ISSN (PRINT) : 2549-9955 ISSN (ONLINE): 2549-9963
JURNAL ILMIAH PENDIDIKAN FISIKA
https://ppjp.ulm.ac.id/journals/index.php/jipf/index

Development of Learning Outcome Test Instruments on Thermodynamics and Waves Based on the Demands of Basic Competencies

Abdul Haris, Khaeruddin, and Sri Reski
Physics Education Study Program, FMIPA, Universitas Negeri Makassar
Makassar, Indonesia
*abd.haris@unm.ac.id

Abstract

Evaluation is a process of determining the results of a learning activity that has been conducted using instruments. This development research aims to develop test instruments for physics learning outcomes that meet valid and reliable criteria. The tests developed were multiple-choice tests in the form of physics material based on the 2013 curriculum, such as thermodynamics and mechanical waves. This research is a research and developmental project that uses the development method according to Djaali and Mulyono, which generally has seven steps: (1) formulating conceptual and operational definitions, (2) designing instruments, (3) reviewing statements, (4) testing, (5) analysis, (6) revising instruments, and (7) instrument assembly into the final instrument. The study's results were analyzed using Microsoft Excel, analyzing the items' validity, reliability, difficulty level, and differentiability level. Validity using the biserial point coefficient, reliability using Kuder Richardson 20, difficulty level using the simple proportion of correct answers, and differentiability using high groups. The research results indicate that the physics learning outcomes test meets valid and reliable criteria. The valid question items were 33 items, and the reliability of the instrument in the first test was 0.77, and in the second test was 0.81. Teachers can use the developed instrument to assess students' learning outcomes because it remains consistent after testing on the same subjects and conditions.

Keywords: instruments; learning outcome; physics

Received : 25 September 2023
Accepted : 7 February 2024
Published: 25 March 2024
DOI : https://doi.org/10.20527/jipf.v8i1.10217
© 2024 Jurnal Ilmiah Pendidikan Fisika

How to cite: Haris, A., Khaeruddin, K., & Reski, S. (2024). Development of learning outcome test instruments on thermodynamics and waves based on the demands of basic competencies. *Jurnal Ilmiah Pendidikan Fisika*, 8(1), 112-122.

INTRODUCTION

Assessment is an important part of the education system that can help learners become more knowledgeable, critical, competent, and responsive. Through assessment, teachers can find out the competencies possessed by students. The competencies that students must have after the learning process are attitude, skills, and knowledge competencies (Kemendikbud, 2013).

The Ministry of Education and Culture, 2013 says that educational assessment standards are criteria regarding mechanisms, procedures, and instruments for assessing student

This is an open access article under the CC–BY-SA license

learning outcomes. Assessment of learning outcomes is the competence of attitudes, skills, and knowledge to determine the competence of each learner and whether it is in accordance with predetermined standards. The content in the assessment of learning outcomes includes, among others, the scope of material, subject competencies, content, or programs and processes. (Ani, 2018).

Competence is a combination of cognitive skills comprising personal and behavioral characteristics and is linked to the quality of students (Hodges & Burchell, 2003; Pop & Khampirat, 2019). Assessment is the application of various methods and instruments to obtain information about the extent of students' learning outcomes or the achievement of learning competencies (Rahman & Nasryah, 2019).

Assessment is the process of collecting and processing information to determine students' learning needs and developmental achievements or learning outcomes (Permendikbud, 2022). Evaluation is the activity or process of assessing and measuring something (Sawitri & Rahayu, 2018). From several definitions of evaluation, it can be concluded that evaluation is a process to determine the results of an activity (learning) that has been carried out using tools.

Compiling learning outcome tests aims to measure students' success in mastering basic competency indicators. Teachers can identify students' weaknesses by looking at the results obtained. Compiling a test that meets the requirements is quite difficult because compiling a test requires relatively high knowledge, skills, and accuracy (Papaya, 2022). Teachers can provide students with experiences covering the affective, cognitive, and psychomotor learning domains, which aligns with the learning outcomes in Bloom's taxonomy (Istiyono, 2020; Roldán-Merino et al., 2019).

Characteristics of a good test require a valid and reliable instrument to be measured (Arifin, 2013; Shirali et al., 2018). Validity relates to whether the test measures what is to be measured, while reliability relates to how reliable and useful the test is (Nurhasanah, 2018). The validity and reliability of the instrument are influenced by the user of the instrument, the research subject or the one being measured, and the instrument itself. Therefore, validity and reliability must be tested before the instrument is used (Yusup, 2018).

In Ko et al., (2017), a useful measure or scale should have the following characteristics: (1) it should be appropriate to the task; (2) it should be valid, that is, it should measure what it purports to measure; (3) it must be accurate and should accurately measure what it purports to measure; (4) it must be reliably reproducible (precise); (5) it should be efficient and easy to use, with little special training; (6) it should be sensitive to change in the underlying condition yet relatively insensitive to symptom fluctuation; and (7) it should be consistent over time, that is, not subject to so-called frame-of-reference shifts (Herndon, 2006).

The research revealed that teachers do not have the ability to develop tests. The preparation of learning outcome tests carried out by teachers has not fulfilled standardized tests, and the alignment between question indicators and question items still has many errors. Moreover, some teachers still use existing tests and then adjust them to the teaching material. On the other hand, most teachers have not been able to compile tests, so they often look for several collections of existing questions (Pakaya, 2022).

One of the serious problems in implementing Curriculum 2013 is the problem of assessing student learning

outcomes. The 2013 curriculum requires that learning outcomes be assessed comprehensively by involving the three main domains of assessment, namely the knowledge domain, the attitude domain, and the skills domain.

At the advanced education level (SMA), the assessment of student learning outcomes is more on the knowledge dimension, followed by the skills and attitude dimensions (Kemendikbud, 2013b). The problems faced by teachers in assessing student learning outcomes are 1) the number of assessments, 2) the complexity of assessments, 3) making assessment instruments, 4) implementing assessments, and 5) reporting learning outcomes (Subagia & Wiratma, 2016).

Based on several journals on research and development of test instruments using research models such as 4-D, developed by Thiagarajan, ten research steps developed by Borg and Gall, and ADDIE. Research has also been conducted on physics subjects, but there has been no development of physics learning outcomes test instruments on thermodynamics and wave material.

Based on the description above, the researcher wants to develop physics learning outcomes test instruments that meet valid and reliable criteria to help teachers make physics instruments in the form of question items that meet valid and reliable criteria. According to Djaali & Muljono (2008), the research used the development model, which was modified with other experts' opinions, resulting in seven development steps.

METHOD

The type of research conducted is research and development (R & D). This theory is similar to the instrument development steps according to Gable, 1986. This research was tested at SMA Negeri 14 Gowa for students in class XI IPA in the 2022/2023 academic year. This development research design uses a test instrument development model according to Djaali and Muljono (2008) and several opinions of experts or other experts who generally have seven steps (Wahyudi, 2018). The test instrument development steps can be seen in Figure 1.

Figure 1 Learning outcome test instrument development steps

The research instruments used were validation sheets for validators and physics learning outcomes test questions. The validation sheet is made for validators to find out suggestions or opinions about the instruments that have been designed. The validated instruments were then analyzed using Gregory's analysis based on the content validity test agreement table between validators. The learning outcome test questions were in the form of multiple-choice questions consisting of 5 choices. Data analysis techniques in this study using descriptive statistical techniques. Data was analyzed using:

Gregory's analysis

Using qualitative descriptive analysis, categorization can be used: without revision, minor revision, major revision, cannot be used/still requires consultation. Analysis was tested on all question items by taking into account the results of the validator's assessment to determine their content validity. The validity of the instrument was evaluated

using the content methods. The content validity was used to determine the relevance of the items in the instrument (Shiralia, 2017). Two experts validated the questions and then underwent the Gregory test using the agreement model. Gregory's test uses the content validity formula (Aryawan et al., 2014).

Item Analysis

Item analysis testing to determine the validity of the learning outcome items after the testing is carried out. The formula used is the biserial point coefficient (r_{bisp}) (Arikunto, 2003). The validity criteria are if t count t table then the item is said to be valid and if t count < t table then the item is invalid. t table is obtained by looking at the predetermined table, where the number of respondents is reduced by two. With α significance level of 0.05, the t table is 1.67.

Instrument Reliability

Testing the instrument's reliability in this study using the Kuder Richardson 20 (KR-20) equation (Sugiyono, 2012). Kuder and Richardson formulated a formula to calculate the reliability of tests consisting of dichotomous items, using the average proportion of subjects scoring 1 ('Ulum, 2016). It is acceptable if the instrument's reliability is above 0.70 (Kholed et al., 2021; Kim et al., 2020; Zainuddin et al., 2020).

Difficulty Level and Different Level

The level of difficulty uses 3 ranges with difficult, medium, and easy criteria. The difficulty level is analyzed based on the proportion of correct answers for the whole group. Meanwhile, the differentiating level uses the high-low group formula. The different levels are divided into two parts: high groups and low groups. The division of high and low groups uses 27% of respondents (Susetyo, 2015).

RESULT AND DISCUSSION

Formulation of Conceptual and Operational Definitions

At this stage, the researcher formulates a conceptual definition of the material to be developed in the Physics class XI Curriculum 2013. The basic competencies used are 3.7) Analyzing changes in the state of an ideal gas by applying the laws of thermodynamics and 3.8) Analyzing the characteristics of waves. While the formulation of operational definitions is to formulate the variables to be developed, the variables of this study are multiple-choice items in the cognitive domain at the level of remembering, understanding, applying, analyzing, and evaluating.

Instrument Design

At the instrument design stage, researchers designed instruments that would be developed. There were 55 items that would then be made indicators of questions based on the operational verbs of each cognitive level. The design of the instrument is referred to as draft I.

Table 1 Learning outcome test instrument design

Cognitive Level	Total
Remembering	10 Items
Understanding	11 Items
Applying	13 Items
Analyzing	17 Items
Evaluate	4 Items
Total	55 Items

The physics learning outcomes test is made based on the learning competency grid and is arranged based on Bloom's Taxonomy in the cognitive domain, starting from remembering, understanding, applying, analyzing, and evaluating.

Statement Review

The instrument at the statement review stage will produce draft II. The test instrument was validated by two different experts and analyzed using

Gregory's table for each test item. Based on the results of the expert assessment, the reliability coefficient of the learning outcomes test is 0.89.

After the instrument was revised based on the expert's assessment, all aspects of the assessment of the learning outcomes test instrument met the feasibility. Several suggestions from experts need to be considered, as well as revision of questions for perfection in the use of testing. The suggestions are:

a. Uses standardized and clear item statements
b. Pay attention to typing errors
c. Correcting unclear images
d. Pay attention to the right answer choices
e. Eliminating items that are not relevant to what is being measured

Figure 2 shows one of the developed learning outcome tests and validator suggestions.

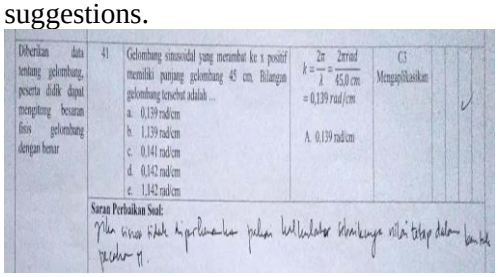

Figure 2 Example of developed learning outcome test

After being validated, the questions were revised:

Gelombang sinusoidal yang merambat ke x positif memiliki panjang gelombang 20 cm. Bilangan gelombang tersebut adalah ...

a. $0{,}1\pi$ rad/cm
b. $0{,}15\pi$ rad/cm
c. $0{,}2\pi$ rad/cm
d. $0{,}25\pi$ rad/cm
e. $0{,}3\pi$ rad/cm

Figure 3 After validation

Revision activities are intended to evaluate and improve the first draft that has been designed. The number of items for the learning outcomes test instrument is shown in Table 2.

Table 2 Number of learning outcome test items after expert validation

Cognitive Level	Total
Remembering	10 Items
Understanding	10 Items
Applying	12 Items
Analyzing	13 Items
Evaluate	4 Items
Total	49 Items

Testing

The first phase of testing was conducted in class XI SMAN 14 Gowa on 58 respondents using 49 items. After the phase testing, we continued analyzing students' physics learning outcomes. After the analysis was conducted, the second phase of testing was continued using questions that met the valid criteria in the first testing analysis, such as 40 items. The second phase of testing was conducted on the same subjects.

Data analysis

After the first and second testing, data analysis was conducted for each learning outcome test instrument. The results of data analysis after the testing showed the number of valid and invalid items, instrument reliability, difficulty level, and differentiation level. The analysis used the following descriptive statistics:

Item analysis

Item analysis is conducted to determine valid and invalid tests. Based on the analysis conducted based on the results of the first testing phase, 40 items meet the valid criteria, as shown in Table 3. In the analysis of the second phase of testing, 33 items met the valid criteria, which can be displayed in Table 3.

Table 3 Number of physics learning outcome test items

Cognitive Level	Total
Remembering	8 Items
Understanding	8 Items
Applying	8 Items

Analyzing	12 Items
Evaluate	4 Items
Total	40 Items

Instrument reliability

Instrument reliability uses the KR-20 formula to determine the reliability of an instrument. The results of the first phase testing showed the reliability of the learning outcomes test instrument was 0.77, while the results of the second phase testing showed the instrument's reliability was 0.81.

Difficulty Level

The results of the level of difficulty analysis were conducted for the first and the second phases of testing. In the first testing phase, the difficulty level developed was seven easy items, 27 medium items, and 15 difficult items. Meanwhile, the second phase of testing of the physics learning outcomes test contained 8 easy items, 25 medium items, and seven difficult items.

Different Level

Differentiating level analysis was conducted for the first and second phase testing. The differentiating level of the items was analyzed based on the criteria used, such as items that have an excellent differentiating level, items that have a good enough differentiating level, items that require minor revision, items that require revision or are set aside, and items that are completely revised or set aside.

The results of the first phase of testing different levels of analysis show that 17 items are quite good, 12 items require minor revisions, six items require revision, and 14 are completely revised. Meanwhile, for the second testing phase, there are questions with different levels: 3 items are very good, 16 items are quite good, six items require minor revisions, seven items require revision or are set aside, and eight items are completely revised.

Instrument Revision

At this stage, the instrument was revised based on the analysis results in the first testing phase. The revised results were then tested again on the same subjects: students of class XI SMAN 14 Gowa. Based on the revision of the instrument, the number of items used in the second phase of testing was 40 items. The question items used are in Table 3.

Arranging Instruments into a Final Instrument

The final instrument is an amalgamation of expert review, pilot test results, analysis, and revision of the instrument referred to as draft III. The number of items of the final physics learning outcomes test instrument is shown in Table 4.

Table 4 Final physics learning outcome test instrument

Cognitive Level	Total
Remembering	7 Items
Understanding	7 Items
Applying	7 Items
Analyzing	9 Items
Evaluate	3 Items
Total	33 Items

There are seven items about remembering, seven items about understanding, seven about applying, seven about analyzing, nine items, and three about evaluating that meet the valid criteria with instrument reliability of 0.81.

This research and development aims to produce instruments that meet valid and reliable criteria for validity, referring to the content validity criteria. In this field, content validity refers to the suitability of the questions with the learning outcome test indicators. The proportion of items that received a rating of 3 or 4 by the experts was accepted as demonstrating content validity (Gül et al., 2023).

Expert agreement between items judged relevant by the experts was analyzed using the Gregory table.

Ideally, item and scale validity should be 0.80–0.90 or better (Gee et al., 2023; Polit & Beck, 2006). The content validity after analysis was 0.89, indicating an ideal instrument.

Test retests were employed to examine the reliability of the instrument; the purpose of reliability retests is to measure whether an instrument can reliably replicate its results in similar situations and populations (Mentzel et al., 2016; Shirali et al., 2018; Zhao et al., 2021). The testing was conducted twice to determine the reliability of the instrument.

Instruments with reliability higher than 0.70 already show adequate consistency and are considered acceptable (Kim et al., 2021; Lasater et al., 2022; Wulff et al., 2021). The reliability obtained in the research results for the first and second tests is 0.77 and 0.81, indicating that the instrument can be used to assess student learning outcomes.

Valid means that the test should be able to measure what should be measured. For this reason, learning outcome tests must be formulated clearly and carefully so that the behaviors to be measured can measure learning objectives. Reliability or consistency means that the measurement results are always consistent when carried out on the same students at different times and conditions (Ali & Khaeruddin, 2012). The reliability test is used to determine the consistency of the measuring instrument and whether the measuring instrument used is reliable and remains consistent if the measurement is repeated. The method used is reliable and remains consistent if repeated measurements are used (Dewi, 2018).

The results of theoretical and empirical validity have shown that the learning outcome test instrument is valid and that valid items are reliable. In theory, the test developed has also met the reliable or stable category.

In addition to validity and reliability, instrument development is also based on analyzing the difficulty and differentiating levels. The difficulty level and differentiating level. The difficulty level uses the criteria of difficult, medium, and easy. The criteria for the difficulty level are: 0-0.24 is a difficult category question, 0.25-0.74 is a medium category question and 0.75-1.00 is an easy category question.

The differentiating level serves to calculate the level of validity. It describes the level of ability of the question in differentiating between students who already understand the material tested and students who have not / do not understand the material tested. The criteria are 0.70-1.00 items have excellent differentiating level, 0.40-0.69 items have good enough different level, 0.30-0.39 items require little or no revision, 0.20-0.29 items require revision or are set aside and 0.19-0.00 items are completely revised or set aside (Susetyo, 2015).

The differentiating level in the first testing there were 4 items, and in the second testing, there were two items that were not in the criteria. Items not within the criteria of being below 0 are questions that are in the invalid category.

Differentiating levels distinguish the ability of high-ability learners from low-ability learners. In this study, some items do not meet the criteria for differentiating levels, so it is recommended that other researchers develop items that are within the criteria for differentiating levels to differentiate students' abilities.

CONCLUSION

The results of developing learning outcome test instruments met the valid criteria after being reviewed by experts and tested. Learning outcome tests that fulfill the valid criteria after the first

phase of testing are 8 items of remembering, nine items of understanding, eight items of applying, 12 items of analyzing, and four items of evaluating. In the second phase, testing is seven items of remembering, seven items of understanding, seven items of applying, nine items of analyzing, and three items of evaluating.

The results of developing learning outcomes test instruments have fulfilled the reliability criteria after testing. Reliability above 0.75 has met the reliability criteria. After the analysis, the reliability value of the physics learning outcomes test in the first phase testing was 0.77, and for the second phase testing, it was 0.81.

The developed instrument can be used by teachers in assessing students' learning outcomes on thermodynamics and waves because it remains consistent after repeated trials on the same subjects and conditions. The development of learning outcome tests can be used as an instrument for assessing students in order to obtain accurate data.

REFERENCES

Ali, S., & Khaeruddin. (2012). *Evaluasi pembelajaran*. Badan Penerbit UNM Makassar.

Ani, Y. (2018). Penilaian autentik dalam kurikulum 2013. *Seminar Nasional Implementasi Kurikulum 2013*, 742–749.

Amelia, M. A. (2016). Analisis soal tes hasil belajar high order thinking skills (hots) matematika materi pecahan untuk kelas 5 materi dasar. *Jurnal Penelitian (Edisi Khusus PGSD)*, *20*(2), 123–131.

Arifin, Z. (2013). *Evaluasi pembelajaran*. PT Remaja Rosdakarya.

Arikunto, S. (2003). *Dasar-dasar evaluasi pendidikan*. Bumi Aksara.

Aryawan, P., Sulastri, M., & Seeknayasa, G. (2014). Pengaruh konseling kelompok dengan pelatihan tutor sebaya terhadap kepemimpinan peserta didik peserta smansa conseling club (ssc) di sma negeri 1 singaraja. *E-Journal Undiksha Jurusan Bimbingan Konseling*, *2*, 1–10.

Asrul, A., Ananda, R., & Rosnita, R. (2014). *Evaluasi pembelajaran*. Citapustaka Media.

Dewi, D. A. N. N. (2018). Modul uji validitas dan hormonal. In *Universitas Diponegoro* (Issue October). https://www.researchgate.net/publication/328600462

Djaali, D., & Mulyono, P. (2008). *Pengukuran dalam bidang pendidikan*. Grasindo.

Gable, R. K. (1986). *Instrument development in affective domain*. Klawer-Nichoff Publishing.

Gee, P. M., Clift, A., Osteyee, J., Katz, E. B. P. C. R., & Collingridge, D. (2023). Fall risk instrument for pediatric behavioral health patients (BHFR-P©): A validity and reliability study. *Journal of Pediatric Nursing*, *69*, 118–120. https://doi.org/10.1016/j.pedn.2023.02.011

Gül, Ş., Avşar, P., Karadağ, A., Karaçay, P., & Gökmen, D. (2023). The validity and reliability of the Pressure Ulcer Quality of Life instrument: A methodological study. *Journal of Tissue Viability*, *32*(1), 102–106. https://doi.org/10.1016/j.jtv.2022.12.003

Gunada, I. W., Sahidu, H., & Sutrio, S. (2015). Pengembangan perangkat pembelajaran fisika berbasis masalah untuk meningkatkan hasil belajar dan sikap ilmiah mahapeserta didik. *Jurnal Pendidikan Fisika Dan Teknologi*, *1*(1), 38–46. https://doi.org/10.29303/jpft.v1i1.233

Hariyanto, Y., & B, I. G. P. A. (2015). Pengaruh metode pembelajaran tipe talking chips terhadap hasil belajar

peserta didik pada kompotensi dasar memahami model atom bahasan semi konduktur di smk negeri 1 jetis mojokerto. *Jurnal Pendidikan Teknik Elektro*, *4*(3), 999–1005.

Herndon, R. M. (2006). *Handbook of neurologic rating scales*. 2nd ed., Demos Medical Publishing.

Hodges, D., & Burchell, N. (2003). Business graduate competencies: Employers' views on importance and performance. *Asia-Pacific Journal of Cooperative Education*, *4*(2), 16–22.

Istiyono, E. (2020). *Penilaian dan analisis hasil belajar fisika*. UNY Press.

Kemendikbud. (2013). *Permendikbud No. 65 tahun 2013 Tentang Standar Proses Pendidikan Dasar dan Menengah*.

Kemendikbud. (2013). *Permendikbud No 66 Tahun 2013 tentang Standar Penilaian Pendidikan*. https://doi.org/10.1016/j.metabol.2009.10.012

Kemendikbud. (2018). *Permendikbud RI Nomor 37 tahun 2018 tentang Perubahan atas Peraturan Menteri Pendidikan dan Kebudayaan Nomor 24 tahun 2016 tentang Kompetensi Inti dan Kompetensi Dasar Pelajaran pada Kurikulum 2013 pada Pendidikan Dasar dan Pendidikan Menengah*.

Kholed, S. N. S., Maon, S. N., & Hassan, N. M. (2021). Reliability and validity of the inter-professional collaboration practice instrument. *Journal of Interprofessional Education and Practice*, *24*(October 2019), 100450. https://doi.org/10.1016/j.xjep.2021.100450

Kim, H. Y., Nho, J. H., Kim, J. Y., & Kim, S. R. (2021). Validity and reliability of the Korean version of the world health organization quality of life instrument-older adults module. *Geriatric Nursing*, *42*(2), 548–554. https://doi.org/10.1016/j.gerinurse.2020.10.006

Kim, M. J., McKenna, H., Park, C. G., Ketefian, S., Park, S. H., Galvin, K., & Burke, L. (2020). Global assessment instrument for quality of nursing doctoral education with a research focus: Validity and reliability study. *Nurse Education Today*, *91*, 104475. https://doi.org/10.1016/j.nedt.2020.104475

Ko, M. M., Lee, M. S., Birch, S., & Lee, J. A. (2017). The reliability and validity of instruments measuring pattern identification in Korean medicine: A systematic review. *European Journal of Integrative Medicine*, *15*(September), 47–63. https://doi.org/10.1016/j.eujim.2017.09.005

Lasater, M. E., Beebe, M., Warren, N. E., Winch, P. J., Soucko, F., Keita, M., Doumbia, S., & Murray, S. M. (2022). Reliability and validity of a perinatal depression screening instrument in rural Mali. *SSM - Mental Health*, *2*(September 2021), 100059. https://doi.org/10.1016/j.ssmmh.2021.100059

Mentzel, T. Q., Lieverse, R., Levens, A., Mentzel, C. L., Tenback, D. E., Bakker, P. R., Daanen, H. A. M., & van Harten, P. N. (2016). Reliability and validity of an instrument for the assessment of bradykinesia. *Psychiatry Research*, *238*, 189–195. https://doi.org/10.1016/j.psychres.2016.02.011

Nurhasanah, N. (2018). Pengembangan tes untuk mengukur kemampuan penalaran mahapeserta didik mata kuliah geometri. *Jurnal Pendidikan PEPATUDZU*, *14*(1), 62. https://doi.org/10.35329/fkip.v14i1.186

Pakaya, A. (2022). Meningkatkan kemampuan guru dalam menyusun tes hasil belajar melalui supervisi

akademik di sdn no. 93 kota utara kota gorontalo. *Jurnal Pendidikan Masyarakat Dan Pengabdian*, *02*(93), 931–938. http://ejurnal.pps.ung.ac.id/index.php/dikmas/article/view/1513

Permendikbud. (2022). Standar Penilaian Pendidikan Permendikbudristek No 21 tahun 2022. *Kementerian Pendidikan Dan Kebudayaan Riset Dan Teknologi*. https://www.gurusumedang.com/2022/06/standar-penilaian-pendidikan.html

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497. https://doi.org/10.1002/nur.20147.

Pop, C., & Khampirat, B. (2019). Self-assessment instrument to measure the competencies of Namibian graduates: Testing of validity and reliability. *Studies in Educational Evaluation*, *60*(December 2018), 130–139. https://doi.org/10.1016/j.stueduc.2018.12.004

Rahman, A. A., & Nasryah, C. E. (2019). Evaluasi Pembelajaran. In *Uwais Inspirasi Indonesia*.

Roldán-Merino, J., Farrés-Tarafa, M., Estrada-Masllorens, J. M., Hurtado-Pardos, B., Miguel-Ruiz, D., Nebot-Bergua, C., Insa-Calderon, E., Grané-Mascarell, N., Bande-Julian, D., Falcó-Pergueroles, A. M., Lluch-Canut, M. T., & Casas, I. (2019). Reliability and validity study of the Spanish adaptation of the "Creighton Simulation Evaluation Instrument (C-SEI)." *Nurse Education in Practice*, *35*(December 2018), 14–20. https://doi.org/10.1016/j.nepr.2018.12.007

Sahriani, S., Palloan P, A. A. (2019). Analisis kesesuaian kompetensi dasar dengan pembelajaran berbasis masalah. *Jurnal Sains Dan Pendidikan Fisika (JSPF)*, *15*(2), 10–17.

Sawitri, D., & Rahayu, E. M. (2018). *Modul PKT. 08 - Penilaian Hasil Belajar*. Kemenristek.

Shirali, G., Shekari, M., & Angali, K. A. (2018). Assessing reliability and validity of an instrument for measuring resilience safety culture in sociotechnical systems. *Safety and Health at Work*, *9*(3), 296–307. https://doi.org/10.1016/j.shaw.2017.07.010

Stiggins, R. J. (1994). *Student-centered classroom assessment*. Macmillan College Publishing Company.

Subagia, I. W., & Wiratma, I. G. L. (2016). Profil penilaian hasil belajar peserta didik berdasarkan kurikulum 2013. *JPI (Jurnal Pendidikan Indonesia)*, *5*(1), 39–54. https://doi.org/10.23887/jpi-undiksha.v5i1.8293

Sugiyono, S. (2012). *Statistika untuk penelitian*. Bumi Aksara.

Susetyo, B. (2015). *Prosedur penyusunan dan analisis tes untuk penilaian hasil belajar bidang kognitif*. PT Refika Aditama.

'ulum, M. (2016). *Uji Validitas dan Uji Reliabilitas* (1st ed.). www.stikeswch-malang.ac.id

Wahyudi, W., Arafah K., & Khaeruddin, K. (2018). Pengembangan instrumen berbasis keterampilan-keterampilan abad ke-21. *Seminar Nasional Fisika*, 1–4.

Winkel, W. S. (1996). *Psikologi pengajaran*. PT Grasindo.

Wulff, N. B., Møller, P. R., Christensen, K. B., Pedersen, S. G., Wessel, I., Dalton, S. O., & Homøe, P. (2021). The Voice-Related Quality of Life (V-RQOL) Instrument: Cross-cultural translation and test of validity and reliability of the danish version. *Journal of Voice*, *35*(5), 806.e7-806.e14.

https://doi.org/10.1016/j.jvoice.2020.01.010

Yusup, F. (2018). Uji validitas dan realibiltas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23. https://doi.org/10.21831/jorpres.v13i1.12884

Zainuddin, M., Syam, Y., & Yusuf, S. (2020). Translation of self-efficacy scale (SCSES) stoma care instruments Indonesia version: Test validity and reliability. *Enfermeria Clinica*, *30*, 439–443. https://doi.org/10.1016/j.enfcli.2019.07.133

Zhao, M., Rodriguez, M. A., Wang, B., Santa Ana, E. J., Friedberg, J., Fang, Y., Allegrante, J. P., & Natarajan, S. (2021). Validity and reliability of a short self-efficacy instrument for hypertension treatment adherence among adults with uncontrolled hypertension. *Patient Education and Counseling*, *104*(7), 1781–1788. https://doi.org/10.1016/j.pec.2020.12.029