



# Development Of Higher Order Thinking Skills (HOTS) Assessment in Plant and Animal Reproduction Systems

Erna Juwita<sup>1</sup>, Prasetyorini<sup>2\*</sup>, Dadang Jaenudin<sup>3</sup>, Vinsensius Felix Putra Andy Liman<sup>4</sup>

<sup>1,3</sup>Department of Science Education, Graduate School, Universitas Pakuan, Indonesia

<sup>4</sup>University of the People, Pasadena, California, United States

✉\* [prasetyorini@unpak.ac.id](mailto:prasetyorini@unpak.ac.id)

Article Info	Abstract
<p><b>Keywords:</b> 4D model; higher order thinking skills; HOTS assessment; item development; reproduction system; science education;</p> <hr/> <p><b>Received:</b> 03/01/2026 <b>Revised:</b> 25/03/2026 <b>Accepted:</b> 12/05/2026</p>	<p><b>Background:</b> Assessment practices in schools are still largely dominated by Lower Order Thinking Skills (LOTS) items, with 72% categorized at the C1 level (remembering) and the remainder at the C2 level (understanding), as confirmed through analysis of teacher-made exam documents and classroom observation records. This condition limits students' opportunities to develop higher-order thinking skills such as analysis, evaluation, and creation. Therefore, an appropriate assessment design strategy is needed to support the development of HOTS, particularly in science learning on the topic of plant and animal reproduction, with implications for improving both assessment quality and student cognitive engagement.</p> <p><b>Methods:</b> This study employed a research and development approach using the Four-D (4D) model (Define, Design, Develop, and Disseminate). Data were collected through analysis of existing assessment practices, expert validation, empirical testing, and responses from students and teachers.</p> <p><b>Results:</b> The instrument produced 30 HOTS items aligned with C4–C6 cognitive levels and supported by contextual stimuli. Expert validation averaged 90.63% (Aiken's V = 0.875). Empirical testing showed 24 valid items with a reliability coefficient of 0.868, moderate difficulty levels, and good discrimination indices. Gender analysis revealed significant score differences (<math>p = 0.000</math>): female students excelled at C4–C5, while male students performed better at C6. Student responses were highly positive (98.70%), and teachers confirmed its practical applicability.</p> <p><b>Conclusion:</b> The HOTS assessment strategy based on the 4D model is effective in producing valid, reliable, and pedagogically relevant instruments, contributing to the improvement of science education quality and the advancement of students' higher-order thinking skills.</p>

## INTRODUCTION

Twenty-first-century learning emphasizes the importance of higher-order thinking skills (HOTS), requiring students to analyze, evaluate, and create solutions to complex problems (Vrbanović Lisac et al., 2025). In this context, HOTS-based assessment becomes a crucial component for measuring and directing students' cognitive development (Anderson & Krathwohl,

2001). Within biology learning, the topic of plant and animal reproductive systems requires assessment instruments capable of evaluating conceptual understanding and advanced cognitive processes (Fuadiyah et al., 2023). Therefore, a systematic and well-structured assessment design is essential to support the achievement of higher-order thinking competence in biology education (Crowe et al., 2019).

One persistent issue in school-based assessment practices is the dominance of low-level cognitive items, which primarily measure factual recall (Zakariya, 2022). Field observations and document analysis of teacher-constructed exam questions at the junior high school level revealed that the majority of test items were classified at the C1 (remembering) and C2 (understanding) levels, with no items targeting analysis, evaluation, or creation. Documentation of these exam questions confirmed this systematic gap. Although basic comprehension questions remain relevant, they are insufficient to stimulate analytical and reflective thinking. Consequently, students often rely on memorization rather than connecting concepts or applying knowledge in real-life contexts. This condition reduces learning quality and contradicts curriculum reform emphasizing 21st-century competencies (Díaz et al., 2022). Thus, developing HOTS-focused assessment strategies is necessary to ensure that learning outcomes are not limited to lower-order cognitive achievements (Taber, 2017).

The reproductive system topic is essential in biology because it relates to the continuity of life and connects to environmental and health-related phenomena (Vineis et al., 2020). The abstract and complex nature of the concepts, along with their contextual relevance, offers opportunities for integrating HOTS-based problem-solving. Through well-designed HOTS items, students can compare reproductive mechanisms, examine biological implications, and propose solutions to contextual issues (Fuadiyah et al., 2023). For this reason, contextual stimuli such as diagrams, tables, and authentic cases are fundamental in constructing meaningful HOTS items (Razmjoo & Kazempourfard, 2012).

Developing HOTS assessment strategies requires identifying learning objectives, formulating indicators, writing items, and ensuring cognitive demand alignment (Khoy, 2025). Teachers must possess the competence to design questions that target analysis, synthesis, and evaluation levels following the revised Bloom's taxonomy. Equally important is ensuring alignment with learning contexts, difficulty balance, and instructional relevance (Ravand et al., 2025). A structured strategy allows teachers to produce valid and comprehensive assessments that accurately measure students' higher-order thinking abilities (Li et al., 2024). The impact of HOTS-based learning design has been documented in several studies: Ritchie et al. (2019) found that HOTS-oriented assessments enhanced conceptual engagement, while Simonsmeier et al. (2025) demonstrated that contextually rich items improve both measurement reliability and student motivation.

Validation involving content experts and educational assessment specialists is essential in ensuring instrument quality. Expert review assesses content accuracy, construct alignment, and readability of the items (Tavakol & Dennick, 2011). Qualitative evaluations help refine contextual clarity, cognitive focus, and linguistic precision (Jarupunphol et al., 2025). Following international research standards, expert validation must also be complemented by empirical analysis such as item validity, reliability, difficulty level, and discrimination power to guarantee psychometric robustness (Han & Xiang, 2025).

A limited trial with students is needed to determine the practicality and effectiveness of the assessment in classroom settings. Student responses typically reflect engagement, readability, and perceived challenge, which are important indicators of instrument feasibility. Furthermore, examining gender-based differences offers additional insight into fairness and cognitive performance patterns, ensuring that assessment results are representative and equitable

(Arendse, 2021). Despite many studies on HOTS assessment, research specifically focusing on the reproductive system topic using a systematic development model remains limited. Existing studies rarely integrate contextual stimuli, gender analysis, and psychometric validation within a single framework. This gap highlights the need for a comprehensive strategy that is both pedagogically sound and empirically validated.

Therefore, this study aims to develop a Higher Order Thinking Skills (HOTS) assessment design strategy for the topic of plant and animal reproduction using a structured development model to produce valid, reliable, feasible, and pedagogically relevant assessment instruments for biology learning. The findings of this study are expected to serve as a practical reference for educators and curriculum developers seeking to improve assessment quality, promote deeper student cognitive engagement, and contribute to the broader advancement of science education at the junior high school level.

## METHODS

This study applied a Research and Development (R&D) approach aimed at producing a valid, practical, and effective HOTS-based assessment strategy for the topic of plant and animal reproductive systems. The development process followed the 4D model consisting of Define, Design, Develop, and Disseminate stages (Thiagarajan et al., 1974).

In the Define phase, curriculum documents, learners' needs, and conceptual difficulties were analyzed. This included direct observation and documentation of teacher-made exam questions currently in use at the target school, which formed the empirical basis for identifying the dominance of lower-order cognitive items. In the Design phase, assessment blueprints were prepared, HOTS-oriented test items were drafted, and validation instruments for experts and students were designed. During the Develop phase, the drafted assessment was evaluated by material, media, and evaluation experts to determine content validity and clarity before undergoing revisions and limited trials with students. Finally, in the Disseminate phase, the improved instrument was implemented and analyzed for broader readiness.

The participants included three expert validators (consisting of two subject-matter lecturers and one assessment specialist) and two experienced science teachers, alongside a group of 35 students selected through purposive sampling. Expert validators assessed the content, construct, and readability of the items, while students participated in the limited field test to determine practicality and effectiveness.

**Table 1. Summary of Data Analysis Techniques**

Analysis Technique	Formula / Tool	Reference
Content Validity (Aiken's V)	$V = \sum s / [n(c-1)]$	Aiken (1985)
Reliability (Cronbach's Alpha)	$\alpha = k / (k-1) \times (1 - \sum \sigma_i^2 / \sigma_t^2)$	Cronbach (1951); Taber (2017)
Item Validity, Difficulty, Discrimination Index	Point-biserial; P; D	Tavakol & Dennick (2011)
Learning Gain (N-Gain)	$g = (\text{Post} - \text{Pre}) / (100 - \text{Pre})$	Hake (1999)
Normality (Kolmogorov-Smirnov)	Sig. > 0.05	Field (2018)
Homogeneity (Levene's Test)	Sig. > 0.05	Field (2018)
Paired Sample t-test	$t = \bar{D} / (SD/\sqrt{n})$	Cohen (1988)
Effect Size (Cohen's d)	$d = (M_{\text{post}} - M_{\text{pre}}) / SD_{\text{pooled}}$	Cohen (1988)

The key formulas applied are as follows. Content validity was computed using Aiken's V:

$$V = \sum s / [n(c - 1)] \quad (1)$$

where  $s = r - l_0$  ( $r$  = expert score;  $l_0$  = lowest possible score),  $n$  = number of experts, and  $c$  = number of rating categories. Values closer to 1.0 indicate stronger content validity. Instrument reliability was assessed using Cronbach's Alpha:

$$\alpha = [k / (k - 1)] \times [1 - (\sum \sigma_i^2 / \sigma_t^2)] \quad (2)$$

where  $k$  = number of items,  $\sigma_i^2$  = item variance, and  $\sigma_t^2$  = total test variance. A coefficient of  $\alpha \geq 0.70$  indicates acceptable internal consistency (Taber, 2017). Learning improvement was measured using the Normalized Gain (N-Gain):

$$N\text{-Gain} = (\text{Posttest} - \text{Pretest}) / (100 - \text{Pretest}) \quad (3)$$

Score interpretation criteria are presented in Table 2a below (Hake, 1999):

<b>N-Gain Score</b>	<b>Category</b>
$g > 0.7$	High
$0.3 \leq g \leq 0.7$	Moderate
$g < 0.3$	Low

Statistical significance of pre–post differences was tested with the paired sample t-test:

$$t = \bar{D} / (SD / \sqrt{n}) \quad (4)$$

where  $\bar{D}$  = mean difference of paired scores,  $SD$  = standard deviation of differences, and  $n$  = number of participants. Significance was set at  $p < 0.05$ . Effect size was calculated using Cohen's  $d$ :

$$d = (M_{\text{post}} - M_{\text{pre}}) / SD_{\text{pooled}} \quad (5)$$

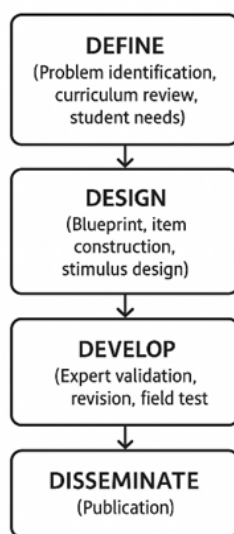
where  $SD_{\text{pooled}} = \sqrt{[(S_{\text{pre}}^2 + S_{\text{post}}^2) / 2]}$ . Effect size interpretation: small (0.2), medium (0.5), large ( $\geq 0.8$ ) (Cohen, 1988).

## RESULT AND DISCUSSION

The development of the HOTS assessment on the topic of plant and animal reproduction was carried out through the Four-D (4D) model. Each stage produced significant findings contributing to the construction of an instrument capable of measuring analytical, evaluative, and creative thinking skills. The Define stage revealed fundamental problems in current assessment practices at the junior high school level. Document analysis of teacher-constructed exam questions and direct classroom observation showed that assessment instruments were still dominated by Lower Order Thinking Skills (LOTS), with approximately 72% of existing questions classified at the C1 level (remembering) and the remainder at the C2 level (understanding). Specific examples of these LOTS questions included tasks that only required students to identify parts of a reproductive organ or define biological terms such as "pollination" without contextual application. No questions were found addressing C4 (analysis), C5 (evaluation), or C6 (creation) levels. These findings illustrate that higher-order competencies are rarely assessed, resulting in students' inability to

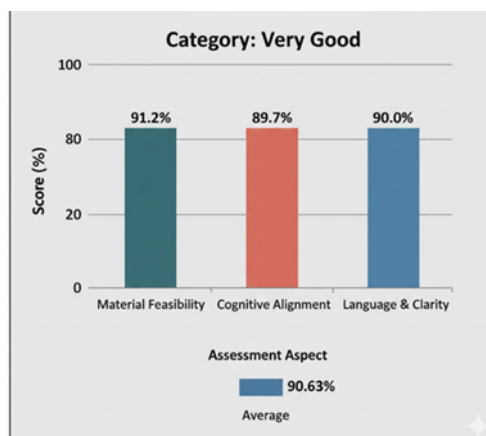
apply, analyze, and evaluate biological phenomena (Musliha et al., 2021; Vrbanović Lisac et al., 2025). These findings confirm that when assessment instruments are limited to lower-order cognitive tasks, students are deprived of meaningful opportunities to practice analysis, evaluation, and creation—the very skills targeted by 21st-century learning frameworks (Khoy, 2025; Fuadiyah et al., 2023).

During the Design stage, 30 HOTS-oriented multiple-choice items were constructed to measure cognitive processes at the C4–C6 levels. Each item was equipped with contextual stimuli, such as diagrams of reproductive organs, graphical data of plant growth, comparative tables of asexual and sexual reproduction, and situational problems related to environmental influences on reproduction. Items followed a stimulus-first approach to ensure that students relied on reasoning rather than memorization. This design approach is in line with Crowe et al. (2019) and Kibble (2017), who highlight that contextual stimuli promote deeper cognitive engagement and improve reasoning accuracy. A complete test blueprint was also formulated to ensure alignment between learning objectives, indicators, material scope, and HOTS cognitive levels.



**Figure 1.** The 4D Development Model

The Develop stage produced empirical evidence of the instrument's quality. Expert validation involved two subject-matter lecturers, one assessment specialist, and two experienced science teachers, yielding an overall average rating of 4.65/5.00 (90.63%), categorized as Very Good. The three key aspects assessed were Material Feasibility (4.70; 91.2%), Cognitive Alignment (4.70; 91.2%), and Language & Clarity (4.63; 90.0%), all falling into the Very Good category.



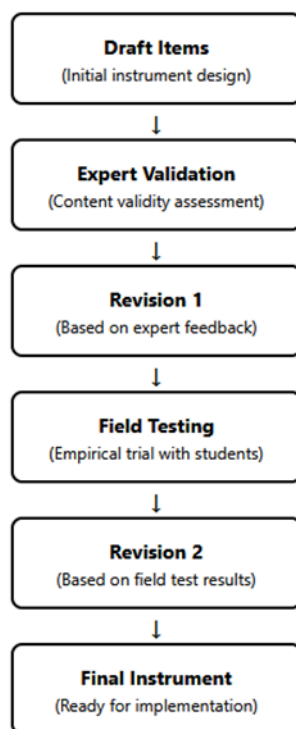
**Figure 2.** Instrument Assessment Scores by Aspect

Furthermore, Aiken's  $V$  coefficient reached 0.875, indicating high agreement among validators regarding content relevance, clarity, cognitive alignment, and feasibility. These findings support Han and Xiang (2025), who emphasize that expert validation plays a critical role in ensuring the appropriateness and accuracy of HOTS assessment items. The empirical test using 35 students provided further evidence of the instrument's robustness. Of 30 items, 24 met the validity requirements. The 6 invalid items were attributed to distractors that were too ambiguous and stimuli that were overly complex for students' current cognitive stage, serving as a reference for future developers to balance scientific depth with language readability.

**Table 2.** Item Analysis Summary

Indicator	Result	Category
Valid Items	24 out of 30	Valid
Reliability (Cronbach's $\alpha$ )	0.868	Very High
Difficulty Level	83% Moderate	Good
Discrimination Power	Good–Very Good	Accepted
Invalid Items (reasons)	6 items	Ambiguous distractors, complex stimuli

The reliability coefficient ( $\alpha = 0.868$ ) indicates excellent internal consistency, confirming stable and dependable measurements. This is consistent with Simonsmeier et al. (2025) and Zakariya (2022), who found that HOTS-based assessments produce high reliability when items include rich contexts and strong stimulus integration.



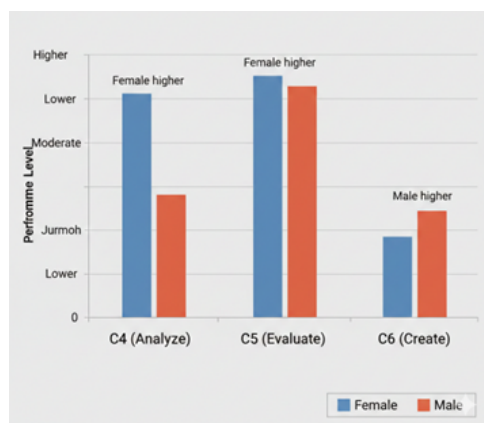
**Figure 3.** Validation–Revision Process

The results reveal distinct cognitive patterns across gender groups. Female students demonstrated higher performance in analysis (C4: 76.8%) and evaluation (C5: 78.2%), while male students performed slightly better in the creation level (C6: 71.5%), which also showed the lowest overall performance among the three HOTS components.

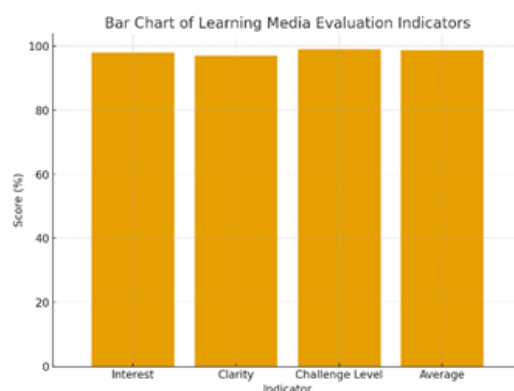
**Table 3.** T-Test Results Between Male and Female Students

Group	Mean Score	Sig. (p)	Interpretation
Female	81.5	0.000	Significant Difference
Male	73.2	0.000	Significant Difference

These findings ( $p < 0.001$ ) indicate that the instrument successfully differentiates cognitive abilities across demographic groups. This aligns with Kuncoro et al. (2022) and Sepriyanti et al. (2022), who similarly reported gender-based variations in HOTS performance and analytical ability among science learners across different gender groups.

**Figure 4.** HOTS Performance Pattern

Student responses toward the instrument indicated an overwhelmingly positive reception, with an average score of 98.70%. Based on an interview conducted with one of the students, it was stated: "The questions were challenging because we had to analyze the data provided in the pictures first, rather than just choosing an answer from memory." This qualitative feedback confirms that the stimulus-first approach effectively promoted analytical reasoning. Learners described the assessment as both challenging and engaging, requiring them to interpret data, compare biological processes, and draw evidence-based conclusions. These results are consistent with Ritchie et al. (2019) and Baloyi (2023), who reported that HOTS-oriented assessments enhance student motivation and strengthen conceptual understanding. Teachers likewise expressed strong approval, noting that the assessment tool was appropriate, practical, and well aligned with the objectives of the science curriculum. They particularly appreciated the clarity of the stimuli, the structure of the questions, and the integration of 21st-century learning skills. This aligns with Nkhoma et al. (2017), who indicated that HOTS-based assessments improve teachers' capacity to evaluate students' reasoning skills and promote inquiry-based learning environments.

**Figure 5.** Bar Chart of Learning Media Evaluation Indicators

As illustrated in Figure 5, all learning media evaluation indicators—feasibility, clarity, functionality, and alignment with HOTS objectives—received consistently high ratings from both students and teachers, demonstrating strong acceptance and practical applicability of the developed assessment.

**Table 4.** Effectiveness Analysis Results

Measure	Value	Category	Reference
N-Gain	0.52	Moderate	Hake (1999)
Paired t-test (sig.)	$p < 0.001$	Significant	Cohen (1988)
Cohen's d	0.74	Medium–Large	Cohen (1988)

Overall, the results demonstrate that the HOTS assessment developed through the 4D model successfully yields a high-quality measurement tool that is valid, reliable, and effective in capturing higher-order cognitive skills within the context of plant and animal reproduction.

## CONCLUSION

The development of the HOTS-oriented biology assessment instrument using the 4D model (Define, Design, Develop, and Disseminate) has successfully addressed the critical need for higher-order thinking evaluation in junior high schools. The study concludes that the developed instrument is highly feasible, with an expert validation average score of 4.65/5.00 (Very Good) and an Aiken's V coefficient of 0.875. Empirical testing further confirmed its quality, yielding a high reliability coefficient of 0.868, with 24 out of 30 items meeting the validity requirements. Analysis of the invalid items revealed that stimulus complexity and distractor ambiguity were the primary factors, providing valuable insights for future test construction. Furthermore, student and teacher feedback indicated that the use of contextual stimuli effectively promotes deeper cognitive engagement and reasoning skills. This instrument not only serves as a valid tool for measuring student performance but also encourages the transition from factual recall to evidence-based analysis in science education. Future implementation should focus on broader dissemination to various educational contexts to further validate its effectiveness in improving 21st-century learning outcomes.

## REFERENCES

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Arendse, D. E. (2021). Addressing gender discrimination in cognitive assessment using the English Comprehension Test. *SA Journal of Industrial Psychology*, 47(1), 1-10.
- Baloyi, O. B. (2023). Development of higher-order thinking skills in nursing students through online problem-based assessment. *Health SA Gesondheid*, 28, Article 2423. <https://doi.org/10.4102/hsag.v28i0.2423>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>

- Crowe, A., Dirks, C., & Wenderoth, M. P. (2019). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, 7(4), 368–381. <https://doi.org/10.1187/cbe.08-05-0024>
- Díaz, M. M. M. B., Lim, J. R., Iborra, C. P., López, E., Rodríguez, H., López, R., ... & Groot, B. (2022). The power of curriculum to transform education: How education systems incorporate 21st century skills to prepare students for today's challenges.
- Fuadiyah, S., Selaras, G. H., Melta, D., & Rahmi, D. (2023). The urgency of Higher Order Thinking Skills (HOTS) base on development of biology assessment instruments for class XI students of SMA/MA. In *Proceedings of the 3rd International Conference on Biology, Science and Education (IcoBioSE 2021)* (pp. 3–8). Atlantis Press. [https://doi.org/10.2991/978-94-6463-166-1\\_2](https://doi.org/10.2991/978-94-6463-166-1_2)
- Hake, R. R. (1999). Analyzing change/gain scores. American Educational Research Association's Division D, Measurement and Research Methodology.
- Han, C., & Xiang, J. (2025). Alignment analysis between China college entrance examination physics test and curriculum standard based on E-SEC model. *International Journal of Science and Mathematics Education*, 23(1), 215–234. <https://doi.org/10.1007/s10763-024-10468-0>
- Jarupunphol, P., Flek, L., & Hui, S. C. (2025). Applying Cronbach's alpha to ensure reliable online testing in e-learning environments. In R. Silhavy & P. Silhavy (Eds.), *Artificial intelligence and system engineering. CoMeSySo 2024. Lecture Notes in Networks and Systems* (Vol. 1490, pp. 91–100). Springer. [https://doi.org/10.1007/978-3-031-96759-7\\_8](https://doi.org/10.1007/978-3-031-96759-7_8)
- Khoy, B. (2025). Unlocking cognitive learning objectives: A comprehensive evaluation of how textbooks and syllabi align with revised Bloom's taxonomy across disciplines. *Current Perspectives*, 45, 189–202. <https://doi.org/10.1007/s41297-024-00295-2>
- Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, 41(1), 110–119. <https://doi.org/10.1152/advan.00116.2016>
- Li, D., Meng, L., & Fan, X. (2024). Development and validation of a higher-order thinking skills scale for major students in the interior design discipline for blended learning. *BMC Medical Education*, 24, Article 287. <https://doi.org/10.1186/s12909-024-05268-5>
- Musliha, S., Ikhsan, M., & Fajrina, S. (2021). The analysis of higher order thinking skills (HOTS) in the test questions constructed by English teachers. *English Language Teaching Educational Journal*, 4(2), 118–127. <https://doi.org/10.12928/eltej.v4i2.3250>
- Nkhoma, M. Z., Lam, T. K., Richardson, J., Kam, B., & Lau, K. H. (2017). Developing case-based learning activities based on the revised Bloom's taxonomy. *Informing Science*, 20, 85–105.
- Ravand, H., Shahi, R., Effatpanah, F., & Moghadamzadeh, A. (2025). Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's Taxonomy. *Frontiers in Education*, 10, Article 1644811.
- Razmjoo, S. A., & Kazempourfard, E. (2012). On the representation of Bloom's revised taxonomy in Interchange coursebooks. *The Journal of Teaching Language Skills*, 4(1), 171–204.
- Ritchie, K. L., Gingerich, A. C., Barr, N., Bateman, W., Boyce, S., Desy, J., Fuller, K., Hecker, K., Cameron, S., MacNeil, J., Oksanen, K., Philpott, B., & Wright, B. (2019). Incorporating higher order and deep learning in a large, lecture-based human physiology course. *Advances in Physiology Education*, 43(3), 359–369. <https://doi.org/10.1152/advan.00126.2019>

- Sepriyanti, N., Nelwati, S., Kustati, M., & Afriadi, J. (2022). The effect of 21st-century learning on higher-order thinking skills (HOTS) and numerical literacy of science students in Indonesia based on gender. *Jurnal Pendidikan IPA Indonesia*, 11(2), 314–321. <https://doi.org/10.15294/jpii.v11i2.36384>
- Simonsmeier, B. A., Peiffer, H., Flaig, M., & Schneider, M. (2025). The Cronbach's alpha of domain-specific knowledge tests before and after learning: A meta-analysis. *Educational Psychology Review*, 37, 1–43. <https://doi.org/10.1007/s10648-024-09982-y>
- Taber, K. S. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children*. University of Minnesota.
- Vineis, P., Avendano-Pabon, M., Barros, H., Bartley, M., Carmeli, C., Carra, L., ... & Zins, M. (2020). Special report: the biology of inequalities in health: the lifepath consortium. *Frontiers in public health*, 8, 495953.
- Vrbanović Lisac, D., Radanović, I., & Šimić Šašić, S. (2025). Enhancing higher order thinking skills to improve understanding of biological content: A quasi-experimental study. *Natural Sciences Education*, 54, e70023. <https://doi.org/10.1002/nse2.70023>
- Zakariya, Y. F. (2022). Improving students' mathematics self-efficacy: A systematic review of intervention studies. *Frontiers in Education*, 7, Article 986622. <https://doi.org/10.3389/educ.2022.986622>