bacon.

Jansen, T., & Möller, J. (2022). Teacher judgments in school exams: Influences of students' lower-order-thinking skills on the assessment of students' higher-order-thinking skills. *Teaching and Teacher Education*, *111*. https://doi.org/10.1016/j.tate.2021.103616

Johansson, E. (2020). The assessment of higher-order thinking skills in online EFL courses. *Nordic Journal of English Studies*, *19*(1), 224–256. https://doi.org/10.35360/njes.519

Khan, H. F., Qayyum, S., Beenish, H., Khan, R. A., Iltaf, S., & Faysal, L. R. (2025). Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. *BMC medical education*, *25*(1), 200. https://doi.org/10.1186/s12909-025-06736-4

Kusuma, M. D., Rosidin, U., Abdurrahman, A., & Suyatna, A. (2017). The development of higher order thinking skill (Hots) instrument assessment in Physics study. *IOSR Journal of Research & Method in Education (IOSRJRME), 07*(01), 26–32. https://doi.org/10.9790/7388-0701052632

Long, M. H. (1990). Maturation constraints on language development. *Studies in Second Language Acquisitions, 12(3)*, 251 – 285. http://www.jstor.org/stable/44488300

Mubarok, H., & Anggraini, D. M. (2020). Literation skill to improve higher-order thinking skills in elementary school students. *Al-Bidayah: Jurnal Pendidikan Dasar Islam*, *12*(1), 31–42. https://doi.org/10.14421/AL-BIDAYAH.V12I1.234.

Pratiwi, N. P., Dewi, N. L., & Paramartha, A. A. (2019). The reflection of Hots in EFL teachers' summative assessment. *Journal of Education Research and Evaluation*, *3*(3), 127. https://doi.org/10.23887/jere.v3i3.21853

Quesada, A. G. (2023). Assessment of young English-language learners. *Revista de Lenguas Modernas*, (36). https://doi.org/10.15517/rlm.v0i36.48313

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

Wiyaka, W., Prastikawati, E. F., & Kusumo Adi, A. P. (2020). Higher-order thinking skills (hots)-based formative assessment: A proposed model for language learning assessment. *Vision: Journal for Language and Foreign Language Learning*, *9*(2), 115–130. https://doi.org/10.21580/vjv9i25859