Jurnal Pendidikan Sains Indonesia
(Indonesian Journal of Science Education)
URL: http://jurnal.usk.ac.id/JPSI/index

11(4), p.853-864, (2023)
e-ISSN: 2615-840X p-ISSN: 2338-4379
DOI: doi.org/**10.24815/jpsi.v11i4.31756**

Development of Wordwall-Based Assessment Instruments to Measure Higher Order Thinking Skills Science Material Temperature, Heat, and Expansion

Anastasya Eka Saputri, Sri Wahyuni*, Diah Wahyuni

Program Studi Pendidikan IPA, FKIP Universitas Jember, Jember, Indonesia

*Email: sriwahyunifkip@unej.ac.id

Article History:
Received date: April 15, 2023
Received in revised from: September 19, 2023
Accepted date: October 7, 2023
Available online:October 15, 2023

Citation:
Saputri, A.E., Wahyuni, S., & Wahyuni, D. 2023. Development of wordwall-based assessment instruments to measure higher order thinking skills science material temperature, heat, and expansion. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 11(4):853-864.

Abstract. The application of higher order thinking skills (HOTS) in evaluation aims to improve the quality of students' thinking, especially about the process of solving a problem. This study aims to develop a wordwall-based assessment instrument to measure HOTS. This study methods used the Borg and Gall development model with 8 stages covering, looking for potential and formulating problems, information gathering, product design, product design validation, design revision, small group trials, product revisions, field trials. Respondents who took the test were divided into 2, namely 25 students for small group trials and all of class VII SMPN 2 Maesan, Bondowoso for field trials. The questionnaires used in this study were interview questionnaires, validation questionnaires and tests to find out students' HOTS. The results showed that the validity of the assessment instrument was 80.99% with a valid category along with the items, and the reliability of the assessment instrument was 0.770. The analysis of the discriminating power of the items produced 3 items in the very good category, 10 items in the good category, 4 items in the sufficient category, and 3 items in the poor category. Then the difficulty level analysis produces 5 questions in the too easy category, 12 questions in the medium category, and 3 questions in the too difficult category. Based on the results of the research, the assessment instrument developed was stated to be valid and suitable for use for subsequent learning evaluations even though there were minor revisions.
Keywords: Higher order thinking skills, wordwall, assessment instrument

Introduction

One of the latest education policies issued by Minister is to replace the national examination (UN) with the national assessment (AN). The national assessment system focuses on the process of evaluating and mapping the education system. The evaluation process applied to students requires higher-order thinking and not just understanding and remembering learning material (Putra et al., 2021). Students must be able to think at a higher level in the hope that it will encourage students to apply new and previous knowledge in solving problems in new situations (Dermawan et al., 2021). This scientific ability is defined as higher order thinking skills (HOTS).

HOTS can be interpreted as thinking activities that list the high hierarchical cognitive level of Bloom's thinking taxonomy, including analyzing, evaluating, and creating levels by Sudibyo et al. (2020). The application of HOTS in evaluation aims to

improve the quality of students' thinking, especially about the process of solving a problem. HOTS aims to improve cognitive thinking skills at a higher level, especially those related to the skill of analyzing learning materials that are very challenging to understand. The output of implementing HOTS is to further develop higher ways of thinking actively, critically, and creatively at a higher level to capture various types of information, solve problems and make decisions (Fadhil & Rokhimawan, 2020).

According to Bloom's cognitive taxonomy, students' thinking skills are classified into six thinking categories including remembering (C1), understanding (C2), applying (C3), analysis (C4), evaluation (C5), and creation (C6). Levels C1 to C3 are LOTS while levels C4 to C6 are HOTS by Saraswati & Agustika (2020). The levels of Bloom's taxonomy can be considered in Table 1.

In fact, in Indonesia, the ability to HOTS in students is still very low. The low ability of HOTS is confirmed by the results of the PISA report in 2018 which describes Indonesia's ranking in Science classified as low, which is ranked 71 out of 79 study participating countries (Hewi & Shaleh, 2020). Based on the PISA results, it can be stated that Indonesia's HOTS level is still low compared to other countries. The poor efforts of the Indonesian state in optimizing the HOTS process in learning and learning evaluation have caused the ranking in PISA to be very low. One aspect that triggers the low level of HOTS ability in students is the lack of application of HOTS-based evaluation questions and the teaching materials and media used do not trigger students to think in HOTS (Wahyuni et al., 2021). Based on interviews obtained at SMPN 2 Maesan, Bondowoso showed that the lack of application of HOTS-based questions and the majority of teachers are still not used to applying HOTS in the evaluation process carried out. This results in students not being familiar with the existence of HOTS-based questions which will later have an impact on the low ability of HOTS.

HOTS-based evaluation can motivate students to master higher-order thinking skills to apply their knowledge in finding solutions to new problems. Evaluation instruments containing HOTS questions can train students' ability to solve problems. Widarta & Artika (2021) stated that the characteristics of HOTS problems include being able to transfer one concept to another, wanting the ability to process and use information, forming the ability to draw relationships from various information, forming the ability to process information to overcome problems, and forming the ability to analyze ideas and information critically.

So an evaluation process is needed whose implementation is interesting and innovative in order to motivate student learning, to be able to meet the desired learning objectives. Kurniati et al. (2021) stated that one way that educators can always be creative and innovative in the evaluation process is to keep up with the times and utilize information and communication technology (ICT). Educators can utilize ICT to be implemented in learning evaluation, such as utilizing various available and easily accessible applications such as Wordwall. Wordwall is a learning media and evaluation that can be implemented in the learning process. In Wordwall there are also interesting and fun quiz-based games. The choice of wordwall as an assessment medium that is quiz-based games is a learning strategy solution that can be used in optimizing student learning outcomes. Thus, it is expected that using wordwall as an assessment instrument can optimize student learning outcomes, especially in the ability of HOTS. Based on the description of the problem above, the purpose of this study is to produce a word wall-based assessment instrument that is valid, reliable, has good distinguishing power and has difficulty. The results of this research are expected to provide ideas and information for developing quality HOTS questions.

Methods

This type of research is development research that aims to create a specific product by examining various testing processes to ensure its usefulness (Destiana et al., 2020). The research design used is the Borg and Gall development model which has systematic steps and is easy to understand. According to Hidayat et al., (2021) the Borg and Gall development model includes 10 general steps. These steps do not have to be followed because they can adjust to the needs of researchers, so this study only uses 8 stages. The stages of this research include (1) looking for potential and formulating problems, (2) collecting various information, (3) designing products, (4) validating related to product design, (5) revising designs, (6) conducting trials for small groups, (7) revising products, (8) conducting direct trials in the field. The flow of the Borg and Gall development model is presented in the flowchart in Figure 1.

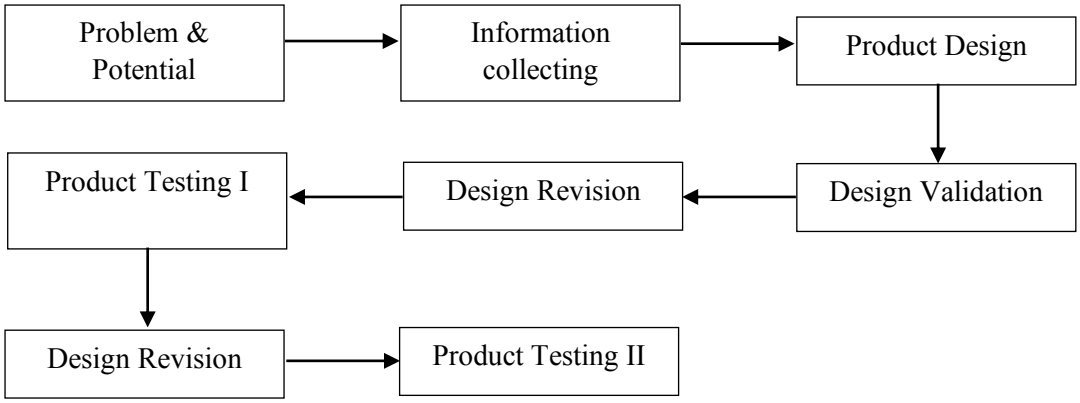


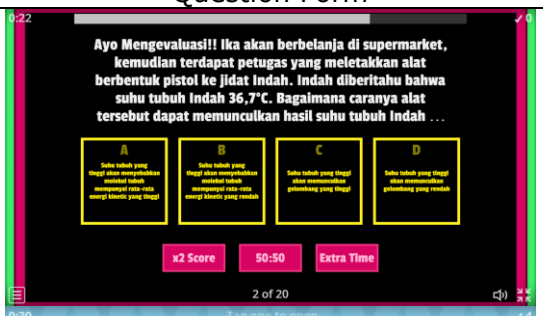
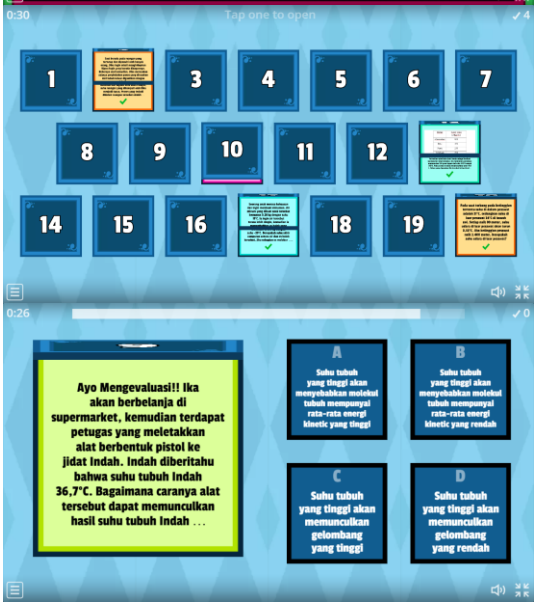
Figure 1. Borg and Gall development research pipeline

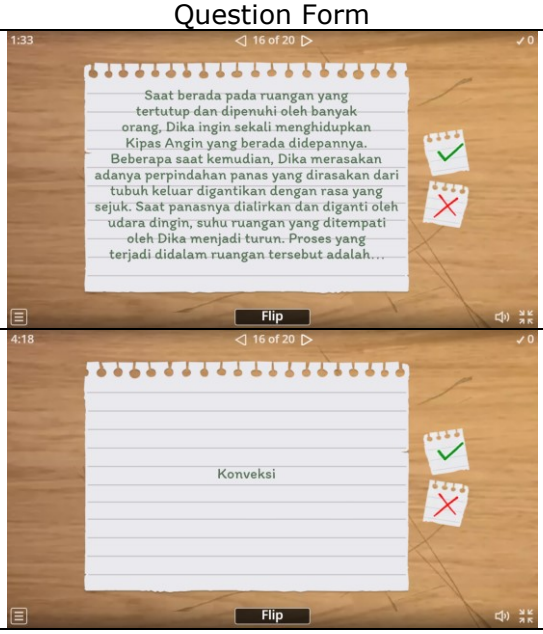
This research was conducted at SMP 2 Maesan Bondowoso in the 2022/2023 academic year in the even semester. The subjects of the study were all grade VII students with details of 25 students for small group trials and 100 students for field trials. Data collection techniques in this study are validation sheets, interviews, observations, tests, and documentation. Data analysis techniques in this study use expert validation analysis, empirical validity (analysis of question items, discriminating power, and level of difficulty), and reliability analysis of word wall-based assessment instruments.

Results and Discussion

The research was conducted at SMPN 2 Maesan with participants from all grade VII students. The results of the questions from the development of the HOTS assessment instrument are implemented in education-based media, namely word wall. The developed wordwall-based assessment can be accessed through several links, one of which is https://wordwall.net/resource/51904356. The link will later be sent to the grub of each class and the test work can be done online with good internet access. Making tests on the wordwall is divided into 4 sessions according to the number of classes in SMP 2 Maesan, namely classes A, B, C, and D. The development of wordwall-based assessment instruments using various templates can be seen in Table 1.

Table 1. Implementation of problems on various types of wordwall templates

Template Type	Question Form
Game Show Quiz	
Open The Box	

Template Type	Question Form
Flash Cards	

After the assessment instrument was developed, it was validated by 3 expert validators from science teachers at SMPN 2 Maesan. Validity relates to the ability of an assessment tool to measure the topic being studied by students, ensuring that it accurately measures what it should be. In accordance with Suseno (2021) stated that assessment instruments function to determine the achievement of learning given to students, so it is necessary to ensure that tools can measure accurately and validly. The results of expert validation of assessment instruments can be seen in Table 2.

Table 2. Expert validation results

Validator	Assessment Aspect			Percentage	Category
	Material	Construction	Language		
I	78	71	77	75.33	Valid
II	89	89	82	86.66	Sangat Valid
III	81	78	84	81	Sangat Valid
Average Score	82.66	79.33	81	80.99	Valid

Based on the results of expert validation of wordwall-based assessment instruments by validators, it can be stated that the instruments made are valid and suitable for use in the process of evaluating science learning of temperature, heat and expansion materials. The assessment of expert validators to determine the feasibility of the instrument is reviewed and assessed from the aspects of material, construction and language which has a total of 15 indicators (Alfiana et al., 2021). There are several suggestions and inputs from validators that must be added by researchers aimed at improving the quality of the assessment instruments developed. The results of suggestions and feedback are described in Table 3.

Table 3. Results of suggestions and feedback

Revised aspects	Questions Number	Suggestion	Before revision	After revision
Construction	1	Add a table about liquids so that students understand the differences between each liquid used	There is no explanatory table about the characteristics of thermometer filler liquid	Add a thermometer filler liquid character table
Material	10	Change the temperature unit of the problem to kelvin to deceive the clue	The temperature unit on the question is still Celsius according to the unit on the table type heat	Convert centigrade units to kelvins on the problem
Material	11	Convert the weight unit of the problem to grams to deceive the problem	The unit of weight on metal still uses kilograms	Convert weight units to grams
Construction	12	Replace 1 or 2 answer choices that were originally united kcal all into KJ or Joule	The unit of answer choice is the same, namely kcal	Change the unit of answer choice by changing KJ or Joule

In addition to being tested with validation based on experts, assessment instruments must also be tested in small group trials to determine the validity of the question items. Good quality question items will be a strong basis for interpreting the test using a high degree of validity. Valid instruments will also obtain valid data (Mamik, 2015), so that questions with valid information will produce precise data, while questions with invalid information will be revised until valid information is produced. Small group trials are conducted on 25 students and the test results will be analyzed with various tests, such as question item validity, reliability, discriminating power, and level of difficulty. The analysis was performed using microsoft excel and SPSS application. Analysis of the validity test of the question items can be seen in Table 4.

Table 4. Test results of validity of small group test items

Question Number	r Calculate	r Table	annotation
1	0.517	0.396	Valid
2	0.524	0.396	Valid
3	0.343	0.396	Invalid
4	0.343	0.396	Invalid
5	0.512	0.396	Valid
6	0.301	0.396	Invalid
7	0.319	0.396	Invalid
8	0.447	0.396	Valid
9	0.405	0.396	Valid
10	0.539	0.396	Valid
11	0.466	0.396	Valid
12	0.471	0.396	Valid
13	0.445	0.396	Valid
14	0.492	0.396	Valid

Question Number	r Calculate	r Table	annotation
15	0.513	0.396	Valid
16	0.427	0.396	Valid
17	0.306	0.396	Invalid
18	0.383	0.396	Invalid
19	0.322	0.396	Invalid
20	0.209	0.396	Invalid

The calculated results will be compared with the r table at t.sig 5% by adjusting the number of participants who take the test. If the calculated results are > r table, then the question items made are valid (Waminton, 2015). Question items that are declared invalid will be revised so that they can be reused in field trials. Furthermore, reliability tests were carried out using the Cronbach alpha formula which resulted in a reliability interval of 0.741. Based on the reliability criteria, the interval of small group test reliability results is included in the high criteria in the range of 0.70-0.90 (Waminton, 2015). Question items that are considered valid and reliable still have to go through further tests with an analysis test of the differentiating power and difficulty of the question items (Azizah et al., 2020). The results of the discriminating power test and the difficult test of the question items are recapitulated in Table 5.

Table 5. Results of recapitulation of discriminating power and difficulty level of question items in small group trials

Quuestion Number	Distinguishing Power		Difficulty Level	
	Differentiating Power Index	Classification	Difficulty Index	Interpretation
1	0.421	Excellent	0.68	Enough
2	0.452	Excellent	0.84	Too Easy
3	0.226	Enough	0.60	Enough
4	0.226	Enough	0.60	Enough
5	0.412	Excellent	0.64	Enough
6	0.181	Bad	0.60	Enough
7	0.203	Enough	0.64	Enough
8	0.341	Good	0.64	Enough
9	0.295	Enough	0.64	Enough
10	0.446	Excellent	0.68	Enough
11	0.381	Good	0.80	Too Easy
12	0.380	Good	0.76	Too Easy
13	0.335	Good	0.48	Enough
14	0.387	Good	0.56	Enough
15	0.410	Excellent	0.56	Enough
16	0.316	Good	0.60	Enough
17	0.185	Bad	0.56	Enough
18	0.272	Enough	0.64	Enough
19	0.201	Enough	0.48	Enough
20	0.084	Bad	0.44	Enough

Based on the results of the distinguishing power test of question items conducted using SPSS, there are still question items that are categorized as bad, so it needs to be revised so that the question items developed can sort out students who understand the material well from those who do not understand the material. A question item is recognized as having poor distinguishing power, when the question can be responded

appropriately by students who master the material or those who do not master the material (Fatimah et al., 2019). The results of the difficulty test of the question items are still in the interpretation of medium questions and easy questions, so it needs to be revised so that the difficulty of the question items is evenly distributed ranging from easy, medium to difficult with an even composition (Puspita et al., 2021). The composition of a good assessment instrument consists of 80% easy questions and 20% difficult questions (Dapiyanta, 2022).

The results of the question item test from small group trials that have not met the category are revised again by researchers so that the assessment instrument can be said to be valid and suitable for use in learning evaluation. Field trials of the final product are carried out when improvements have been made to the assessment instruments developed (Linda et al., 2021). Field trials were carried out on 100 students from a total of all grade VII students of SMPN 2 Maesan. The results of the data obtained will be tested again with the validity of the question items, reliability, and distinguishing power to test the level of difficulty using Microsoft excel and SPSS. Analysis of the validity test of question items in field trials can be seen in Table 6.

Table 6. The results of the validity test of the field trial question items

No Problem	r Calculate	r Table	Information
1	0.512	0.195	Valid
2	0.427	0.195	Valid
3	0.269	0.195	Valid
4	0.590	0.195	Valid
5	0.450	0.195	Valid
6	0.421	0.195	Valid
7	0.467	0.195	Valid
8	0.562	0.195	Valid
9	0.491	0.195	Valid
10	0.472	0.195	Valid
11	0.414	0.195	Valid
12	0.349	0.195	Valid
13	0.447	0.195	Valid
14	0.355	0.195	Valid
15	0.487	0.195	Valid
16	0.372	0.195	Valid
17	0.468	0.195	Valid
18	0.382	0.195	Valid
19	0.358	0.195	Valid
20	0.309	0.195	Valid

Based on the results of the validity test, the question items in the field trial data have produced 20 valid questions. Changes in question items that are invalid from small group trials turn into valid in field trials because the invalidity of questions can be affected by several factors, namely factors from internal tests, external test factors, and factors derived from students. Invalid questions made by researchers because they are influenced by factors from students. Factors from these students such as students are not careful in doing tests and deliberately answer randomly. Furthermore, reliability tests were carried out on field trial data resulting in 0.770 which is included in the high reliability interval. The increase in reliability results when tested in the field is due to the addition of students when taking the test which was originally only 25 people to 100 people and besides that there is also a need for revisions to other analyses that affect the increase in reliability test results (Husna et al., 2017). The existence of increased results

in the results of field trials does not make the assessment instrument invalid, but is still declared reliable and can be used in further learning evaluation (Hasanah & Tsutaoka, 2019). In accordance with Arikunto's opinion (2018), that tests tested on many students will affect the results of the reliability of the tests developed. Furthermore, differentiating power tests and difficulty tests were carried out on field trial data. The results of the discriminating power test and the difficulty level test of the question items are recapitulated in Table 7.

Table 7. Results of recapitulation of discriminating power and difficulty level of question items in field trials

Question Number	Differentiating Power		Difficulty Level	
	Differentiating Power Index	Classification	Indeks Kesukaran	Differentiating Power Index
1	0.431	Excellent	0.78	Too Easy
2	0.327	Good	0.69	Enough
3	0.164	Bad	0.26	Too Hard
4	0.502	Excellent	0.59	Enough
5	0.357	Good	0.74	Too Easy
6	0.318	Good	0.67	Enough
7	0.366	Good	0.63	Enough
8	0.468	Excellent	0.52	Enough
9	0.395	Good	0.66	Enough
10	0.373	Good	0.65	Enough
11	0.314	Good	0.71	Too Easy
12	0.248	Enough	0.27	Too Hard
13	0.343	Good	0.61	Enough
14	0.251	Enough	0.70	Enough
15	0.386	Good	0.58	Enough
16	0.272	Bad	0.73	Too Easy
17	0.374	Good	0.71	Too Easy
18	0.273	Enough	0.62	Enough
19	0.254	Enough	0.29	Too Hard
20	0.195	Bad	0.61	Enough

Question items in an assessment instrument must be able to distinguish between students who master the topic and those who do not so that it can be said that the assessment tool has excellent distinguishing power. Question items that have a high discriminating power index close to 1.00 are declared to have good discriminating power, while those with a low index close to 0 to negative are recognized as having poor discriminating power (Kusaeri, 2014). In the results of field trials, there are still question items that have poor distinguishing power. There are several factors that affect question items that have poor distinguishing power, namely students only do it and then choose the answer randomly but get the right answer, so that the question item is difficult to distinguish which students already understand the material from those who do not understand the material (Fithri et al., 2021).

The results of the difficulty test will be interpreted in several categories, namely the question is too easy when getting a difficulty index above 0.70, the question is considered medium when getting a difficulty index between 0.30-0.70, while the question is considered too difficult when getting a difficulty index less than 0.30. The composition of a good assessment instrument consists of 80% easy questions and 20% difficult questions (Dapiyanta, 2022). An assessment instrument can be said to be good when it

has a proportion of questions that are evenly divided between easy and difficult questions (Indriaty & Setyoko. 2018). The results of field trials have a composition that is in accordance with theory, which has 17 easy questions in which there are easy and medium questions and 3 difficult questions.

Conclusion

Based on the results of the analysis of the research that has been conducted, the assessment instrument has been considered valid and suitable for use after being tested through tests of validity, reliability, discriminating power, and level of difficulty. The validity obtained is 80.99%, the reliability of the instrument is 0.770, and the questions already have good distinguishing power, so that the level of difficulty of the entire question is evenly divided, so it can be concluded that the assessment instruments developed can be used in the evaluation of science learning to measure higher order thinking skills.

References

Acesta, A. 2020. Analisis kemampuan higher order thinking skills (HOTS) siswa materi IPA di Sekolah Dasar*. Quagga: Jurnal Pendidikan dan Biologi,* 12(2):170-175*.* https://doi.org/10.25134/quagga.v12i2.2831

Alfiana, R., Parno, & Yogihati, C. I. 2021. Development of ILAU based on PBL-STEM model with formative assessment as an opportunity to improve problem solving skills in heat and temperature topics. Journal of Physics: Conference Series, 1747:1-8.

Arikunto, S. 2018. *Dasar-Dasar Evaluasi Pendidikan*. Jakarta: PT Bumi Aksara.

Azam, I.F. & Rokhimawan, M.A. 2020. Analisis materi IPA kelas IV tema indahnya kebersamaan dengan HOTS*. Jurnal Ilmiah Didaktika: Media Ilmiah Pendidikan dan Pengajaran*, 21(1):100-110. http://dx.doi.org/10.22373/jid.v21i1.5970

Azizah, W.A., Sarwi, S., & Ellianawati. 2020. Development of stream-based teaching materials in training students' process skills through science project activities. *Advances in Social Science, Education and Humanities Research*, 574:147-156

Dapiyanta, F.X. 2022. *Evaluasi Hasil Pembelajaran Pendidikan Agama Katolik*. Depok: PT Kanisius.

Dermawan, D.D., Wardani, S., & Pranoto, Y.K.S. 2021. Pengembangan instrumen assesmen HOTS pada pembelajaran IPA kelas V Sekolah Dasar. *Jurnal Ilmiah Ilmu Kependidikan*, 5(3):387-394. https://doi.org/10.26858/jkp.v5i3.21869

Destiana, D., Suchyadi, Y., & Anjaswuri, F. 2020. Pengembangan instrumen penilaian untuk meningkatkan kualitas pembelajaran produktif di Sekolah Dasar. *Jurnal Pendidikan dan Pengajaran Guru Sekolah Dasar (JPPGuseda)*, 3(2):119-123. https://doi: https://doi.org/10.55215/jppguseda.v3i2.2720

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. https://doi.org/10.36668/jal.v8i2.115

Fithri, S., Pada, A.U.T., Artika, W., Nurmaliah, C., & Hasanuddin. 2021. Implementasi LKPD berbasis stem untuk meningkatkan keterampilan berpikir kritis peserta didik. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 9(4): 555–564. https://doi.org/10.24815/jpsi.v9i4.20816

Hasanah, U., & Tsutaoka, T. 2019. An outline of worldwide barriers in science, technology, engineering and mathematics (STEM) education. *Jurnal Pendidikan IPA Indonesia*, 8(2):193–200. https://doi.org/10.15294/jpii.v8i2.18350

Hidayat, A.S., F. Dlis, & S. Hanief. 2021. *Pengembangan Model Pembelajaran Atletik Nomor Lari Berbasis Permainan Pada Siswa Sekolah Dasar*. Purwodadi: CV. Sarnu Untung.

Hewi, L. & Shaleh, M. 2020. Refleksi hasil PISA (the programme for international student assesment): upaya perbaikan bertumpu pada pendidikan anak usia dini. *Jurnal Golden Age*, 4(1):30-41. https://doi.org/10.29408/goldenage.v4i01.2018

Husna, N., Sari, S.A., & Halim, A. 2017. Pengembangan media puzzle materi pencemaran lingkungan di SMP Negeri 4 Banda Aceh. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 5(1):66–71. http://jurnal.unsyiah.ac.id/JPSI/article/view/8413

Indriaty & Setyoko. 2018. Pengembangan bahan ajar ekologi berbasis problem based learning di Program Studi Pendidikan Biologi Universitas Samudra. *JIPI (Jurnal IPA dan Pembelajaran IPA)*, 2(1):29-35

Linda, R., Zulfarina, Mas'ud, & Putra, T.P. 2021. Peningkatan kemandirian dan hasil belajar peserta didik melalui implementasi e-modul interaktif IPA terpadu tipe connected pada materi energi SMP/MTS. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 9(2):191-200. https://doi.org/10.24815/jpsi.v9i2.19012

Kusaeri. 2014. *Acuan dan Teknik Penilaian Proses dan Hasil Belajar dalam Kurikulum* 2013. Yogyakarta: Ar-Ruzz Media.

Mamik. 2015. *Metodologi Kualitatif*. Sidoarjo: Zifatama Publishing.

Puspita, K., Nazar, M., Hanum, L., & Reza, M. 2021. Pengembangan e-modul praktikum kimia dasar menggunakan aplikasi canva design. *JIPI (Jurnal IPA dan Pembelajaran IPA)*, 5(2):151-161. https://doi.org/10.24815/jipi.v5i2.20334.

Putra, P.D.A., Ahmad, N., Wahyuni, S., & Narulita, E. 2021. An analysis of the factors influencing of pre-service science teacher in conceptualization of stem education: self-efficacy and content knowledge. *Jurnal Penelitian Pendidikan IPA*, 7(specialIssue):225–230. https://doi.org/10.29303/jppipa.v7iSpecialIssue.877.

Sudibyo, E., Nurita, T., Qosyim, A., & Sabtiawan, W.B. 2020. Pendampingan pengembangan instrumen penilaian berstandar nasional bagi guru IPA SMP Kabupaten Kediri. *Jurnal Abdi: Media Pengabdian Kepada Masyarakat*, 5(2):127-133. https://doi.org/10.26740/ja.v5n2.p127-133

Suseno, E. & Susongko, P. 2021. *Mengukur Validitas Tes*. Kediri: Edukreatif.

Wahyuni, S., Rizki, L.K., Budiarso, A.S., Putra, P.D.A., & Narulita, E. 2021. The development of e-student worksheet on environmental pollution to improve critical thinking skills of Junior High School students. *Jurnal Penelitian Pendidikan IPA*. 7(4):723–728. https://doi.org/10.29303/jppipa.v7i4.870

Waminton, R. 2015. *Evaluasi Hasil Belajar Matematika*. Yogyakarta: Media Akademi.

Widarta, F.O. & Artika, W. 2021. Analisis bentuk stimulus, dimensi kognitif, dan karakteristik HOTS pada instrumen evaluasi mata pelajaran IPA karya guru. *JIPI (Jurnal IPA & Pembelajaran IPA)*, 5(3):197-208. https://doi.org/10.24815/jipi.v5i3.21429