ISSN : 2549-2764 (ONLINE)
ISSN : 2337-604X (PRINT)

BERKALA ILMIAH PENDIDIKAN FISIKA

https://ppjp.ulm.ac.id/journal/index.php/bipf

Development of *Higher Order Thinking Skills* (HOTS) Test Instruments on Parabolic and Circular Motion Materials in High Schools

Muhammad Aizri Fadillah* and Sahyar
Physics Education, Medan State University, Medan, Indonesia
*m.aizrifadillah@gmail.com

DOI:10.20527/bipf.v11i3.16697

Received: 1 July 2023 Accepted: 5 November 2023 Published: 28 December 2023

Abstract

This research intends to produce HOTS test instruments on high school parabolic and circular motion materials. The Borg and Gall model of research and development (R&D) includes six stages: (1) preliminary investigation, (2) planning and equipment preparation, (3) expert initial product testing and revision, (4) limited field testing, (5) revision based on the results field testing, and (6) broader field test. The data analysis used was qualitative and quantitative. The HOTS test instrument consists of 20 multiple-choice questions, including ten on parabolic and circular motion. According to three experts from material, construction, and language, the outcomes of HOTS instrument validation showed valid and feasible to use with an average value of 0.95 (high validity). The characteristics of HOTS questions have met the criteria of validity as many as 18 questions, high reliability with a value of 0.71, nine questions with a medium difficulty level, an average differentiating power of 0.41 with a good category, and 17 questions have the effectiveness of effective deception. This research concluded that HOTS-based parabolic and circular motion instruments are feasible. Thus, the instrument can be used as training material and assessment of students' HOTS in high schools.

Keywords: Development; HOTS; Instrument; Parabolic and circular motion

How to cite: Fadillah, M. A., & Sahyar, S. (2023) Development of *higher order thinking skills* (HOTS) test instruments on parabolic and circular motion materials in high schools. *Berkala Ilmiah Pendidikan Fisika*, 11(3), 329-338.

INTRODUCTION

The education world faces increasingly complex 21st-century challenges, where all activities are integrated with massive use of technology. Education needs to prepare students to learn, master, and develop essential 21st-century skills (Hujjatusnaini et al., 2022; Sofyan, 2019; Varghese & Musthafa, 2021). Partnership for 21st-century skills formulated a learning framework to describe the skills learners need to face 21st-century challenges, including critical thinking, creative thinking, problem-solving, and decision-making. These four abilities are referred to as Higher-Order Thinking Skills (HOTS) (Afandi et al., 2019; Borrowski, 2019; González-Pérez & Ramírez-Montoya, 2022; Ramírez-Montoya et al., 2022).

HOTS differs from Higher-Order Thinking (HOT) (Sani, 2019). Figure 1 shows that HOT is related to cognitive abilities in analysis, evaluation, and creation. Meanwhile, HOTS relates to thinking critically and creatively, solving problems, and making decisions. In general, HOTS includes HOT. For example, to solve problems, students must be able to analyze and evaluate.

Learners must be able to reason, consider, analyze, think of alternatives, and evaluate to make decisions or think critically. Likewise, to produce new/developed concepts, ideas, or products, learners must be able to think creatively (Sani, 2019).

In essence, every learner must be equipped with high-level thinking skills to become a more independent, critical, and productive individual. Learners who have HOTS will not easily accept information without strong reasons or evidence, accept differences, are independent in thinking and acting, and are not easily influenced by various factors so that they can work and benefit the community. Therefore, HOTS is needed in education, especially in physics learning.

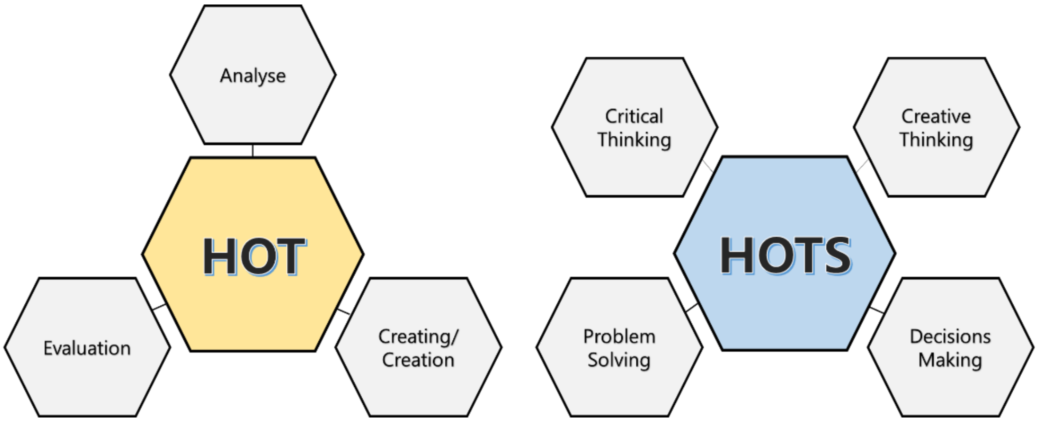

Figure 1 Differences between HOT and HOTS (Sani, 2019)

Physics is a science that studies, understands, and analyzes nature systematically. In physics learning, students are required to have low-level thinking skills (LOTS) and up to HOTS (Asysyifa et al., 2019; Santyasa et al., 2020; Yusuf & Widyaningsih, 2019). Through physics learning, students' higher-order thinking skills can be formed through assessment with various techniques and instruments. However, test instruments specifically designed based on HOTS are still lacking in schools (Damanik & Irfandi, 2022; Hanafi et al., 2022; Hikmah & Amin, 2019; Jayanti, 2020).

The limited number of high-level thinking skills results in students being less trained to deal with various physics problems that require high-level thinking skills, one of which is Parabolic and circular motion material. Parabolic and circular motion are common everyday physics phenomena that require a high level of skill due to their complexity. Understanding gravity and Newton's laws of motion is essential for parabolic motion (Jörges et al., 2021). In contrast, centripetal force, centripetal acceleration, and angular velocity principles are critical for circular motion (Luchembe & Shumba, 2019). In addition, students often face challenges when working with these two materials because they require complex mathematical calculations, such as trigonometry and vectors, for motion analysis and modeling. Designing appropriate problem-solving approaches also requires critical and creative thinking, problem-solving, and decision-making skills. Therefore, developing a feasible HOTS-based testing instrument is imperative to assess students' proficiency in these materials.

METHOD

This research is an R&D utilizing the Borg & Gall expansion framework, converted into six exploration stages according to the objectives and interests of the research. A stage consists of (1) preliminary investigation, (2) planning and equipment preparation, (3) expert initial product testing and revision, (4) limited field testing, (5) revision based on the results of field testing, and (6) broader field test (Rahielanamy & Kamaludin, 2022).

Instruments are prepared through the stages of grid preparation, determination of stimuli, preparation of test items, determination of keys, and answer scores. The instrument tool is a multiple choice (binary 0 and 1) with five answer options and pertains to HOTS indicators (Setiawan et al., 2021). The test instrument consisted of 20 items with a division of 10 parabolic motions and 10 items of circular motion. The test instrument is structured based on the stimulus: (1) Contextual, indicating actual situations in day-to-day existence so learners can employ educational ideas in the classroom to resolve the challenges provided. (2) Argument, meaning that it can invite students to identify, evaluate, and build new ways of solving problems based on a case. (3) Reasoning, meaning that the test instrument can invite students to conclude from statements by examining the relationship between statements or data.

The suitability of the HOTS instrument tool on parabolic and circular motion content is established from the initial validation conducted by three instrument experts, which was subsequently modified during the first phase. The modified version was then trialed on a small scale at SMA Negeri 13 Medan class XI with a sample size of 10 students, and further revisions were made during the second phase. The revised version from the second phase was then tested on a larger scale at SMA Negeri 13 Medan class XI with a sample size of 60 students.

Data analysis was performed qualitatively and quantitatively, and the data were processed utilizing Microsoft Exel software. A qualitative examination was carried out by three instrument experts, who assessed the content validation sheets from various perspectives, including material, structural, and linguistic aspects. The content validity index was calculated using Aiken's formula V (Desilva et al., 2020). Additionally, a quantitative analysis was carried out to ascertain the characteristics of the HOTS question based on the student's responses, which included assessing the validity, reliability, difficulty, differentiating power, and effectiveness of deception.

RESULT AND DISCUSSION

The preliminary assessment was conducted by giving validation documents to instrument experts, then assessed using Aiken's index of validity formula to determine the credibility of the assets. The results of the instrument experts are shown in Table 1, with an average score of 0.95 (high validity), and all items were deemed suitable for use. The evaluation of content validity is conducted using the criteria $0 \leq V < 0.4$ (indicating low validity), $0.4 \leq V \leq 0.8$ (indicating medium validity), and $0.8 \leq V \leq 1.0$ (indicating high validity) (Astuti & Retnawati, 2018; Retnawati, 2016).

Table 1 Expert validation results

Question Number	Material	V Aiken	Description
1	Parabolic motion	0.95	Worth using
2	Parabolic motion	0.91	Worth using
3	Parabolic motion	0.96	Worth using
4	Parabolic motion	0.91	Worth using
5	Parabolic motion	0.91	Worth using

Question Number	Material	V Aiken	Description
6	Parabolic motion	0.95	Worth using
7	Parabolic motion	0.95	Worth using
8	Parabolic motion	0.97	Worth using
9	Parabolic motion	0.97	Worth using
10	Parabolic motion	0.98	Worth using
11	Circular motion	0.96	Worth using
12	Circular motion	0.98	Worth using
13	Circular motion	0.94	Worth using
14	Circular motion	0.97	Worth using
15	Circular motion	0.96	Worth using
16	Circular motion	0.97	Worth using
17	Circular motion	0.94	Worth using
18	Circular motion	0.97	Worth using
19	Circular motion	0.95	Worth using
20	Circular motion	0.98	Worth using
Average		0.95	Worth using

The following research was carried out at SMA Negeri 13 Medan class XI to ascertain the characteristics of HOTS questions. The validity of HOTS questions is carried out to determine whether the questions follow the rules measured. The calculated validity value is obtained using the rough product moment correlation formula. The question is declared valid if Rcount > Rtable and invalid if Rcount > Rtable with a significant level of Rcount 5% (Joko, 2018; Putri et al., 2020; Supena et al., 2021). The validity results of the questions are displayed in Table 2. The validity of HOTS questions obtained 18 valid questions with N = 60 and a Rtable of 0.2108. It can be inferred that the HOTS questions concerning parabolic and circular motion materials are valid, as the proportion of valid items exceeds 50% (Setiawan et al., 2021). A valid test item means that the item has good construction and covers the entire material to be measured (Adom et al., 2020; Nurhayati et al., 2020).

Table 2 Validity test results

R_{count}	Question Number	Total	Criteria	Percentage
> 0.2108	1,2,3,4,6,7,8,9,10,11,13,14,15,16,17,18,19,20	18	Valid	90%
≤ 0.2108	5,12	2	Invalid	10%

The reliability of HOTS questions utilized the Kuder Richardson (KR-20) formula with the following standards: $0.91 \leq R \leq 1.00$ (very high), $0.71 \leq R \leq 0.90$ (high), $0.41 \leq R \leq 0.70$ (moderate), $0.21 \leq R \leq 0.40$ (low), and $0.00 \leq R \leq 0.20$ (very low). The results of the HOTS question reliability test are displayed in Table 3. The reliability value of the question is 0.71, and it has a high level of reliability. This demonstrates that HOTS questions can be utilized in gathering research information (Cahyono & Darsini, 2022; Sürücü & Maslakçi, 2020).

The difficulty was analyzed on the test instrument to identify the difficult, medium, and easy questions (Desstya et al., 2019; Harjo et al., 2019; Sa'idah et al., 2018). The difficulty is calculated using the formula P = B/J with the criteria: $0.00 \leq P \leq 0.30$ (difficult), $0.31 \leq P \leq 0.70$ (medium), and $0.71 \leq P \leq 1.00$ (easy) (Arikunto, 2014). The calculation results obtained 11 difficult items, nine medium items, and no easy category items, as displayed in Table 4.

Table 3 Reliability test results

Number of Items	Reliability Value	Description
20	0.71	High reliability

Table 4 Difficulty test results

Criteria	Question Number	Total	Percentage
Difficult	3,5,8,9,10,11,12,13,17,18,19	11	55%
Medium	1,2,4,6,7,14,15,16,20	9	45%
Easy	-	-	-

The differentiating power test is conducted to ascertain the ability of test items to distinguish test takers who understand the material from test takers who do not understand the material (Blegur et al., 2023; Erfan et al., 2020). The differentiating power index is calculated using the formula Dp=(BA/JA)-(BB/JB) with the criteria: $0.71 \leq Dp \leq 1.00$ (very good), $0.41 \leq Dp \leq 0.70$ (good), $0.21 \leq Dp \leq 0.40$ (simply), and $0.00 \leq Dp \leq 0.20$ (not good) (Joko, 2018). The calculation results obtained 1 item in the very good category, 9 in the good category, 8 in the simple category, and 2 in the not good category. Table 5 displays the results of the differentiating power test.

Table 5 Differentiating power test results

Criteria	Question Number	Total	Percentage
Very good	1	1	5%
Good	2,6,7,11,14,15,16,17,20	9	45%
Simple	4,5,8,9,10,13,18,19	8	40%
Not good	3,12	2	10%

Deception effectiveness is a critical component affecting a test item's overall quality. Excerpts' effectiveness is designed to contain plausible but incorrect answers based on the supposed concept to determine the test takers' material mastery level (Gierl et al., 2017; Shin et al., 2019). The index of the effectiveness of deception is calculated using the formula IP = P / N x 100% and functions if it is selected by at least 5% of participants. An exemption is effective if 3-5 exemptions function and ineffective if 1-2 exemptions function (Kadir, 2015; Tanjung & Bakar, 2019). The calculation results obtained from 17 questions have effective deception effectiveness, as shown in Table 6. Ambiguous sentences can cause HOTS questions to be ineffective, so the deception cannot function properly (Scully, 2017).

Table 6 Results of the deception effectiveness test

Functioning Deceptions	Criteria	Question Number	Total	Percentage
3-5	Effective	3,4,5,7,8,9,10,11,12,13,14,15,16, 17,18,19,20	17	85%
1-2	Ineffective	1,2,6	3	15%

Table 7 summarizes the test results conducted to determine the characteristics of HOTS questions. Based on the test, it was found that two items were invalid but had a reliability level in the moderate category. This finding differs from (Aisah & Pahlevi, 2020) (Setiawan et al., 2021), who found that the more valid items, the higher the reliability level. Despite high validity, the primary cause of low reliability is the considerable variation in student responses to items. Diverse student responses to the same item can significantly decrease answers' consistency, ultimately leading to low reliability (Tanjung & Dwiana, 2019).

Table 7 Recapitulation of test results

Validity		**Reliability**	**Level of Difficulty**			**Differentiating Power**				**Deception Effectiveness**	
V	**IV**		**M**	**D**	**E**	**VG**	**G**	**S**	**NG**	**E**	**IE**
18	2	High	9	11	0	1	9	8	2	17	3

A relationship was also found between the deception's effectiveness, difficulty level, and differentiating power. The test results found that the effectiveness of effective deception has a high level of difficulty and poor differentiating power. This finding differs from (Akhmadi, 2021), where the effectiveness of good deception means that the index of differentiating power and difficulty level is also good. If the effectiveness of deception is effective, the differentiating power is good, and the difficulty level is not too difficult or too easy.

The efficacy of a sample in an examination is typically assessed on two key grounds: difficulty and discriminatory ability (Akhmadi, 2021). Effective prompts are usually intricate and perplexing, intentionally formulated to thoroughly test exam-takers' comprehension. This can inspire examinees to cogitate critically and apply their knowledge more seriously. Nevertheless, if most examinees answer accurately, the item's discriminatory ability is inadequate. This phenomenon may result from some test takers guessing correctly or possessing sufficient knowledge to surmount the test's high difficulty level. Consequently, a question that is effective regarding difficulty may still have limited discriminatory power as it fails to distinguish between test takers with profound comprehension and those without. Therefore, a skilled test-taker must strike a balance between difficulty and discriminative power to provide an exact assessment of their understanding of the tested material.

In addition, the relationship between difficulty level and differentiating power was also found, where the greater the difficulty level, the lower the differentiating power. Tables 4 and 5 show that 11 items have a high level of difficulty, and 9 of them have a differentiating power in the categories of simple and not good. This aligns with (Fatimah & Alfath, 2019), who stated that the higher the difficulty level, the lower the differentiating power.

The test's differentiating power measures the extent to which a test can separate participants with different abilities or understandings (Erfan et al., 2020). When a test is highly challenging, participant separation ability is restricted due to low scores amongst almost all participants. Consequently, the higher the difficulty level, the lower the test's differentiating power. Ideally, a test should balance difficulty and the ability to differentiate between high and low performers to provide valuable insights into participants' grasp of the tested material.

HOTS questions require mastery of concepts in solving them, so test takers who need help understanding the concept of material will have difficulty working on the problems given and tend to guess. The need for more understanding of material concepts in test takers results in varied test results. For example, items considered not too difficult or easy become items with difficult categories. The problems experienced can be caused because the material on the test instrument has passed, so test takers have difficulty working on it. This aligns with research conducted by (Tanjung & Dwiana, 2019) that the instrument being tested becomes less effective if there is an interval between the material studied by the test takers and the material on the instrument.

CONCLUSION

This research produced a HOTS test instrument totaling 20 items with a division of 10 items of parabolic motion and ten items of circular motion. The test instrument is structured based on context, argument, and reasoning. According to three experts from material, construction, and language, the outcomes of HOTS instrument validation showed valid and feasible to use with an average value of 0.95 (high validity). The characteristics of HOTS questions obtained are 18 questions of valid and high reliability, 0,71, 9 questions in the medium category, the average test item differentiator is 0.41 in the good category, and 17 questions have the effectiveness of effective deception.

REFERENCES

Adom, D., Adu-Mensah, J., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(1), 109. https://doi.org/10.11591/ijere.v9i1.20457

Afandi, A., Sajidan, S., Akhyar, M., & Suryani, N. (2019). Development frameworks of the indonesian partnership 21st-century skills standards for prospective science teachers: A delphi study. *Jurnal Pendidikan IPA Indonesia*, *8*(1). https://doi.org/10.15294/jpii.v8i1.11647

Aisah, S., & Pahlevi, T. (2020). Pengembangan instrumen penilaian higher order thinking skills (hots) pada mata pelajaran korespondensi kelas x otp di smk negeri 1 jombang. *Jurnal Pendidikan Administrasi Perkantoran (JPAP)*, *8*(1), 146–156.

Akhmadi, M. N. (2021). Analisis butir soal evaluasi tema 1 kelas 4 sdn plumbungan menggunakan program anates. *Ed-Humanistics: Jurnal Ilmu Pendidikan*, *6*(1), 799–806.

Arikunto, S. (2014). Dasar-dasarevaluasi pendidikan. Jakarta: PT Bumi Aksara. *Prosedur Penelitian Suatu Pendekatan Praktik, Jakarta: Renika Cipta*.

Astuti, F., & Retnawati, H. (2018). Students' school-based assessment scores relationship with students' readiness, motivation, and national examination scores of mathematics subject. *JPP (Jurnal Pendidikan Dan Pembelajaran)*, *24*(2), 64–72. https://doi.org/10.17977/um047v24i22017p064

Asysyifa, D. S., Jumadi, Wilujeng, İ., & Kuswanto, H. (2019). Analysis of students critical thinking skills using partial credit models (pcm) in physics learning. *International Journal of Educational Research Review*, *4*(2), 245–253. https://doi.org/10.24331/ijere.518068

Blegur, J., Rajagukguk, C. P. M., Sjioen, A. E., & Souisa, M. (2023). Innovation of analytical thinking skills instrument for throwing and catching game activities for elementary school students. *International Journal of Instruction*, *16*(1).

Borrowski, T. (2019). The battelle for kids p21 framework for 21st century learning. *University of Illinois at Chicago*.

Cahyono, E. A., & Darsini, D. (2022). Sikap masyarakat terhadap program vaksinasi booster covid-19 (validitas dan reliabilitas kuesioner penelitian). *Pengembangan Ilmu Dan Praktik Kesehatan*, *1*(1), 1–21.

Damanik, F. C., & Irfandi, I. (2022). Pengembangan instrumen tes berbasis higher order thinking skill (hots) bentuk pilihan ganda berdasarkan taksonomi bloom revisi pada materi gelombang bunyi. *Inpafi (Inovasi Pembelajaran Fisika)*, *10*(1).

Desilva, D., Sakti, I., & Medriati, R. (2020). Pengembangan instrumen penilaian hasil belajar fisika berorientasi hots (higher order thinking skills) pada materi elastisitas dan hukum hooke. *Jurnal Kumparan Fisika*, *3*(1 April), 41–50.

Desstya, A., Prasetyo, Z. K., Suyanta, S., Susila, I., & Irwanto, I. (2019). Developing an instrument to detect science misconception of an elementary school teacher. *International Journal of Instruction*, *12*(3), 201–218. https://doi.org/10.29333/iji.2019.12313a

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model rasch. *Indonesian Journal Of Educational Research and Review*, 11–19.

Fatimah, L. U., & Alfath, K. (2019). Analisis kesukaran soal, daya pembeda dan fungsi distraktor. *AL-MANAR: Jurnal Komunikasi Dan Pendidikan Islam*, *8*(2), 37–64.

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

González-Pérez, L. I., & Ramírez-Montoya, M. S. (2022). Components of education 4.0 in 21st century skills frameworks: systematic review. *Sustainability*, *14*(3), 1493. https://doi.org/10.3390/su14031493

Hanafi, M., Syamsuri, S., & Mutaqin, A. (2022). Pengembangan instrumen soal higher order thinking skills (hots) matematika berdasarkan brookhart konteks motif batik pandegelang pada siswa mts. *Media Pendidikan Matematika*, *10*(1), 43–59.

Harjo, B., Kartowagiran, B., & Mahmudi, A. (2019). Development of critical thinking skill instruments on mathematical learning high school. *International Journal of Instruction*, *12*(4), 149–166. https://doi.org/10.29333/iji.2019.12410a

Hikmah, H., & Amin, N. (2019). Pengembangan instrumen untuk mengukur kemampuan berpikir tingkat tinggi dalam mata pelajaran matematika di sma kabupaten majene. *SAINTIFIK*, *5*(1), 1–7.

Hujjatusnaini, N., Corebima, A. D., Prawiro, S. R., & Gofur, A. (2022). The effect of blended project-based learning integrated with 21st-century skills on pre-service biology teachers' higher-order thinking skills. *Jurnal Pendidikan IPA Indonesia*, *11*(1), 104–118.

https://doi.org/10.15294/jpii.v11i1.27148

Jayanti, E. (2020). Instrumen tes higher order thinking skill pada materi kimia SMA. *Orbital: Jurnal Pendidikan Kimia*, *4*(2), 135–149.

Joko, W. (2018). *Evaluasi Pembelajaran (Sesuai dengan Kurikulum 2013)*. UNIPMA Press.

Jörges, B., La Scaleia, B., López-Moliner, J., Lacquaniti, F., & Zago, M. (2021). Perceptual judgments of duration of parabolic motions. *Scientific Reports*, *11*(1), 7108. https://doi.org/10.1038/s41598-021-86428-3

Kadir, A. (2015). Menyusun dan menganalisis tes hasil belajar. *Al-TA'DIB: Jurnal Kajian Ilmu Kependidikan*, *8*(2), 70–81.

Luchembe, D., & Shumba, O. (2019). Practical work, simulations and feedback to address undergraduate physics students' challenges in understanding circular and rotational motion. *Education Quarterly Reviews*, *2*(3).

Nurhayati, Y., Rubini, B., & Pursitasari, I. D. (2020). Pengembangan asesmen autentik pada materi sistem pernapasan. *JSEP (Journal of Science Education and Practice)*, *3*(2), 56–69.

Putri, H. E., Wahyudy, M. A., Yuliyanto, A., & Nuraeni, F. (2020). Development of instruments to measure mathematical anxiety of elementary school students. *International Journal of Learning, Teaching and Educational Research*, *19*(6), 282–302. https://doi.org/10.26803/ijlter.19.6.17

Rahielanamy, R., & Kamaludin, A. (2022). Development of higher order thinking skill qualified test on colligative properties of solutions. *Jurnal Pendidikan Sains Indonesia*, *10*(2), 294–311.

Ramírez-Montoya, M. S., Castillo-Martínez, I. M., Sanabria-Z, J., & Miranda, J. (2022). Complex thinking in the framework of education 4.0 and open innovation—a systematic literature review. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(1), 4. https://doi.org/10.3390/joitmc8010004

Retnawati, H. (2016). Validitas reliabilitas dan karakteristik butir. *Yogyakarta: Parama Publishing*.

Sa'idah, N., Yulistianti, H. D., & Megawati, E. (2018). Analisis instrumen tes higher order thinking matematika smp. *Jurnal Pendidikan Matematika*, *13*(1), 41–54. https://doi.org/10.22342/jpm.13.1.6619.41-54

Sani, R. A. (2019). *Pembelajaran berbasis hots edisi revisi: higher order thinking skills* (Vol. 1). Tira Smart.

Santyasa, I. W., Rapi, N. K., & Sara, I. W. W. (2020). Project based learning and academic procrastination of students in learning physics. *International Journal of Instruction*, *13*(1), 489–508. https://doi.org/10.29333/iji.2020.13132a

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 4.

Setiawan, J., Sudrajat, A., & Kumalasari, D. (2021). Development of higher order thinking skill assessment instruments in learning indonesian history. *International Journal of Evaluation and Research in Education*, *10*(2), 545–552.

Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, *10*, 825.

Sofyan, F. A. (2019). Implementasi HOTS pada kurikulum 2013. *INVENTA: Jurnal Pendidikan Guru Sekolah Dasar*, *3*(1), 1–9.

Supena, I., Darmuki, A., & Hariyadi, A. (2021). The influence of 4c (constructive, critical, creativity, collaborative) learning model on students' learning outcomes. *International Journal of Instruction*, *14*(3), 873–892. https://doi.org/10.29333/iji.2021.14351a

Sürücü, L., & Maslakçi, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal*, *8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Tanjung, Y. I., & Bakar, A. (2019). Development of physical test instruments based on the conceptual knowledge dimension of the revision bloom taxonomy. *Jurnal Pendidikan Fisika*, *8*(2), 141–148.

Tanjung, Y. I., & Dwiana, Y. A. (2019). Pengembangan instrumen tes berbasis critical thinking skill pada materi gerak lurus. *INPAFI (Inovasi Pembelajaran Fisika)*, *7*(4).

Varghese, J., & Musthafa, M. N. M. A. (2021). Why the optimism misses? an analysis on the gaps and lags of teachers' perceptions of 21st century skills. *Shanlax International Journal of Education*, *10*(1), 68–75. https://doi.org/10.34293/education.v10i1.4322

Yusuf, I., & Widyaningsih, S. W. (2019). HOTS profile of physics education students in STEM-based classes using PhET media. *Journal of Physics: Conference Series*, *1157*, 032021. https://doi.org/10.1088/1742-6596/1157/3/032021