



Jurnal

Penelitian dan Evaluasi Pendidikan



Jurnal

Penelitian dan Evaluasi Pendidikan

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA (HEPI)
in cooperation with
GRADUATE SCHOOL OF UNIVERSITAS NEGERI YOGYAKARTA
Kampus Karangmalang, Yogyakarta 55281. Phone. 0274 550836 Fax : 0274 520326

Website: <http://journal.uny.ac.id/index.php/jpep>
e-mail: jurnalhepi@uny.ac.id



9 772338 606001



9 772685 711007

Jurnal Penelitian dan Evaluasi Pendidikan

Volume 25, No 1, June 2021

Developing assessment in improving students' digital literacy skills
- - Erlida Ammie; Undang Rosidin; Kartini Herlina; Abdurrahman

Cognitive domain analysis (LOTS and HOTS) assessment instruments made by primary school teachers
- - Puji Hartini; Hari Setiadi; Ernawati

Character education strengthening model during learning from home: Ki Hajar Dewantara's scaffolding concept
- - Hasti Robiasih; Ari Setiawan; Hanandyo Dardjito

Developing the flipped learning instrument in an ESL context: The experts' perspective
- - Wahyu Hidayat; Mohammad Musab bin A. Ali; Nur Asmawati Lawabid; Mujabidah

Developing self and peer assessment to improve student's appreciative critical ability in learning drama appreciation
- - Khafidatur Rohmah; Endah Tri Priyatni; Heri Suwignyo

Evaluation of TOEFL preparation course program to improve students' test score
- - Mega Selvi Maharani; Nur Hidayanto Pancoro S. Putro

A psychometric evaluation of the career decision making self-efficacy scale
- - Chandra Yudistira Purnama; Linda Ernawati

Determinant factors affecting the improvement of education index
- - Jalil Setiawan Jamal; Muslim Salam; A. Nixia Tenriawaru; Didi Rukmana; Mubammad Hatta Jamil; Saadah

Applying Item Response Theory model for evaluating item and test properties of academic potential test for students with disability
- - Sukaesi Marianti; Dian Putri Permatasari; Unita Werdi Rahajeng

Analysis of mathematics test items quality for high school
- - Budi Manfaat; Ayu Nurazizah; Mubamad Ali Misri



Jurnal

Penelitian dan Evaluasi Pendidikan
ISSN 2685-7111 (print) | ISSN 2338-6061 (online)

Publisher

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA
In Cooperation With
PROGRAM PASCASARJANA UNIVERSITAS NEGERI YOGYAKARTA
(MOU Nomor 5835a/UN34.17/DN/2018)

Editor in Chief

Edi Istiyono, *Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta*

Associate Editor

Syukrul Hamdi, *Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta*

Editors

Siti Salina Mustakim, *Faculty of Educational Studies, Univesiti Putra Malaysia*

Heru Widiatmo, *American College Testing*

Wiel Veugelers, *Department of Education, University of Humanistic Studies*

Badrun Kartowagiran, *Faculty of Engineering, Universitas Negeri Yogyakarta*

Samsul Hadi, *Faculty of Engineering, Universitas Negeri Yogyakarta*

Heri Retnawati, *Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta*

Sudiyatno, *Faculty of Engineering, Universitas Negeri Yogyakarta*

Jailani, *Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta*

Undang Rosidin, *Faculty of Teacher Training and Educational Sciences, Universitas Lampung*

Wasis, *Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya*

Risky Setiawan, *Faculty of Social Sciences, Universitas Negeri Yogyakarta*

Alita Arifiana Anisa, *Lembaga Pendidikan dan Pengembangan Profesi Indonesia*

Correspondence:

Graduate School of Universitas Negeri Yogyakarta
Kampus Karangmalang, Yogyakarta, 55281, Telp. (0274) 550835, Fax. (0274) 520326

Homepage: <http://journal.uny.ac.id/index.php/jjep>

e-mail: jurnalhepi@uny.ac.id

FOREWORDS

We are very pleased that *Jurnal Penelitian dan Evaluasi Pendidikan* is releasing its issue **Volume 25, No 1, June 2021**. We are also very excited that the journal has been attracting papers from many institutions in Indonesia and many foreign countries. *Jurnal Penelitian dan Evaluasi Pendidikan* was first published in **1998** and since then regularly published online and in print twice a year: June and December.

Jurnal Penelitian dan Evaluasi Pendidikan with ISSN 2338-6061 (online) has been **re-accredited** by the Ministry of Research, Technology, and Higher Education of Republic of Indonesia under the Decree Number 30/E/KPT/2018 which is valid for 5 (five) years since enacted on 24 October 2018 (Vol. 20, No 2, 2016 until Vol. 25, No 2, 2021). *Jurnal Penelitian dan Evaluasi Pendidikan* successfully achieved accreditation in four periods in a row (in 2007, 2010, 2014, & 2018).

Jurnal Penelitian dan Evaluasi Pendidikan is a showcase of original, rigorously conducted educational evaluation, measurement and assessment from primary, secondary, and higher education institutions. Each issue of this journal is not limited to comprehensive syntheses of studies towards developing new understandings of educational evaluation, measurement and assessment only, but also explores scholarly analyses of issues and trends in the field.

Yogyakarta, June 2021

Editor in Chief

TABLE OF CONTENT

<i>Erlida Amnie, Undang Rosidin, Kartini Herlina, Abdurrahman</i>	Developing assessment in improving students' digital literacy skills	1-15
<i>Puji Hartini, Hari Setiadi, Ernawati</i>	Cognitive domain analysis (LOTS and HOTS) assessment instruments made by primary school teachers	16-24
<i>Hasti Robiasih, Ari Setiawan, Hanandyo Dardjito</i>	Character education strengthening model during learning from home: Ki Hajar Dewantara's scaffolding concept	25-34
<i>Wahyu Hidayat, Mohammad Musab bin A. Ali, Nur Asmawati Lawahid, Mujahidah</i>	Developing the flipped learning instrument in an ESL context: The experts' perspective	35-48
<i>Khafidatur Rohmah, Endah Tri Priyatni, Heri Suwignyo</i>	Developing self and peer assessment to improve student's appreciative critical ability in learning drama appreciation	49-62
<i>Mega Selvi Maharani, Nur Hidayanto Pancoro S. Putro</i>	Evaluation of TOEFL preparation course program to improve students' test score	63-76
<i>Chandra Yudistira Purnama, Linda Ernawati</i>	A psychometric evaluation of the career decision making self-efficacy scale	77-87
<i>Jalil Setiawan Jamal, Muslim Salam, A. Nixia Tenriawaru, Didi Rukmana, Muhammad Hatta Jamil, Saadab</i>	Determinant factors affecting the improvement of education index	88-96
<i>Sukaesi Marianti, Dian Putri Permatasari, Unita Wendi Rahajeng</i>	Applying Item Response Theory model for evaluating item and test properties of academic potential test for students with disability	97-107
<i>Budi Manfaat, Ayu Nurazizah, Muhamad Ali Misri</i>	Analysis of mathematics test items quality for high school	108-117

Developing assessment in improving students' digital literacy skills

Erlida Amnie*; Undang Rosidin; Kartini Herlina; Abdurrahman

Universitas Lampung

Jl. Prof. Dr. Ir. Sumantri Brojonegoro, Gedong Meneng, Rajabasa, Kota Bandar Lampung, Lampung
35141, Indonesia.

*Corresponding Author. E-mail: erlida.amnie@gmail.com

ARTICLE INFO

Article History

Submitted:

28 July 2020

Revised:

11 January 2021

Accepted:

14 January 2021

Keywords

assessment; problem
solving; digital literacy

Scan Me:



ABSTRACT

This research aims to develop an assessment of problem-solving skills in improving students' digital literacy skills. This research is directed to produce physics learning assessments that can improve students' digital literacy using problem-solving skills assessments. The use of assessments that refer to the problem-solving skills stage is expected to improve students' digital literacy. The study is an R & D research with the Borg and Gall development model. On the preliminary research, a questionnaire is needed to detect need analysis of an assessment that can help improving student digital literacy skills. Questionnaires were used to collect expert review data, while cognitive tests were used to collect data on students' problem-solving skills. Cognitive tests by posttest form were held to find out the progress of students understanding during the learning process using the products. The results of content validity by Aiken's V is 0.80. The factor analysis is seen by the Bartlett test value with Chi-squares = 1.604 and significance at 0.659. Therefore, the problem-solving skills assessment from the aspect of content and construction has valid criteria and is suitable for use. The N-Gain test results of students problem-solving skills in the experimental class by 0.3 with a quite effective category higher than the control class of 0.12 with a quite effective category. The results show that the use of assessment of problem-solving skills effectively improved students digital literacy skills.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



How to cite:

Amnie, E., Rosidin, U., Herlina, K., & Abdurrahman, A. (2021). Developing assessment in improving students' digital literacy skills. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 1-15.

doi:<https://doi.org/10.21831/jpep.v25i1.33600>

INTRODUCTION

Problem Solving Skills Assessment and Students Digital Literacy Skills

Outline assessment is one of the most important parts in teaching. Educators can determine the student skill and knowledge level through assessment (Taras, 2005). The existence of the assessment is one of the main components in the learning and evaluation process. Educators make the assessment as a means of measuring the understanding achievement and the student's skill level.

Assessment has an important role in education and a critical role in the teaching process (Tosuncuoglu, 2018). Assessment is a measuring tool for the achievement of learning targets in the scope of education. Assessment is one way to improve the quality of educational outcomes. The application of assessment in learning requires a renewal of the assessment used so it makes authentic assessment an alternative assessment carried out to measure students skills.

The alternative assessment has been presented as an alternative to standard testing and all problem topics found by the assessment. This assessment is different from the traditional assessment in its activities. Assessment directs students to show what they can do (Richards & Renandya, 2002). Authentic assessment is also known as an alternative assessment.



The purpose of authentic assessment is to measure students's skill in answering the assignment or test given in form the problems found in real life (Frey & Schmitt, 2007). Authentic assessment has the ability to measure students's skill trough learning process. Unfortunately, the teacher gets a difficulties to arrange the assessment. A research by Kartowagiran and Jaedun shows that only a few teachers have done authentic assessment because it must be addressed immediately, so that teachers can reveal students' real abilities.

Good authentic judgment will be able to improve the quality of learning and improve the quality of student performance, the more so, when the teacher provides feedback to the students (Kartowagiran & Jaedun, 2016). The use of assessments is expected to improve students' learning skills. Based on the questionnaire, it can be seen that 68% of students stated that the Physics teacher gave an assessment at the end of each learning chapter. It is just that the assessment conducted by the teacher, according to the students, is still dominant in the form of a written test. Although 73% of students are aware that Physics teachers use assessments, so that students become active and engaged in thinking activities, this is still a concern. As many as 65% of students stated that they only learned the material if the teacher had given an assignment or there would be an assessment. The assessment aspects observed were dominant in students' cognitive skills. The use of assessments that are relevant to students thinking activities is not yet diverse. Thinking skills that are part of the 2013 curriculum learning component in Indonesia have not yet been realized optimally. One part is the problem-solving skills.

Problem solving was first introduced by Heller et al. (1992). Through the article, it is introduced the problem solving strategies in guiding students to have thinking skills that are more complex than cognitive. Problem solving strategies are the strategies used by Heller et al. (1992) to bring up students problem solving skills, especially high school students in that research. In their research, it has investigated how diverse the problem solving performance of high school students is after applying problem solving strategies in a physics class cooperative group. The problem solving strategy with the contextual content of the problem uses five stages, including visualizing the problem; physics description, planning a solution; executing plan; checking and evaluating. The strategies make the students have the skills to solve a problem that they have to face by learning process.

The existence of integrated problem solving skills in an assessment can also bring up other aspects of skills for students, one of them is students' skills in literacy. Literacy is one of the skills to access, understand, integrate, communicate, evaluate, and compile information obtained based on the study of sources. Source studies can be done with various types of media, including printed media, electronic media, and digital media. One kind of literacy that can be explored by problem solving skills is digital literacy.

Digital literacy has been defined by various researchers and practitioners with the same meaning, but has different focuses (Son et al., 2017). As digital technology develops in the world, understanding of digital literacy is also expanding. The focus of the meaning of digital literacy is to be richer in digital media sources that are the source of literacy information.

Digital literacy is the skills to use technology and information from digital devices effectively and efficiently in various contexts such as academics, careers and everyday life (Gilster, 1997). In line with Gilster, according to Ng (2012), digital literacy is related to the variety of literacy associated with the use of digital technology. This technology is divided into hardware and software that individuals use for educational, social, and entertainment purposes at school and home.

Digital literacy is rooted. It is developed in the 1980s when microcomputers became more widespread. Information literacy was widespread in the 1990s, while the information was more easily compiled, accessed, and disseminated through networked information technology (Bawden, 2001). Digital literacy should be more than just the skills to use multiple digital sources effectively. Digital literacy is also a particular form of thinking (Eshet, 2002).

Digital literacy encompasses a wide variety of cognitive, motor, sociological and complex emotional skills that will be needed effectively in a digital environment (Eshet, 2004). Digital literacy is very closely related to the use of digital information technology. The use of supporting applications certainly becomes a necessity in realizing life skills of the current generation.

Problem Solving Skills Assessment in Improving Students Digital Literacy Skills

This research needs to be done in realizing the learning process with the aim of eliciting students high-level thinking skills. This research was conducted as a recommendation in order to achieve a better physics learning process. The 21st century skills observed throughout this study are the students' problem solving skills.

Problem solving skills are considered necessary to be owned by students, especially high school students, because these skills can help students make the right decisions, careful, systematic, logical, and consider various points of view. Students can carry out various activities without knowing the purpose and reason for doing so if they do not have these skills.

In general, the problem solving skills aspects consist of five indicators: students are able to define problems, examine problems, find solutions, implement plans that have been made, and evaluate (Novitasari et al., 2015). Referring to a research on problem solving strategy by Heller et al. (1992), the aspect of defining a problem is nothing but following the problem visualization (visualizing problem) stage, which is the stage where students can arrange problem statements in visual and verbal form in accordance with the condition of the problem.

The aspect of examining a problem is similar to the description stage (physics description), which is the stage that directs students to use the understanding they have to analyze the problem. The aspect of finding a solution is the same as the stage of planning a solution (planning a solution) where students begin to determine what solution is to be applied to the problem and arrange a plan in implementing the solution. The aspects of implementing a plan that have been created are similar to carrying out a plan (executing plan). Finally, the evaluation aspect follows the stages of checking and evaluating. One concept aspect of the problem solving skills used in learning today is the development of the problem solving strategy that Heller et al. (1992) have applied in their research.

The problem defining skills is student skills to determine what problems are the topic of observation. The problem checking is done by analyzing the problems that have been identified previously. Furthermore, students are directed to find solutions by referring to the results of problem analysis. The solutions that have been obtained are then planned and implemented. Students' skills in problem solving become intact after students evaluate the problem solving activities that have been carried out starting from determining the problem to the implementation of the solution.

In this study, increasing literacy skills with the assistance of digital media is the goal of developing students problem solving skills assessment assessments. This research is directed to produce physics learning assessments that can improve students digital literacy, namely by using problem solving skills assessments. The use of assessments that refer to the problem solving skills stage is expected to improve students digital literacy. The aspects of digital literacy skills that are considered during the learning implementation process by using the assessment of problem solving skills in this research are meaning making, analyzing, persona, using, decoding, creativity, operational skills, information skills, and ICT literacy.

In a complex system of class activities, assessments cannot be abstracted and used without reference to other components. Changes in assessment practice will certainly have an effect, either good or bad effect (James, 2015). Assessments always change in harmony with the changing atmosphere of teaching and learning in the post compulsory education period (Rust, 2002). The assessment relates to other components in the learning process. The development of the assessment that is used has supported the achievement of the educational targets to be

achieved. The implementation of the 2013 curriculum at all levels and types of education currently leads to the emergence of various assessment models as an informative medium for achieving student competencies.

Authentic assessment is one of the assessments that needs to be carried out along with the ineffectiveness of traditional assessments that have often been used. Traditional assessment is considered to ignore the real world context and does not adequately describe students skills holistically.

Table 1. Digital Literacy Rubric

No	Aspect	Dimensions
1	Meaning Making	Reading, Relating, Expressing (agency of the learner as a participant in the construction of the text; reflexive process in which content, style and purpose of the text is in dialogue with the prior experience knowledge and responses of the reader; implies both understanding and interpretation)
2	Analyzing	Deconstructing, Selecting, Interrogating (developing the skills to make informed judgements and choices in the digital domain; applying critical, aesthetic, and ethical perspectives to the production and consumption of digitized material)
3	Persona	Identity Building, Managing Reputation, Participating (sensitivity to the issues of reputation, identity and membership within different digital contexts; purposeful management and calibration of one's online presence; developing a sense of belonging and a confident participant role)
4	Using	Finding, Applying, Problem Solving, Creating (developing the skills to deploy digital tools appropriately and effectively for the task in hand; solving practical problems dynamically and flexibly as they arise using a range of methods and approaches both individually and as part of communities)
5	Decoding	Navigation, Conventions, Operations, Stylistics, Modalities (developing familiarity with the structures and conventions of digital media; sensitivity to the different modes at work within digital artifacts; confidently using the operational frameworks within which they exist)
(Hinrichsen & Coombs, 2014)		
6	Creativity	Generates content and constructs knowledge Publishes and peer reviews Exhibits creative thinking using digital
7	Operational Skills	Develops formal computer and internet skills and navigation and orientation skills
8	Information Skills	Identifies, accesses, manages, and transforms using online public services and applications
(Jimoyiannis, 2015)		
9	ICT Literacy	Adopts, adapts, and uses digital devices, applications, and services
(JISC, 2014)		

Authentic assessments direct students to become effective actors when gaining knowledge ([Wiggins, 1990](#)). Authentic assessment is expected to help the ongoing measurement of student learning achievement effectively from various aspects (knowledge, skills, and also attitudes). Through global assessments, it is hoped that the evaluation and decision making process will be more precise along with the precise improvement of the learning scenarios taken.

Assessment requires congruence on the learning point of view if it wants to be a valid, trusted assessment condition, to be applied to students ([James, 2008](#)). The use of assessment relates to targeted learning directions. Determination of the assessment can also consider the characteristics of the assessment to be used in learning.

Assessment can be supported by raising students' thinking skills. Problem solving skills assessment is an assessment that directs students to solve problems using what was previously understood. Problem solving skills will be unified if expressed in the form of assessments.

This assessment leads students to recall each material learned and apply it in everyday life. Problem solving skills assessment refers to the five stages of problem solving learning strategies proposed by Heller et al. (1992) and adjusting the latest stages based on a research carried out by Novitasari et al. (2015). Successful application of problem solving skills occurs when students successfully define problems, examine problems, plan solutions, carry out plans that have been made, and evaluate (Cullinane & Liston, 2011). Students are guided to get used to defining problems that exist during the learning process. Students are directed to examine the problem so that the skills to plan, implement, and evaluate solutions that have been stated previously was formed.

Since the existence of integrated problem solving skills in an assessment can also bring up students' skills in literacy, in this study, increasing literacy skills with the assistance of digital media is the goal of developing students' problem solving skills assessments. The indicators of achieving digital literacy skills of students refer to nine of the 20 aspects found in three different article sources. Table 1 presents an explanation of the aspects of digital literacy skills to be achieved. Thus, considering the aforementioned points and the previous related studies that have been conducted, this research aims to develop an assessment of problem-solving skills in improving students' digital literacy skills.

RESEARCH METHOD

This research was conducted to develop an assessment that guides problem solving skills in improving digital literacy of students in learning physics. This research uses research and development (R & D) method. The development of the problem solving skills assessment was carried out considering two aspects. The first aspect to note is that there are significant differences in the problem solving skills of students before and after using problem solving skills assessment. The second aspect of using problem solving skills assessment is that there are significant differences in students digital literacy skills before and after using problem solving skills assessments.

Development is carried out with the development model of Borg and Gall (2003). The Borg & Gall Model consists of ten stages of activities, namely: (1) research and data collection; (2) planning; (3) initial product development; (4) initial field trials; (5) revision of trial results; (6) field trials; (7) improvement of product yields; (8) field implementation test; (9) improvement of the final product; (10) dissemination and implementation. The implementation of assessment activities is carried out by the teacher following the stages that have been prepared in the assessment.

Firstly, on the research and data collection, the need analysis is a way to know what the need for improving students' digital literacy skills. In the initial stage, the researchers conduct a needs analysis to explore the problems that exist in the school, namely about what students and teachers need. Needs analysis data were taken from students and teachers using a questionnaire. This was done to determine the needs of teachers and students for problem solving skills assessment in improving student digital literacy. The material to be taught was analyzed according to the core competent and basic competent studies and student needs. Then, the results of the analysis are used as a basis in preparing the background of the problem.

At this stage, the first step was to determine the type of assessment used, taking into account the strengths and weaknesses and adjusting the needs of students. The material used as literacy material is material that has a lot of information on the development of study material on the internet as a source of digital literacy.

On the planning stage, the problem solving skills assessment was targeted as an assessment used for improving students' digital literacy skills. Meanwhile, the assessment has to bring up students' problem solving skills in a time. The problem solving skills are the skill that can be related to students' digital literacy skills.

On the initial product development, a draft I of the product was arranged, which is an assessment referring to problem solving skills to improve student digital literacy. The product draft was then validated by expert as the initial fields trials. The initial product validation process was done by validation questionnaire. Three expert lecturers and a teacher checked the product validity. Product development validation was focused on content and design validation. Then, a one-on-one test was done to determine the product validity. The result of the assessment was obtained by calculating the total score through the validation questionnaire.

The assessment has to be tested about content validity and construct validity. Content validity can be proved by using CVR and CVI, known as V Aiken Coefficient. [Lawshe \(1975\)](#) proposed content validity ratio (CVR) to measure the degree of agreement between experts of one item and which can express the level of content validity through a single indicator that ranges from -1 to 1. The CVR formula is presented in Formula (1), where ne = the number of SME (Subject Matter Experts) that assess an item as 'essential', and n = the number of SME who conducted the assessment ([Aiken, 1985](#)).

$$CVR = \frac{2ne}{n} - 1 \quad (1)$$

Content validity can be seen by validity standard through rater frequency and the scale ([Aiken, 1985](#)). The CVR value ranges from -1 to 1. If half of the SMEs are essential, the CVR would be 0. CVR would be worth 1 if all SMEs are essential for an item. Overall, the validity value of the test can be determined using the CVI (Content Validity Index), with the formula presented in Formula (2), where CVR = Content Validity Ratio of each item, and k = the number of items ([Aiken, 1985](#)).

$$CVI = \frac{\sum CVR}{k} \quad (2)$$

The number of rating categories affects the content validity standard set by Aiken. The smallest rate of each categories formulated by Aiken is two and the highest is seven ([Aiken, 1985](#)). This research uses four rating categories and three raters. Content validity analysis using the Aiken V coefficient was used to test the validity of the observation sheet instrument, with the content validity introduced by Lawshe to test the content validity of the test instrument.

The Aiken validity coefficient is calculated using the raw score of n experts, while the content validity using the V Aiken coefficient by Formula (3), where r = the score given by the rater, l_0 = the lowest rate, c = the highest rate, n = the number of raters, and i = integers from 1, 2, 3 to n ([Aiken, 1985](#)).

$$V = \frac{\sum(r_i - l_0)}{[n(c - 1)]} \quad (3)$$

The validity and reliability of the constructs of the indicators (items) formed latent constructs by conducting Confirmatory Factor Analysis (CFA) ([Latan, 2012](#)). The construct validity was tested by using KMO and Bartlett's test of sphericity.

Based on the content and construct validity, the product from the trial results needs to be revised in terms of its content. The revised product was named as draft II. The next stage was field trials, on which a Physics teacher was given the role to analyze the product before it was tried out in the classroom. The product was revised again as draft III in the improvement of product yields stage. After going through an assessment by a construction expert and material, the next stage involved class XI students as research subjects in the field implementation test. During the learning process, the final product improvement stage was carried out simultaneously with the implementation of learning using the revised product.

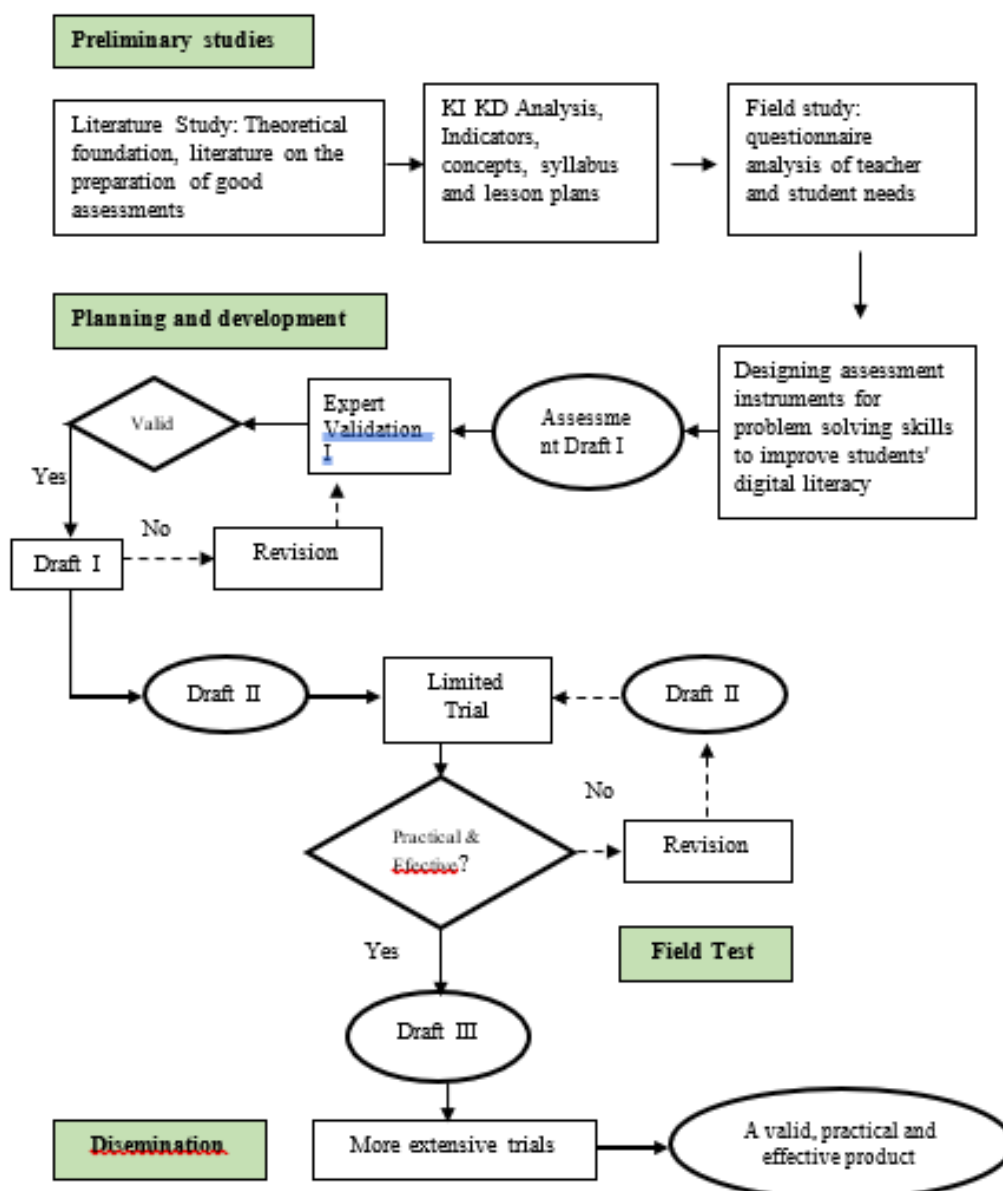


Figure 1. Research Flow Chart

The implementation phase of the research was carried out by following the development model of Borg and Gall (2003), where then the ten steps were grouped into four stages by making adjustments, namely a preliminary study; planning and development; field test; and dissemination. The steps in conducting this research can be seen in Figure 1.

FINDINGS AND DISCUSSION

This research was conducted by developing a product that will be used in learning process. The product is students' problem skills assessment. The assessment uses the student problem solving skills as a way to improve students' digital literacy skills on each stage of problem solving skills. The research was carried out using a self-assessment that gave rise to five stages of problem solving skills. Students are evaluated on what has been achieved and understood in learning. Assessment by focusing on students skills to remember and conclude what is being analyzed with prior understanding. This assessment is realized in class activities by involving reflection activities and adjusting previously owned concepts.

Through alternative assessments, learning becomes more authentic because it helps improve students' decision-making skills and problem solving skills (Timmins, 1996). Authentic assessment result data includes information on the strengths and weaknesses of students during the learning process. Data collection generally uses assignments that are close to real day-to-day activities. Authentic assessment helps educators gather views on the effectiveness of learning towards students and make changes to further learning if needed.

The implementation of assessment activities is carried out by the teacher following the stages that have been prepared in the assessment. Before carrying out the assessment activities, the teacher observes students. Observations were made on students in the experimental class and the control class. The teacher pays attention to the students initial conditions before starting learning. Observations are also made by giving several opening questions (pre-test) to students. The results of the initial observation data and pre-test become the initial baseline report before conducting treatment in learning.

Enforcement of learning by using the assessment was carried out in the experimental class, while in the control class is applied as usual learning. In the control class, students use books as a source of information when they want to carry out learning activities.

The students in the experimental class carry out self-assessment activities of problem solving skills by using video learning media that are prepared beforehand the site address and several sources of articles that have been checked before the accuracy of the information on the site. When learning takes place, the teacher makes observations again to measure the extent to which activities that bring up the problem solving skills are carried out among students. The measured problem solving skills of students are used by assisting digital literacy of students, namely by using information sources and learning videos obtained from the internet.

In the aspect of gathering information about a problem, students are directed to gather information with the help of the internet. The teacher provides a link option for digital information sources that students can use to collect information and students are given the opportunity to look for additional information independently from other sources. The process of collecting digital-assisted information is carried out to realize meaning making aspects for students. Students who can collect information about problems through digital information well can be concluded that they can compose meaning in digital information as well. In the aspect of examining problems and compliance with the principles of fluid dynamics that have been studied using information obtained digitally, the teacher observes the students skills to trace and associate information that has been obtained digitally with the problems that are the topic of discussion during learning. In the aspect of finding digital-assisted solutions, the teacher observes students attitudes when conducting a library review of digital information sources that have been obtained previously. In this aspect, the teacher also observes the students to discuss digital information obtained with other students so that a more robust literature study analysis of the digital information sources that has been obtained previously will be formed.

The teacher also observes nine aspects of students digital literacy skills (making meaning, analyzing, persona, use, decoding, creativity, operational skills, information skills, ICT literacy). The observation process was carried out using observation instruments of student behavior by loading nine aspects of the digital literacy skills.

The posttest assessment process took place in both classes using the same instrument at the end of the lesson. The results of observations and tests that have been given are taken into consideration in measuring the improvement of students digital literacy skills in the experimental class.

This research involves class XI students as research subjects, consisting of one class as an experimental class and one control class. The experimental class was treated with learning by using problem solving skills assessment, while the control class carried out learning using books and worksheets as the main source of information throughout the learning process. The design of this study uses the pretest-posttest design from Fraenkel and Wallen (2012), as

shown in Table 2, where E = experimental class, C = control class, Y1 = pretest on experimental class, Y2 = posttest on experimental class, X1 = learning with problem solving assessment, X2 = learning without problem solving assessment, Y3 = pretest on control class, and Y4 = posttest on control class.

Table 2. Pretest-Posttest Design

Group	Pretest	Dependent Variable	Posttest
E	Y ₁	X ₁	Y ₂
C	Y ₃	X ₂	Y ₄

The assessment that was developed was problem solving skills assessment in the Fluid Dynamics subtopic which included the task of tracking the cause of the aircraft being unable to fly through the internet media. The assessment contains learning scenarios, grids, instrument shapes (observation sheets), instrument rubrics, and scoring guidelines to obtain final scores on aspects of students problem solving skills.

The preparation of problem solving skills assessment begins with determining the objectives of the preparation of the assessment, the assessment lattice, the shape and format of the assessment, and the scoring guidelines. Initial testing of the assessment is carried out by initial testing of product design on a limited scale, namely the expert validation test. Lawshe (1975) stated that the minimum recommended number of evaluators is 5, and in order for the item to be accepted, all the raters must admit that the item is essential.

Based on the research, the number of raters for the validation assessment was three, and all raters stated that the item is essential. Based on Formula (1), CVR index from the assessment is 1. CVR of each item is 1 and there are 13 items to assess. Based on Formula (2), Content Validity Index is 1. The CVI value obtained from the average CVR is 1. Based on the CVR value that exceeds 0.99, all items are declared valid (Lawshe, 1975, p. 568) and are suitable for use for further research.

Table 3. Content Validity Analysis

Item	Rater	Aiken's V
1	3	0.89
2	3	0.67
3	3	0.67
4	3	0.78
5	3	0.78
6	3	1
7	3	0.89
8	3	0.67
9	3	0.78
10	3	0.78
11	3	0.89
12	3	0.78
13	3	0.89

The content validity was analyzed by Aiken's V for each item, in which the result can be seen in Table 3. Based on Table 3, the results of the validation of each validator can be said to be very high. The average of content validation by Aiken's V is 0.80. The construct validity is analyzed with SPSS 25 by using KMO and Bartlett's Test.

Based on Table 4, the results of the SPSS output display show that the value of KMO = 0.517, so that it means that the factor analysis can be carried out. The Bartlett test value with Chi-squares = 1.604 and significant at 0.659, so it can be concluded that the factor analysis test can be continued.

Table 4. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.517
Bartlett's Test of Sphericity	Approx. Chi-Square	1.604
	df	3
	Sig.	.659

The assessment can be continued to be used with a slight improvement in the learning scenario section. Previous learning scenarios contain questions that must be answered by students but are not included in the column to answer the question but because it is felt to be incorrect, each question is included with a place to answer the question. It aims to make students more challenged to explore their answers.

The number of aspects of the observation that originally contained 18 items was reduced to 12 items because if a teacher uses a problem solving skills assessment that contains 18 assessment items for one student then the teacher will run out of time to assess all students in the class. This is considered ineffective, so the researchers make 12 items on the observation aspect. The researchers improve the scale used in the scoring guidelines and recapitulates the final grade, rubric, writing system, and language as suggested. Overall, on the assessment problem solving skills development results is feasible in terms of construction.

The assessment is suitable to be used after being revised according to the validator's direction. Revised improvements based on the three validator's suggestions for the problem solving skills assessment, which is to add an observation link source or observation that will be done through a video tutorial, clarify the description of the stages of problem solving skills that students want to do, and provide clearer details about the assessment rubric on the assessment. The revision product the known as draft II.

After the product is revised, a field test is then performed. This field test is given to physics subject teachers. This trial was conducted with the aim to determine the validity, effectiveness, and practicality of the use of assessments felt by the user, namely the teacher.

The use of problem solving skills assessment by involving information retrieval using digital sources that have been done also pay attention to aspects of students digital literacy skills at the time of the assessment. Problem solving skills assessment is used in an experimental class by the subject teacher. The teacher does the apperception and presentation of learning material related to the use of the assessment. Use of the assessment is carried out at the second meeting after the teacher ensures students understanding of the material.

Students read and understand the preliminary activities section and description of the first activity in eliciting problem solving skills, namely defining the problem. The students can move to the second stage if they have followed the steps in the first stage, one of them, they are asked to open sources of learning videos and articles related to the main problem.

Students can understand and follow the instructions in the steps easily. The main problem lies in the implementation of the stages in the assessment. In the first stage, the average student takes approximately 40 minutes (one hour of learning) to identify problems based on article searches and results of watching videos.

The implementation of the first stage also has obstacles for some students. The first stage requires internet activation as a medium for information retrieval. Some students claimed not to have a quota to conduct their own searches. Finally, some students did a group search. The information retrieval process carried out is also far from good and long lasting.

Some female students find it hard to watch learning videos and are lazy to read the source of the article provided. They did not like studying Physics from the beginning and already felt dizzy first when studying Physics. In this condition, the teacher just let it go because most students have been able to follow the process of finding information in the first stage.

Students who have completed the activity in the first stage then proceed to stage two, examining the problem. At this stage, each writes the results of the previous search. When writing down the cause of a plane not flying suddenly, there are some students who repeat the learning video that has been watched previously. Some students also reopen articles that have been searched and some students rely on the results of a one-time search.

During stages one to three, the teacher acts as a facilitator and makes a thorough observation of each student's behavior. The teacher makes observations about aspects of problem solving skills and aspects of students digital literacy skills during the activity. Observations are made by referring to the observation instruments that have been prepared beforehand.

In the fourth stage, students are asked in groups in one bench to discuss the results of each writing from stages two and three. During the discussion, each group draws a plan that is appropriate for the solution of the problem raised. In the final stage of the activity, all students are invited to discuss by describing the results of the discussion on the bench to the search results in stages one to three. At this stage, all students conduct a global check and evaluation of the problems that have been raised.

At the end of the lesson, the teacher gives a posttest to measure student understanding related to the material that corresponds to the problem. In the posttest, the teacher included five questions related to students responses to the assessments and learning activities at that time. Students are also given one question space to write suggestions and hopes for future physics learning. All students understand the stages that exist in the learning process. Students are just not used to it because learning activities involve using the internet directly in class with a time limit. Some students better understand the concepts of physics through simple learning video media that are searched online. Some other students experience technical difficulties because they have not become routine learning activities in the classroom.

The results of the analysis of the experimental class using the stages of digital literacy are the planning stages carried out by students on average by 77.2%, the implementation stage by an average of 75%, and the evaluation stage by an average of 81%. The analysis of the posttest and pretest results are listed in Table 5.

Table 5. T-Test of Pretest Posttest for Control Class

Data	Pretest	Posttest
N	30	30
Lowest Score	17	26
Highest Score	41	44
Average	33.13	34.35
T		-1.033
Sig. 2- tailed		0.120

The lowest score for the control class at the time of the pretest was 17 and the highest score was 41. For the posttest, the lowest score was 26 and the highest score was 44. The results of the T test show sig. $0.120 > 0.05$, so H_0 is accepted, meaning that the average of the two populations is the same or there is no difference between the pretest and posttest. The results of the T-test in the experimental class are summarized in Table 6.

Table 6. T-Test of Pretest Posttest for Experimental Class

Data	Pretest	Posttest
N	30	30
Lowest Score	26	58
Highest Score	42	82
Average	33.57	70.70
T		-1.033
Sig. 2- tailed		0.000

Based on Table 6, the lowest score for the experimental class at the time of the pretest was 26 and the highest score was 42. For the posttest, the lowest score was 58 and the highest score was 82. The results of paired sample t-test analysis of the experimental class shows sig. 2-tailed $0.000 < 0.05$ in the experimental class, so it can be concluded that H_0 is rejected.

This means that the average of the two populations is different or there is a difference between pretest and posttest. Based on the aforementioned results of analysis, there is no difference between the pretest and posttest in the control class, while the experimental class shows the difference between the pretest and posttest. The N-Gain test results of students problem-solving skills in the experimental class by 0.3 with a quite effective category higher than the control class of 0.12 with a quite effective category.

Problem solving skills assessment using aspects of digital literacy skills is a way of integrating digital literacy and problem solving in a single learning activity. Through this assessment, students are directed to be skilled in finding and compiling information that will be used to solve some of the problems that have been raised by the teacher previously. Based on the description of the practicality and effectiveness of using the previous assessment of problem solving skills, it is known that students' digital literacy skills have increased, although slowly.

This is in accordance the statement of [LINC Regional Professional Development Center for Adult Education \(2015\)](#) that integrating digital literacy and problem solving in an instruction (learning activities) can accelerate the student learning process by increasing the use of technology to solve problems raised during learning and will certainly increase students' digital literacy and access. Thus, it can be concluded that the instrument of problem solving skills assessment can help students improve digital literacy so that the learning outcomes of students also increase.

The digital literacy referred to in this research is the learning independence of students using the help of digital information in finding a solution to the problems raised by during the learning process. The aspects that cover digital literacy skills include planning, implementation and evaluation.

In the control class, learning is carried out in a conventional way and in the experimental class, learning is carried out using the problem solving skills assessment instrument. The results of the pretest and posttest showed that in the control class there was no significant increase in scores between the pretest and posttest. The technology used in classroom learning can actually help carry out learning activities better. During the process of learning activities using problem solving skills assessment and covering aspects of digital literacy activities, students are initially directed to use their technology as a source of fulfilling the information needed in learning. This is still felt very slowly during the learning process.

Along with the increasingly skilled students in using technology, it can be ascertained that students' skills to understand learning and digital (information) literacy skills will increase. This is in accordance with the research conducted by [Katz et al. \(2008\)](#) which states that students are able to understand learning materials better and faster along with the skillful use of information technology (digital literacy), are able to generate problem solving during learning, and are able to make better learning process.

Based on the trial results shown in Figure 2, the assessment validity level was 86.67%, the assessment effectiveness level was 71.33%, and the practicality level of the assessment was 71.33%. These results can be said that the level of suitability of the assessment is high, the level of ease of the assessment is high, the level of instrument usability and the level of practicality of the assessment is high.

The assessment is considered to be very practical because the aspects of the skills observed in the assessment are already quite practical in use, meaning that the number of skill aspects in the assessment is not too much but has been able to cover all aspects needed in the assessment therefore the time used in conducting the assessment using the assessment is also not too long.

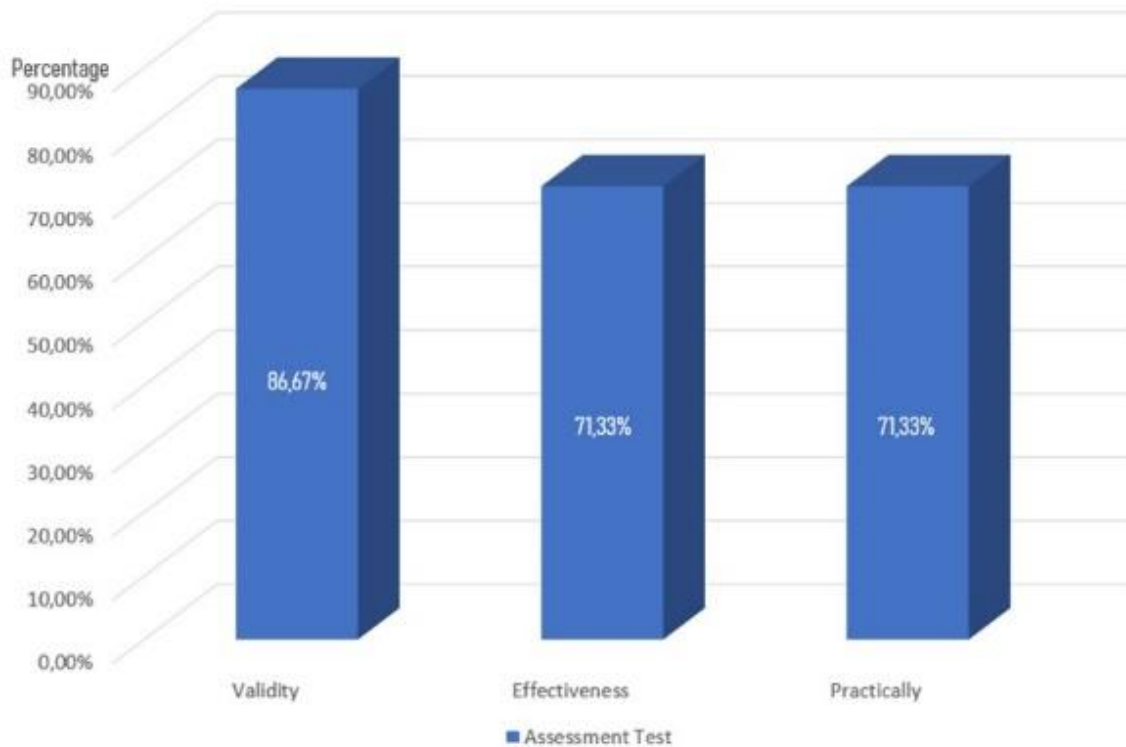


Figure 2. Assessment Test

The scoring aspect of the assessment is also quite practical, because the rubric used is also easy to understand so the assessment is practical to use. The scope of contents and also the design of the assessment is quite practical. The sequential instrument design of the grid, the shape of the assessment, and the assessment rubric make this instrument practical to use and also easy to administer. The problem solving skills assessment is also practical in assessing digital literacy because aspects of digital literacy have been included in this instrument.

The results of the effectiveness test are based on Figure 2 when converted to 71.3%, which means it is very appropriate. The effectiveness of the use of assessments for problem solving skills in improving digital literacy can be determined by comparing learning outcomes or the pretest-posttest scores of the control class and the experimental class. The average pretest score in the control class was 33.13 and the average posttest score was 34.35, while the average pretest score in the experimental class was 33.57 and the average posttest score was 70.70.

CONCLUSION

The average of content validation by using Aiken's V is 0.80. The construct validity is analyzed with SPSS 25 by using KMO and Bartlett's Test. The results of the SPSS output display show that the value of KMO = 0.517, so it means that the factor analysis can be carried out. The Bartlett test value with Chi-squares = 1.604 and significance at 0.659, so it can be concluded that the factor analysis test can be continued.

The lowest score for the control class at the pretest was 17 and the highest score was 41. Besides, for the posttest, the lowest score was 26 and the highest score was 44. The results of the T test show sig. 0.120 > 0.05, so it can be concluded that H_0 is accepted. The lowest score for the experimental class at the pretest was 26 and the highest score was 42, while for the posttest, the lowest score was 58 and the highest score was 82. The results of the T test show sig. 0.120 < 0.05, so it can be concluded that H_0 is accepted.

At the product revision stage based on the use trial, the appraiser responded that this assessment still took a long time because the assessor had to observe each student according to the aspects contained in the assessment rubric. The stage that is assessed cannot be done in one learning process. Each meeting is only able to assess one or two stages of problem solving activities that integrate digital literacy skills. This is one of the reasons for the effectiveness of the use of the assessment developed quite effectively. Teachers and students are not familiar with the assessment of problem solving skills that are developed by integrating these digital literacy skills properly.

REFERENCES

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Bawden, D. (2001). Information and digital literacies: A review of concepts. *Journal of Documentation*, 57(2), 218–259. <https://doi.org/10.1108/EUM000000007083>
- Borg, W. R., & Gall, M. D. (2003). *Educational research: An introduction*. Longman.
- Cullinane, A., & Liston, M. (2011). *Two-tier multiple choice questions: An alternative method of formative assessment for first year undergraduate Biology students*. Limerick: National Center for Excellence in Mathematics and Education Science Teaching and Learning (NCE-MSTL).
- Eshet, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Educational Multimedia and Hypermedia*, 13(1), 93–106. <https://www.learntechlib.org/primary/p/4793/>
- Eshet, Y. (2002). Digital literacy: A new terminology framework and its application to the design of meaningful technology-based learning environments. In P. Barker & S. Rebelsky (Eds.), *Proceedings of ED-MEDIA 2002--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 493–498). Association for the Advancement of Computing in Education (AACE). <https://www.learntechlib.org/primary/p/10316/>
- Fraenkel, J., & Wallen, N. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill Higher Education.
- Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics*, 18(3), 402–423. <https://doi.org/10.4219/jaa-2007-495>
- Gilster, P. (1997). *Digital literacy*. Wiley Computer Pub.
- Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American Journal of Physics*, 60(7), 627–636. <https://doi.org/10.1119/1.17117>
- Hinrichsen, J., & Coombs, A. (2014). The five resources of critical digital literacy: A framework for curriculum integration. *Research in Learning Technology*, 21, 21334–21350. <https://doi.org/10.3402/rlt.v21.21334>
- James, M. (2008). Assessment and learning. In S. Swaffield (Ed.), *Unlocking assessment: Understanding for reflection and application* (1st ed., pp. 20–36). Routledge.
- James, M. (2015). Educational assessment: Overview. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 161–171). Elsevier.

- Jimoyiannis, A. (2015). Digital literacy and adult learners. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (pp. 213–216). SAGE Publication.
- JISC. (2014). *Developing digital literacies*. Joint Information Systems Committee. <https://www.jisc.ac.uk/full-guide/developing-digital-literacies>
- Kartowagiran, B., & Jaedun, A. (2016). Model asesmen autentik untuk menilai hasil belajar siswa sekolah menengah pertama (SMP): Implementasi asesmen autentik di SMP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 20(2), 131–141. <https://doi.org/10.21831/pep.v20i2.10063>
- Katz, I. R., Elliot, N., Attali, Y., Scharf, D., Powers, D., Huey, H., Joshi, K., & Briller, V. (2008). The assessment of information literacy: A case study. *ETS Research Report Series*, 2008(1), 1–34. <https://doi.org/10.1002/j.2333-8504.2008.tb02119.x>
- Latan, H. (2012). *Structural Equation Modeling: Konsep dan aplikasi menggunakan LISREL 8,80*. Alfabeta.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- LINCS Regional Professional Development Center for Adult Education. (2015). *Integrating digital literacy and problem solving into instruction*. LINCS Regional Professional Development Center for Adult Education. <https://lincs.ed.gov/professional-development/resource-collections/profile-820>
- Ng, W. (2012). Can we teach digital natives digital literacy? *Computers & Education*, 59(3), 1065–1078. <https://doi.org/10.1016/j.compedu.2012.04.016>
- Novitasari, N., Ramli, M., & Maridi, M. (2015). Penyusunan assessment problem solving skills untuk siswa SMA pada materi Lingkungan. *Prosiding Seminar Nasional XII Pendidikan Biologi FKIP UNS 2015: Biologi, Sains, Lingkungan, Dan Pembelajarannya*, 519–525.
- Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching: An anthology of current practice*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511667190>
- Rust, C. (2002). The impact of assessment on student learning: How can the research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices? *Active Learning in Higher Education*, 3(2), 145–158. <https://doi.org/10.1177/1469787402003002004>
- Son, J.-B., Park, S.-S., & Park, M. (2017). Digital literacy of language learners in two different contexts. *The JALT CALL Journal*, 13(2), 77–96. <https://doi.org/10.29140/jaltcall.v13n2.213>
- Taras, M. (2005). Assessment – Summative and formative – Some theoretical reflections. *British Journal of Educational Studies*, 53(4), 466–478. <https://doi.org/10.1111/j.1467-8527.2005.00307.x>
- Timmins, A. C. B. (1996). Multiple intelligences: Gardner's theory. *Practical Assessment, Research, and Evaluation*, 5(1). <https://scholarworks.umass.edu/pare/vol5/iss1/10>
- Tosuncuoglu, I. (2018). Importance of assessment in ELT. *Journal of Education and Training Studies*, 6(9), 163–167. <https://doi.org/10.11114/jets.v6i9.3443>
- Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, 2(1), 1–6. <https://doi.org/10.7275/ffb1-mm19>

Cognitive domain analysis (LOTS and HOTS) assessment instruments made by primary school teachers

Puji Hartini; Hari Setiadi; Ernawati*

Universitas Muhammadiyah Prof. Dr. HAMKA

Jl. Limau II, Kramat Pela, Kebayoran Baru, Kota Jakarta Selatan, Jakarta 12130, Indonesia.

*Corresponding Author. E-mail: ernawati.pep@uhamka.ac.id

ARTICLE INFO

Article History

Submitted:

12 September 2020

Revised:

23 January 2021

Accepted:

25 January 2021

Keywords

cognitive domain analysis;
assessment instruments;
science

Scan Me:



ABSTRACT

This research aims to qualitatively analyze the validity of items and the suitability of cognitive domains (LOTS and HOTS) assessment instruments on natural science subjects made by elementary school teachers in East Jakarta. The method used is descriptive qualitative method with analysis of observations, documents in the form of teacher-made assessment instruments, interviews, and results of expert validation which are analyzed by comparison analysis techniques. The observation results show that all schools use the questions that are available in textbooks owned by students for assessment and the results of analysis of teacher-made assessment instruments validated by experts, there are 81.25% items included in the LOTS category, while 18.75% are included in the HOTS category, so it can be concluded that: (1) the instruments used by elementary school teachers in East Jakarta have fulfilled the content validity, (2) the cognitive domain (LOTS & HOTS) on the instruments used by elementary school teachers are proportional, (3) the quality of assessments conducted by elementary school teachers in East Jakarta is good with a record of improvement, (4) the implementation of assessments conducted by elementary school teachers in East Jakarta has followed the assessment standards provided by the government.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



How to cite:

Hartini, P., Setiadi, H., & Ernawati, E. (2021). Cognitive domain analysis (LOTS and HOTS) assessment instruments made by primary school teachers. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 16-24.

doi:<https://doi.org/10.21831/pep.v25i1.34411>

INTRODUCTION

The competency standards that students at the basic level must have are listed in the [Regulation of the Minister of Education and Culture No. 21 of 2016](#) concerning the content standards for primary and secondary education. The skills that must be possessed are the thinking and acting skills, including creative, productive, critical, independent, collaborative, and communicative in clear, systematic, logical and critical language, in aesthetic works, movements that reflect healthy children, and actions that reflect children's behavior according to their developmental stage.

[Regulation of the Minister of Education and Culture No. 20 of 2017](#) concerning the team of performance assessor in the Ministry of Education and Culture and [Regulation of the Minister of Education and Culture No. 23 of 2016](#) concerning education assessment standards have mandated that learning outcomes assessment by educators aims to monitor the process to improve the effectiveness of learning. Entering the 21st century is marked by the rapid advancement of technology in various fields of everyday life which triggers a demand for an era that continues to move from various aspects of life, one of them is in terms of education.

The world of education continues to receive the spotlight, especially regarding success, even the effective process of learning. In the process, an educator is highly required to be able

to prepare varied and innovative learning facilities so that it is hoped that education will be able to motivate students to learn and achieve so that they can improve their personal, school and education qualities globally. This can be seen from the results of the effective assessment by educators of the learning process being carried out.

The learning process will run well, if all readiness is done well by the teacher, including the way of assessment to see the level of success and student achievement. The ability to think at a higher level will be truly measurable, if you use the right measuring tool or instrument so that it needs to be considered in the preparation of the instrument, with the aim of maximizing the expected achievement. The change in the education climate is expected to be able to produce future development candidates who are competent, independent, critical, intelligent, creative and ready to face various kinds of challenges (Mulyasa, 2017).

According to Mulyasa (2017), in the implementation of the 2013 curriculum the learning process is required to apply the HOTS-based learning concept with various innovative, creative, collaborative, problem-based and problem solving learning models. This means that the learning that has been implemented has made it possible to apply HOTS-based assessments.

Assessment, according to the Regulation of the Minister of Education and Culture No. 23 of 2016, is the process of collecting and processing information to measure the achievement of student learning outcomes. The results of this measurement process will be used as a reference for the success rate of learning. Meanwhile, education assessment standards in the scope of education assessment in primary and secondary education consist of an assessment of learning outcomes by educators; assessment of learning outcomes by educational units; and assessment of learning outcomes by the Government.

Marzano and Pickering (1997) in Setiawati et al. (2019) explain that in the dimensions of how to think and act, students are directed to have the ability to think critically, creatively and self-regulate in thinking. These learning processes are oriented towards the quality of education. One of the ways to improve the quality of students is through improving the quality of learning that is oriented towards higher order thinking skills. The quality of learning also needs to be measured by an assessment that is oriented towards higher order thinking skills (HOTS).

Based on the results of the 2019 National Examination, the Ministry of Education and Culture's center for educational assessment explains that students are still weak in higher order thinking skills, such as reasoning, analyzing, and evaluating so it is necessary for teachers to be able to carry out HOTS-based assessments so that students are familiar with the questions and learning oriented to higher order thinking skills in order to encourage critical thinking skills. This can be applied in daily assessments in learning.

The realization of an appropriate assessment is inseparable from the quality of the form of the instrument used to measure students' higher order thinking skills. In connection with the demands of the era regarding the ability to think at a high level as a step to prepare a golden generation in 2045, an educator is expected to be able to become the main force spearhead to achieve these national goals.

Regulation of the Minister of Education and Culture No. 104 of 2014 describes an assessment instrument, which is a measuring tool used to assess the learning outcomes of ideal participants with a test and attitude scale. Amirono and Daryanto (2016) explain that assessment is the application of various methods and the use of various assessment tools to obtain information about the extent to which students' learning outcomes or the achievement of students' competencies (series of abilities). It is also explained by Angelo (1991) in Amirono and Daryanto (2016) that assessment is a simple method that can use faculty (schools) to collect feedback, early and after, on how well their students learn what they teach. Then, Kurniasih and Berlin (2016) explain that assessment is a step taken for, as, and for learning. Boyer and Ewel in Amirono and Daryanto (2016) define assessment as a process that provides information about individual learners about curriculum or programs, about institutions, or everything related to the institutional system.

Based on the aforementioned experts' opinion, it can be concluded that the instrument is a measuring tool for assessing, while the assessment is the process of interpreting the measurement result data which is used to determine the level of achievement of the learning process. Thus, the instrument is a tool used to measure a measuring object with the process of data collection and data interpretation.

Amiriono and Daryanto (2016) explain that a test is considered valid if it can accurately measure what it should measure. Arikunto (2011) states that validity is the ability of a measuring instrument to measure its measuring target. It is supported by Maolani and Cahyana (2015) that:

Validity is a quality that shows the suitability of the measuring instrument with the objectives to be measured/what should be measured. The validity that is meant is the validity of the assessment used by the teacher in the learning process that has been designed to be implemented so that learning achievement can be known.

Based on the aforementioned experts' opinions, it can be concluded that the validity of the assessment is the quality of the measuring instrument for the assessment to be measured in accordance with the learning objectives.

Amiriono and Daryanto (2016) suggest that content validity indicates a condition of an instrument arranged based on the content of the subject matter being evaluated which is structured to measure the specific objectives of the given subject matter. It is in line with the opinion of Yusup (2018) regarding the validity of content which focuses on providing evidence on the elements that exist in measuring instruments and is processed with rational analysis so that the assessment will be easier to do. Suryanto and Sutinah (2011) restate that content validity is needed to answer the question to what extent the items in the test can measure the overall material that has been taught. Some examples of the elements that are assessed in content validity in Yusup (2018) are as follows: (1) operational definition of a variable, (2) representation of questions according to the variables to be studied, (3) number of questions, (4) answer format, (5) scale on the instrument, (6) scoring, (7) instructions for charging the instrument, (8) processing time, (9) population sample, (10) grammar, (11) writing layout (writing format).

The Minister of Education and Culture (2013) in Wardhani and Putra (2016) explains that "educational assessment must have basic principles, namely valid, objective, fair, integrated, comprehensive, sustainable, systematic based on criteria, economical, accountable and educational." Bloom (1956) in Situmorang (2018) asserts that the cognitive domain or cognitive domain are behaviors that emphasize intellectual aspects, such as knowledge, understanding, and thinking skills. Bloom et al. (1956) state that the cognitive domain is related to knowledge which involves the process of recalling specific and universal things, recalling methods and processes or recalling patterns, structures or settings. Bloom in Situmorang (2018) also argues that the cognition domain is divided into six hierarchical levels, which are then divided into two parts, namely Lower Order Thinking Skills (LOTS) consisting of knowledge and understanding. The second is Higher Order Thinking Skills (HOTS) which consists of application, analysis, synthesis, and evaluation. Anderson in Situmorang (2018) adds that the ability to think creates as the highest level, after the ability to evaluate is included in the HOTS category.

The material studied at the elementary school level includes material that is factual or based on facts found in everyday life so it is necessary for a teacher to direct the HOTS learning process, this is what students encounter in everyday life can be understood and learn with ease and fun in the hope that students are able to explore and apply high-level thinking skills from an early age, and can be used as provisions for the next level of education. Low-level thinking proposed by Situmorang (2018) consists of the ability to know and understand which is the most basic level of thinking from the cognitive aspect or domain. Sudjana (2010) in Prasetya (2012) suggests that the cognitive domain is a domain related to intellectual learning outcomes which includes six aspects, namely knowledge or memory, understanding, applica-

tion, analysis, synthesis, and evaluation. The first two aspects are called low-level cognitive, namely knowledge and understanding. Besides, Anderson and Krathwohl in Pi'i (2016) explain that LOTS is a low-level thinking which includes the dimensions of knowing (C1) and understanding (C2) thinking processes that measure factual, conceptual and procedural knowledge. In addition, Brookhart (2010) in her book explains that higher-order thinking skills are divided into three categories: HOTS as a transfer process, HOTS as critical thinking skills, and HOTS as a problem solving.

Definitions that I find helpful fall into three categories: (1) those that define higher-order thinking in terms of transfer, (2) those that define it in terms of critical thinking, and (3) those that define it in terms of problem solving, terms of problem solving (Brookhart, 2010).

A good assessment instrument must meet the appropriate criteria. Therefore, Arikunto (1992) in Amirano and Daryanto (2016) explains that a good measuring instrument must have validity, reliability, objectivity, practicality, and economics. Purwanto (2011) in Wijayanto et al. (2016) believes that in an assessment instrument, there is a need for curricular validity based on content or content related to the material to be measured in accordance with the curriculum, syllabus, and Learning Process Plan, then the use of language in the assessment instrument will affect the level of difficulty of the items arranged so that it must pay attention to grammar in accordance with the correct spelling. Wijayanto et al. (2016) also explain that the use of appropriate language will make it easier for students to understand the meaning of the questions well so that the assessment instruments that are arranged can measure what they want to measure. Sudijono (1991) in Khaerudin (2015) suggests that a valid instrument must be logical and empirical. Based on the description of the expert's opinion, it can be concluded that qualitatively, the assessment instrument must have validity, reliability, objectivity, practicality, economics and pay attention to the logical and empirical linguistic arrangement so that it can be said to be a qualitatively quality instrument.

In fact, there are still many teachers who have not implemented HOTS-based assessments so that the learning process and HOTS-based learning outcomes are still low. Wachyudi et al. (2015) state that the government has made efforts to change the assessment on cognitive, affective and psychomotor aspects, but has not shown maximum results. It is in accordance with Nurani et al. (2019) who state that the cognitive assessment in the 2013 curriculum is the most complicated and confusing assessment so that the assessment made by the teacher is only based on the teacher's understanding and knowledge. Setiadi (2016) states that the assessment in the 2013 curriculum is considered more complicated than the assessment in the previous curriculum. To overcome this, the government has pursued various strategies to implement HOTS-based learning processes and assessments in accordance with the demands of the times. Another factor causing the low achievement of HOTS in Indonesia is that the students are not used to working on HOTS questions. Many teachers find it difficult to compile HOTS questions so they use existing and previously-made questions that are still in the LOTS (Lower Order Thinking Skills) category. This factor is one of the factors for untrained children in solving HOTS-based questions. However, in an assessment tool the number of questions presented is the comparison of the proportions between LOTS and HOTS as agreed by each school, 80% LOTS-based and 20% HOTS-based as the implementation of the assessment set in the 2013 curriculum that students must be trained and accustomed to doing questions based on HOTS. These factors also affect the achievement of HOTS ability because the one who provides an assessment to see the achievement of HOTS ability is the teacher so that when HOTS ability is low, the teacher needs to re-examine the learning process until the assessment process used includes the type of instrument used. Situmorang (2018) explains that high-order thinking skills are a competitive advantage for students, thus higher-order thinking skills need to be developed in learning, so it is important for teachers to understand correctly how to assess these abilities.

The difficulty of teachers in understanding the differences in student abilities is also an obstacle for teachers in preparing lesson plans that contain assessments to be carried out. This difficulty affected the preparation of HOTS questions. However, the difficulties experienced by this teacher can also be caused by the difficulty of the teacher in understanding how to prepare HOTS-based assessment instruments used in learning.

At the international level, there are several tests used to measure students' ability in the HOTS form, such as those held by PISA (Program for International Student Assessment) and PIRLS (Progress in International Reading Literacy Study). The achievements of Indonesian students are not satisfactory and only reached level two of the six levels contained in PISA. This low achievement is possible due to several factors, including the learning process or even the assessment used by the teacher, so students are not familiar with the HOTS questions. Assessment is part of evaluating the achievement of students and teachers in teaching (Nugroho, 2018). Based on the aforementioned description, it is important for an educator to master the preparation of assessment instruments, so that this research is very necessary to be conducted.

RESEARCH METHOD

The method used in this research is descriptive qualitative method by qualitatively describing the data obtained from the field. The research was conducted at eight state elementary schools in East Jakarta from January to February 2020. The data collection was carried out by observation, documents, and interviews, analyzed using comparative analysis techniques for checking the validity of the data. Examination of the research data must be carried out to ensure and confirm the results of the research before making conclusions. According to Guba (1981), there are four aspects of the validity or quality of qualitative research. The four aspects can be seen in Table 1.

Table 1. Aspects of Validity of Qualitative Research Perspective by Guba (1981)

Aspect	Scientific Term	Naturalistic Term
Truth Value	Internal Validity	Credibility
Applicability	Eksternal Validity Generalizability	Transferability
Consistency	Reliability	Dependability
Neutrality	Objectivity	Confirmability

Source: Guba (1981)

FINDINGS AND DISCUSSION

The findings based on observations were made at eight public elementary schools in East Jakarta which were used as research sites by entering the classroom when the teacher was teaching and giving an assessment that eight teachers at the schools only used questions available in textbooks owned by students because they were considered more practical and easy for teachers to do.

The next finding was that the results of interviews with grade XI teaching teachers turned out that there was one teacher who had not attended HOTS-based instrument preparation training, while seven teachers had attended the training but they said they had problems in its implementation because it was still something new for elementary school level teachers. This is the reason teachers do not make their own assessment instruments.

Documents taken as data were in the form of assessment instruments made by eight public elementary school teachers in East Jakarta, each of which consisted of ten multiple choice questions in science subject of grade XI with the reproductive system material so that all the items were 80 items, then are validated by two expert. The validation results are presented in Table 2. Then, the document was also validated by a second expert which can be seen in Table 3.

Table 2. The Result of Teacher Evaluation Instrument Validation by Expert 1

School	LOTS	HOTS	Other
School A	6	4	0
School B	8	2	0
School C	8	2	0
School D	7	3	0
School E	10	0	0
School F	8	2	0
School G	5	5	0
School H	10	0	0
Total	62	18	0

Table 3. The Result of Teacher Assessment Instrument Validation by Expert 2

School	LOTS	HOTS	Other
School A	9	1	0
School B	6	4	0
School C	9	1	0
School D	6	4	0
School E	4	4	2
School F	8	1	1
School G	7	3	0
School H	10	0	0
Total	59	18	3

Based on Table 2, there are 62 items included in the LOTS category and 18 items in the HOTS category. This categorization is based on the cognitive domain. Validation is carried out based on the suitability of the content or material to be measured, namely related to the reproductive system.

Validation conducted by the second expert in Table 3 shows that there are 59 items in the LOTS category and 18 items in the HOTS-based category, with three items that are unclear and not included in the reproductive system material. These three items were found in school E totaling two items and school F totaling one item.

Table 2 and Table 3 present different validation results, because there are different understandings between the two experts, but as a whole, they do not show a significant difference. Overall, the items made by the teacher who were included in the LOTS category were 81.25% of the total, and 18.75% of the items were included in the HOTS category. This is similar to the research that was conducted by Samosir et al. (2019) with the results of the number of HOTS quality questions of 51% and LOTS of 49%, the difference is that the number of HOTS in this study was less than that of Samosir et al. (2019) material, and the research was also carried out in different places. It is contradictory to the context of the research that was carried out by Himmah (2019), in which she also analyzed the level of the MOTS, with the end-semester assessment questions on mathematics subject. However, the method used in her research was the same as this research, namely, descriptive analysis. Moreover, other similar researches on item analysis were also conducted by Cahyono and Adilah (2016), and also Muklis and Oktora (2015) who used the cognitive level category of knowing, applying, and reasoning.

Data analysis that was conducted using Anates software to find out the validity, reliability, differentiation power, and level of difficulty in this study distinguishes it from the research that was conducted by Sudrajat (2018) which deals manually with the results of 30 questions: there are 14 valid questions, and 16 other questions are not valid. After conducting interviews with elementary school teachers in East Jakarta, it turned out that the same thing was found in the research of Sudrajat (2018), namely, the problem of educators who neglected their duties and functions, such as not analyzing the questions given to students.

Proportionally, the HOTS items were arranged in a smaller number because each assessment tool had to be adjusted to the processing time, so that there were no major obstacles for students who worked on it. In addition, the teacher still has difficulty in compiling the HOTS questions. It is possibly because the teacher has not attended the HOTS question preparation training.

Based on interviews with the teachers of grade XI students, all of them pay attention to the validity of the content when compiling the items because the suitability of the material is the main benchmark. Then linked to the results of validation by experts, the validity of the content contained in each item has also been completed, but what needs to be improved in the assessment instrument is the Operational Verb or *Kata Kerja Operasional* (KKO) on indicators that do not match the form of questions that appear in the questions.

Furthermore, the questions that teachers usually use are questions that are already available in the textbook that has been provided for student, especially in daily tests. However, there were also items that the teacher modified according to the conditions and material presented in the classroom. Testing the assessment instrument before being used as a measuring tool for the success of the learning process must indeed be done so that the learning objectives can be achieved properly, testing in this study was carried out as research supporting data to determine the quality of the items compiled, which in previous research was conducted by [Hartuti and Handayani \(2019\)](#) with no testing, so they only know the results of the analysis. In their research, it was found that the implementation of the 2013 curriculum assessment, in general, was in accordance with the 2013 curriculum assessment standards by making HOTS questions from the daily tests, mid-semester tests, and end-semester tests were in accordance with the syllabus, lesson plans, teacher books, and 2013 curriculum standards.

CONCLUSION

Based on the data analysis and research findings, it is concluded that (1) the instrument used by elementary school teachers in East Jakarta has fulfilled the content validity, (2) the cognitive domain (LOTS & HOTS) of the instruments used by elementary school teachers is proportional to 81.25% LOTS and 18.75% HOTS, (3) the quality of the assessment carried out by elementary school teachers in East Jakarta is generally good, but there are three items, namely two items for school E and one item for school F, which are not in accordance with basic competencies, and (4) implementation of assessments carried out by elementary school teachers in East Jakarta has followed the assessment standards given by the government. In addition, it is suggested that regular training for teachers should be conducted to monitor their success, and it is necessary to carry out comprehensive further training at all levels of school and in all subject areas as a step to maximize the implementation of the demands of the curriculum in the learning process. Assistance for teachers in training and developing themselves to maximize the learning process, including the assessments carried out, can also support their success.

REFERENCES

- Amirono, A., & Daryanto, D. (2016). *Evaluasi dan penilaian pembelajaran Kurikulum 2013*. Pustaka Pelajar.
- Arikunto, S. (2011). *Prosedur penelitian: Suatu pendekatan praktik*. Rineka Cipta.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals* (B. S. Bloom (ed.)). Longman.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

- Cahyono, B., & Adilah, N. (2016). Analisis soal dalam buku siswa Matematika Kurikulum 2013 kelas VIII semester I berdasarkan dimensi kognitif dari TIMSS. *Jurnal Review Pembelajaran Matematika*, 1(1), 86–98. <https://doi.org/10.15642/jrpm.2016.1.1.86-98>
- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology*, 29(2), 75–91. <https://www.jstor.org/stable/30219811>
- Hartuti, M., & Handayani, D. E. (2019). Analisis penilaian kognitif Kurikulum 2013 kelas rendah MI Sabilul Ulum Mayong Jepara. *El-Ibtidaiy: Journal of Primary Education*, 2(1), 1–8. <https://doi.org/10.24014/ejpe.v2i1.7370>
- Himmah, W. I. (2019). Analisis soal penilaian akhir semester mata pelajaran Matematika berdasarkan level berpikir. *Journal of Medives: Journal of Mathematics Education IKIP Veteran Semarang*, 3(1), 55–63. <https://doi.org/10.31331/medivesveteran.v3i1.698>
- Khaerudin, K. (2015). Kualitas instrumen tes hasil belajar. *Jurnal Ilmiah Madaniyah*, 5(2), 212–235. <https://journal.stitpemelang.ac.id/index.php/madaniyah/article/view/26>
- Kurniasih, I., & Berlin, S. (2016). *Revisi kurikulum 2013: Implementasi dan konsep penerapan*. Kata Pena.
- Maolani, R. A., & Cahyana, U. (2015). *Metodologi penelitian pendidikan*. Rajawali Pers.
- Muklis, Y. M., & Oktora, S. R. (2015). Analisis deskriptif soal-soal dalam buku siswa Kurikulum 2013 (edisi revisi) dan BSE pelajaran Matematika SMP kelas VII ditinjau dari domain kognitif TIMSS 2011. *Prosiding Sempoa: Seminar Nasional, Pameran Alat Peraga, Dan Olimpiade Matematika 1 2015*, 71–78. <https://publikasiilmiah.ums.ac.id/handle/11617/6132>
- Mulyasa, M. (2017). *Pengembangan dan implementasi Kurikulum 2013*. PT Remaja Rosdakarya.
- Nugroho, R. A. (2018). *HOTS: Higher order thinking skill*. PT Gramedia Widiasarana Indonesia.
- Nurani, H., Artharina, F. P., & Kiswoyo, K. (2019). Analisis pelaksanaan penilaian kognitif berbasis kurikulum 2013 Sabiul Ulum Mayonglor Kabupaten Jepara. *Indonesian Journal of Educational Research and Review*, 2(2), 172–181. <https://ejournal.undiksha.ac.id/index.php/IJERR/article/view/17625>
- Pi'i, P. (2016). Mengembangkan pembelajaran dan penilaian berpikir tingkat tinggi pada mata pelajaran Sejarah SMA. *Sejarah Dan Budaya: Jurnal Sejarah, Budaya, Dan Pengajarannya*, 10(2), 197–208. <https://doi.org/10.17977/um020v10i22016p197>
- Prasetya, T. I. (2012). Meningkatkan keterampilan menyusun instrumen hasil belajar berbasis modul interaktif bagi guru-guru IPA SMP N Kota Magelang. *Journal of Research and Educational Research Evaluation*, 1(2), 106–112. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/873>
- Regulation of the Minister of Education and Culture No. 104 of 2014 on the Learning Outcome Assessment by Educators in Primary and Secondary Education Levels, (2014).
- Regulation of the Minister of Education and Culture No. 20 of 2017 on the Team of Performance Assessor in the Ministry of Education and Culture, (2017).
- Regulation of the Minister of Education and Culture No. 21 of 2016 on the Content Standard of Primary and Secondary Education, (2016).
- Regulation of the Minister of Education and Culture No. 23 of 2016 on Educational Assessment Standard, (2016).

- Samosir, A., Hasruddin, H., & Dongoran, H. (2019). Analisis kuantitas dan kualitas pertanyaan guru Biologi dan siswa materi Sistem Eksresi. *Jurnal Pelita Pendidikan*, 7(1), 9–15. <https://doi.org/10.24114/jpp.v7i1.10523>
- Setiadi, H. (2016). Pelaksanaan penilaian pada Kurikulum 2013. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 20(2), 166–178. <https://doi.org/10.21831/pep.v20i2.7173>
- Setiawati, W., Asmira, O., Ariyana, Y., Bestary, R., & Pudjiastuti, A. (2019). *Buku penilaian berorientasi higher order thinking skills*. Direktorat Jenderal Guru dan Tenaga Kependidikan, Kementerian Pendidikan dan Kebudayaan.
- Situmorang, J. (2018). *Higher order thinking skills: Pengembangan keterampilan berfikir tingkat tinggi*. MDP Media.
- Sudrajat, H. (2018). *Analisis alat evaluasi pada mata pelajaran IPA di kelas V Sekolah Dasar Negeri 3 Nyerot Kecamatan Jonggat Lombok Tengah* [Graduate thesis, Universitas Maulana Malik Ibrahim, Malang]. <http://etheses.uin-malang.ac.id/11101/1/15761011.pdf>
- Suryanto, B., & Sutinah, S. (2011). *Metode penelitian sosial*. Kencana Media Group.
- Wachyudi, I., Sukestiyarno, S., & Waluya, B. (2015). Pengembangan instrumen penilaian unjuk kerja pada pembelajaran dengan model problem solving berbasis TIK. *Journal of Research and Educational Research Evaluation*, 4(1), 20–27. <https://journal.unnes.ac.id/sju/index.php/jere/article/view/6928>
- Wardhani, D. F., & Putra, A. P. (2016). Pengembangan instrumen tes standar kognitif pada mata pelajaran IPA kelas 7 SMP di Kabupaten Banjar. *Proceeding Biology Education Conference*, 75–82. <https://jurnal.uns.ac.id/prosbi/article/view/5658>
- Wijayanto, P. A., Allifah, A., & Amirrudin, A. (2016). Evaluasi kualitas instrumen tes dalam pembelajaran Geografi di MAN 2 Kota Batu. *Jurnal Geografi*, 13(2), 101–113. <https://journal.unnes.ac.id/nju/index.php/JG/article/view/7969/5523>
- Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, 7(1), 17–23. <https://doi.org/10.18592/tarbiyah.v7i1.2100>

Character education strengthening model during learning from home: Ki Hajar Dewantara's scaffolding concept

Hasti Robiasih; Ari Setiawan; Hanandyo Dardjito*

Universitas Sarjanawiyata Tamansiswa

Jl. Kusumanegara No.157, Muja Muju, Umbulharjo, Yogyakarta 55165, Indonesia.

*Corresponding Author. E-mail: dardjit@gmail.com

ARTICLE INFO

Article History

Submitted:

9 December 2020

Revised:

31 January 2021

Accepted:

17 March 2021

Keywords

character education;
learning from home;
model; Ki Hajar
Dewantara; scaffolding

Scan Me:



How to cite:

Robiasih, H., Setiawan, A., & Dardjito, H. (2021). Character education strengthening model during learning from home: Ki Hajar Dewantara's scaffolding concept. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 25-34.

doi:<https://doi.org/10.21831/pep.v25i1.36385>

ABSTRACT

In providing educational and learning services to students during the Covid-19 Pandemic, learning is carried out from home. The disparity in learning achievements, especially related to attitude competencies, is strongly experienced during learning from home. This study aims to integrate attitude values through learning with Ki Hajar Dewantara's scaffolding concept: Identifying, Imitating, Developing, Disseminating. This scaffolding is applied to strengthen the character of junior high school students in Yogyakarta Special Region to become Pancasila students. This is a research and development study that applies a qualitative approach. The respondents of this study were teachers, principals, students, parents, and supervisors of junior high schools in this province. Data were collected using focus group discussion, interview, and observation, then were analyzed using Miles and Huberman models, which comprised data reduction, data display, and inference stages. The results of this study is a character education strengthening model during learning from home applying Ki Hajar Dewantara's scaffolding concept. This model contributes to the policymaking of character education strengthening during the learning from home online.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



INTRODUCTION

In response to the Covid-19 pandemic, which urged the change of the teaching-learning process to be online, this study discusses the development of online character education, implying Ki Hajar Dewantara's scaffolding concept. Teaching knowledge in teaching practice becomes the focus of online teaching during the students' learning from home while character education is hardly conducted. This consideration leads to the idea of developing a character education model that contributes to the practical approach to teaching characters.

Education has a paramount role in the character and civilization formation of a nation. Education will consistently maintain the character values of a nation's civilization during dynamic social change. Education serves not only to encourage how to know and how to do, but also how to manifest it, which becomes the most important thing in social reality. Education is also the key to human resource development. The human resource quality is the key to the realization of Indonesia Emas 2045, the nation's goal that is fair and prosperous, safe and peaceful, and advanced and worldwide. Education also determines the direction of this nation's future. Education improves the relevant life order in the current change without having to lose its national identity personality.

[Law of Republic of Indonesia No. 20 of 2003](#) concerning the national education system, article 1 paragraph (1) states that:

Education is a conscious and planned effort to realize the atmosphere of learning and learning process so that learners actively develop their potential to have religious-spiritual power, self-control, personality, intelligence, noble character, as well as the necessary skills of themselves, society, nation, and state.

Furthermore, article 3 of the law states that national education serves to develop and form dignified national character and civilization to educate the life of the nation, aiming to enhance the potential of learners to become human beings who are believing in God and having noble character, healthy, knowledgeable, capable, creative, independent, and become democratic and responsible citizens; this is the profile of *Pancasila* (five pillars of Indonesia) students ([Regulation of the Minister of Education and Culture No. 22 of 2020](#)).

The key elements of *Pancasila* students who are diverse are knowing and appreciating different cultures, intercultural communication skills in interacting with others, and reflection and responsibility for the experience of diversity. Indonesian students recognize the value of *gotong royong*, which means a willingness to work collaboratively. This value aims to carry out the workload smoothly, easily, and lightly. The elements of working together are collaboration, caring, and sharing. Independent Indonesian students are responsible for their learning processes and outcomes. The key independence elements consist of self-awareness and situations sensitivity, and self-regulation. Students with critical reasoning can objectively process information qualitatively and quantitatively, building links between various information, analyzing it, and evaluating and concluding it. The elements of critical reasoning are obtaining and processing information and ideas, analyzing and evaluating reasoning, reflecting thought and thought processes, and making decisions, while creative learners are students who can modify and produce something original, meaningful, useful, and impactful. The key elements of being creative consist of generating original ideas and producing original works and actions.

Currently, all regions of the Republic of Indonesia are affected by the spread of Covid-19. Under any circumstances, the country is obliged to protect the entire nation, promote the general welfare, and educate the people. Therefore, the country is obliged to find a way out of the continuity of education in schools. Realizing the geographical location of Indonesia as an archipelago with different circumstances, a regulation that can become a solution needs to be formulated so that learning activities can still be carried out properly during any emergency conditions. Learning should have never stopped, whatever happens. In case of emergency, learning activities cannot normally run as usual, but students must still get education and learning services.

The government of Indonesia in the world of education seeks to break the chain of transmission of this virus by organizing home learning for all levels of education through the [Circular Letter of the Ministry of Education and Culture No. 4 of 2020](#) concerning the implementation of educational policy in the emergency period of the spread of Corona Virus Disease-19. This surprises the institutions, teaching staff, learners, and parents. Universities, not only in Indonesia but around the world as the highest educational institutions, also participated in the shock. Researches report on how educators and educational institutions are doing and sharing methods as well as strategies to deal with the Covid-19 pandemic ([Daniel, 2020](#); [Romero-Ivanova et al., 2020](#); [Shenoy et al., 2020](#); [Zhang et al., 2020](#)). These studies discuss knowledge learning, but character learning online is still very limited. The researchers believe that character learning, during learning from home, has not been studied. Considering this limitation, this study proposes a model for character education during learning from home, applying Ki Hajar Dewantara's scaffolding concept.

Before discussing learning from home any further, education centers need to be recognized first. A basic understanding of educational agents needs to be understood that the responsibility of education lies in three educational centers. This was initiated by Ki Hajar De-

wantara in the concept of *Tri Pusat Pendidikan* (three centers of education), namely, family, school, and community (Dewantara, 2013). It is further explained that families play a role in educating children to have good character and ethics. Schools play a role in giving children knowledge. When the first and second centers have adequately equipped children, the community plays the social and social education role. Based on the three education centers' understanding, the thought that education is entirely the school's responsibility is certainly not justified. The first center mentioned is the family. The parent or guardian plays a leading role in the learners' character education. However, the school is seen as playing a role to help parents in educating children's character. The concept of involving parents in educating the students is in line with some current researches that parents/guardians and schools can jointly educate children's character during the challenging time of the Covid-19 virus emergence (Asbari et al., 2019; Erol & Danyal, 2020; Lake & Olson, 2020).

From some information circulating in the community about the implementation of learning from home, it is known that not all schools can run full online learning activities. Most schools host off-line learning. Some obstacles found include limited human resources, limited facilities, the difficulty for parents in doing mentoring, and so on. Not all students have a computer or smartphone. Students also have difficulty accessing the internet and limited internet quota. These are true in the Indonesian context, as pointed out by Siron et al. (2020). In addition, the implementation of learning from home during the Covid-19 emergency period between one school and another school varies greatly, in accordance with the perception and readiness of each school.

According to the Decree of the Minister of Education and Culture No. 719/P/2020, there needs to be a paradigm change in learning planning, implementation of learning, and assessment of learning outcomes. Learning from home activities demand collaboration, participation, and active communication between teachers, parents, and students. This collaboration must be a unity that supports each other, on the principle that all of us are teachers, all of us are students, and all places are classrooms. Learning from home, which is online, not only meets the demands of competence in the curriculum, but is emphasized more on character development, noble character, and student independence. Teachers must be more creative and innovative in presenting subject matter and assigning assignments to students, to realize meaningful, inspiring, and enjoyable learning. Thus, students are expected not to experience the boredom of learning from home.

Researches suggest that teaching and learning during the Covid-19 outbreak need contingency strategies (Bao, 2020; Daniel, 2020). Many studies show that the instructional process experienced and is experiencing shock, particularly among the community with low technology literacy (Romero-Ivanova et al., 2020; Shenoy et al., 2020; Siron et al., 2020). These researches report different teaching contexts during the outbreak. They emphasized the knowledge teaching during the pandemic; but, the character teaching was left behind. The current education system in Indonesia urges more teaching on character education to complement knowledge teaching (Regulation of the Minister of Education and Culture No. 22 of 2020).

Character education has a paramount role in building a strong baseline for a community. It provides basic education and self-control concerning the students' local cultural value for facing this open-access information era as the emergence of the internet (Harun et al., 2020; Hermino, 2020; Rosmiati et al., 2016). Character education in some contexts is integrated into learning subjects, such as math and multimedia (Kadek, 2020; Suyitno et al., 2019). Some models of character education were developed (Dewia & Alam, 2020; Septiani, 2020) to pave the way the teaching and learning; however, a contextual model implicating Indonesian local wisdom has not been developed or was limited.

Ki Hajar Dewantara's scaffolding concept emphasizes that teachers must be able to accommodate the development aspects of attitude and scientific values in the learning process with the implementation of 4Ns concept, namely; identifying concepts and values (*NITENI*),

imitating (*NIROKKE*), implementing the concepts and contextually developing the implemented values (*NAMBAHI*), and disseminating them (*NULARKE*) in the wider community (Boentarsono et al., 2016; Dewantara, 2013). The research was, therefore, conducted aiming at developing a character value strengthening model using the concept of Ki Hajar Dewantara's 4N scaffolding when learning from home.

In regard to the aforementioned concept, this study looks at the following research problem. (1) What are the characters needed to be strengthened? (2) How is the model developed applying Ki Hajar Dewantara's scaffolding concept to strengthen the character?

RESEARCH METHOD

This study applied the research and development (R&D) method, which consisted of two stages: need analysis and model development. The need analysis looked for the character needed to be strengthened using the model. The model development is the stage where the Ki Hajar Dewantara scaffolding concept is applied to teach the character found in the need analysis. Figure 1 illustrates the stages in this research.



Figure 1. Research Stages

This study involved junior high school setting in three areas in Yogyakarta Special Region that conducted learning from home: Bantul Regency, Sleman Regency, and Yogyakarta City. There were nine schools altogether. Respondents involved in this study were 12 junior high school teachers, three principals, three supervisors, 15 students, and three parents. The researchers selected the respondents for various considerations such as school accreditation status, media for online used during learning from home, and representation of the school area. This method provided comprehensive and complete data related to the need analysis of character strengthening during learning from home.

Data collection comprised three stages, namely the exploration phase using the Focus Group Discussion (FGD) method supported by the results review of questionnaires, interviews, and observation findings. Triangulation in this study covered data and source triangulation method to develop a comprehensive understanding of phenomena. After completing the exploration phase, the next was the needs analysis to create the model, such as the teaching materials that had integrated the value of the attitude for the profile of *Pancasila* students. The next step is product design by developing model prototypes, implementing prototype models to three schools in Sleman Regency, Bantul Regency, and Yogyakarta City, sampling three subject teachers, namely Mathematics, Bahasa Indonesia, and English subjects. The collected data were analyzed using Miles and Huberman technique. Miles et al. (2014) state that the analysis consists of three flows of activities that move simultaneously: data reduction, data presentation, and conclusion drawing/verification. Qualitative research explores and understands the meaning of individuals associated with social problems (Creswell & Creswell, 2018).

Data analysis in the need analysis stage was validated by an expert who comprised of teacher, principal, school district supervisor, and lecturer by applying expert judgment. In the stage of model development, the analysis applied expert judgment and field testing of the model. This study was done by analyzing the implementation results from prototypes, then

conducted deeper offline FGD by applying strict health protocols. The results of the analysis and the results of the interview proceeded to data reduction by making abstractions to obtain conclusions. This step was an attempt to summarize the core, process, and statements that need to be maintained to stay in it.

FINDINGS AND DISCUSSION

Along with the recent development of technology and information in the era of industrial revolution 4.0, the future workforce's world will certainly be very different from the current situation. Major changes in the era of industrial revolution 4.0 formed a different world of work in terms of structure, technology, and concept of self-actualization. The structure of the work will be more flexible, knows no geographical boundaries, and is unbound. This will result in the worker not being tied to just one institution throughout his or her career.

Based on the questionnaires filled out by the principal, most schools set character grades that had to be strengthened to students in the school's curriculum: school vision and mission (Law of Republic of Indonesia No. 20 of 2003; Regulation of the Minister of Education and Culture No. 20 of 2018; Regulation of the Minister of Education and Culture No. 37 of 2018). The Ministry of Education and Culture has determined these character values in the 2020/2024 Strategic Plan (Regulation of the Minister of Education and Culture No. 22 of 2020). Of the six key elements of *Pancasila* student profile, the teachers had strengthened the character values such as believing in God, being independent, creative, critical thinking, collaboration, and having an awareness of diversity. During the FGD, the respondent confirmed that "believing in God" had been taught by all teachers and confirmed by the students, principals, and district supervisors. On the other hand, "diversity awareness" was identified to be taught the least.

The teacher documented the character strengthening in the lesson plan. The implementation of character strengthening was mostly done by asking students to pray at the beginning and closing of the learning activity. The teachers argued that believing in God and other noble characters was usually carried out by the teacher during face-to-face meetings. Based on the data, it still needed to be strengthened by the teacher to provide other character strengthening. The teachers believed that the students understood the strengthening of character so that students would be able to practice the values in their daily lives. According to 75% of the teachers, the strengthening process was challenging in an online setting. It was hard to do because many parents were busy or working. The parents were having difficulties monitoring the development of their child's character. The support of the principal and supervisor to the teachers to strengthen the students' character was paramount. The principal's monitoring and evaluation process was mostly administratively oriented, namely checking whether the character to be strengthened had already been included in the teachers' lesson plan. The headmaster sometimes also visited the classroom during lessons and assessments and delivered them in regular meetings. They could not organize the observation process because of the learning from home condition.

Students confirmed the character values strengthening data by their teacher. They stated that the character values which the teacher reinforced were religious, disciplined, independent, and creative. Additionally, their teachers always asked them to pray, be disciplined during learning, be creative, and be independent, especially in doing their homework. This demonstrated that the students did not recognize the meaning of other character values strengthened by the teachers.

The character strengthening education conducted by the school during learning from home seemed incidental. Also, it did not integrate with both the learning and the conformity with the students' condition. The analysis of interview data conducted with teachers, students, and parents implied these phenomena. The data analysis from the students revealed that char-

acter education was mostly about praying and reminding around certain characters without providing meaningful guidance for those who violated the character value. For example, according to the students, teachers did not reprimand students who were late collecting assignments or were undisciplined while learning from home.

The parents who accompany the child justified the situation. They stated there was no effective communication related to character disseminating during learning from home. Most discussions between teachers and parents were about school assignments using WhatsApp. The communication made by teachers to parents was mostly through WhatsApp group, and its contents were mostly tasks that had to be done by students at home. The teachers were also aware of that. The interviews with teachers from three regions implied that some teachers also found it difficult to include character strengthening and conduct their online learning assessments. This is because no planned model or strategy was mutually endorsed in schools and passed from supervisors. Conducting monitoring on character strengthening was also challenging for the supervisors because most consulted issues focused on conducting the learning and assessment process during learning from home.

The findings show that teachers wanted a new strategy that was a design or model that teachers can use to integrate the character values through learning and reduce the boredom of teachers, students and parents. Teachers hoped that the school could collaborate with the department or universities to create a learning design to integrate character values.

Further findings indicate that the characters that had to be strengthened during learning from home turned out to get a mixed response. From the results of interviews and study of lesson plan documents, teachers emphasized disciplined character, honesty, and independence. It was believed that the character of discipline was necessary because in the opinion of Nucci (2008, p. 197), “developmental discipline can help teachers build the trusting relationships necessary for all students to learn and develop academically and morally.” The development of discipline carried out by teachers was able to shape students into disciplined people while also improving student achievement academically. Students who had a disciplined character in their daily lives, of course, would also be disciplined in learning at school and home so that it indirectly affected the improvement of achievement in school.

Students reported different characters values delivered by their teachers. Students said supervisors and principals emphasized independent character and hard work. Self-reliance concerning an independent person and having confidence could enable the students to adapt and take care of things by themselves (Parker, 2006, pp. 226–227). Children’s ability to self-help needs to be grown because teachers could not help when learning from home. The students also argued that self-reliance was the main provision during learning from home. However, the value of the character had not become a habit and permanent yet. This character had not been a special student profile yet.

Based on the above discussion, it can be confirmed that character strengthening cannot occur if the teacher only transfers knowledge and occasionally reminds students to be good. Character strengthening must be done intentionally (by design) to seek the transformation of values in forming the character of the nation’s children. According to Ki Hajar Dewantara, Strengthening character education can be done with the *Trisentra* System or *Tripusat* Education that is family, college, and community (Dewantara, 2013). All three have the task and obligation to form the students’ intellectuality and noble ethics (character).

In practice, families, communities, and schools have strengthened character education by habituating excellent behavior practices that must be continuous until they finally form habits. However, this applies to character grades that are common and can be carried out by students regularly. Strengthening the value of character values that can develop their competitiveness to live in the future as stated in the profile of *Pancasila* students must be pursued and planned more seriously and integrated into the learning process. Ki Hajar Dewantara’s 4N scaffolding concept; *Niteni* (paying close attention/observing), *Nirokke* (imitating), *Nambahi*

(adding), and *Nularke* (communicating), can integrate character values in the teaching. This scaffolding stage of 4N is an effort to pave a way to strengthen students' character so that the value of character values that they have understood can be a catalyst and generate the students' willingness to produce behavior regularly to form a permanent habit.

The internalization process of a concept or teaching will be possible when students carefully see/observe (*Niteni*) the concept, consider it, and then imitates (*Nirokke*) it. Students will develop it or add it (*Nambahi*) according to its context in the next stage. In the end, students can convey/communicate (*Nularke*) the value of the concept to others. The result of model development by applying Ki Hajar Dewantara's 4N scaffolding concept for strengthening the character can be seen in Figure 2.

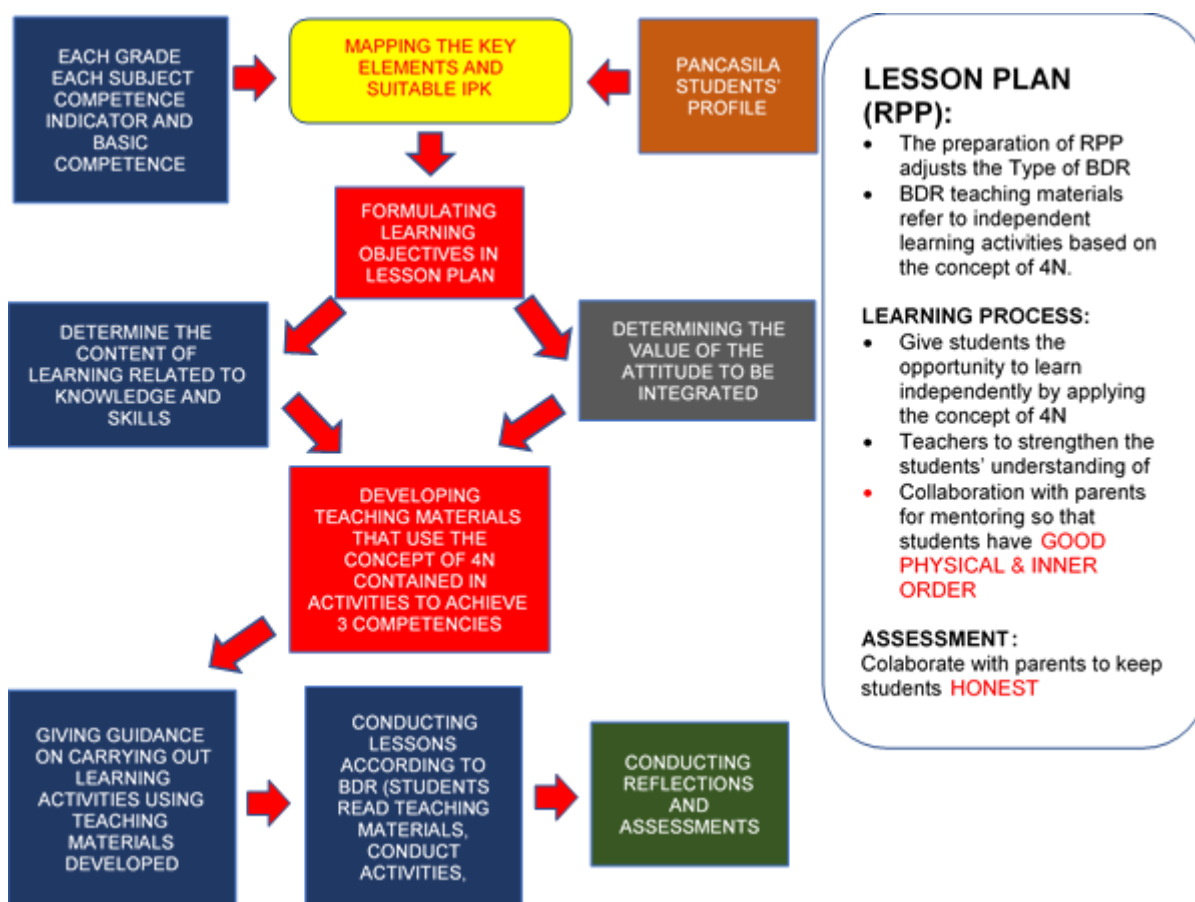


Figure 2. Model by Applying Ki Hajar Dewantara's 4N Scaffolding Concept

The stages in the model illustrated in Figure 2 are as follows. (1) Select the core competence, formulate the indicators of the competency achievement, and select key elements of *Pancasila* students. (2) In each lesson, focus on a number of attitude values to be developed. (3) Formulate the learning objectives that will be outlined in the lesson plan document. (4) Develop teaching materials that have contained character value (it can be in the form of text, modules, videos, sound recordings, etc.). (5) Character values that will be strengthened in teaching materials (it can be written in objectives, activities (including text), exercises, summaries, reflections, assessments, utilization of blank space or delivered in impressions or sounds). (6) Make instructions to carry out learning activities using teaching materials that have been developed. (7) Allow students to learn independently guided by teachers and accompanied by parents, by emphasizing the order of birth and inner and honesty. (8) At the end of weekly activities, students are asked to self-report what has been observed (*Niteni*), what has been done

(*Nirokke*), what other things can be developed (*Nambahi*), and given the challenge of what benefits for others (*Nularke*). (9) Strengthen the meaningful things that have been learned and valuable for their lives. (10) Do reflection and follow-up.

The development of this model was to assist teachers in creating learning tools for the lesson plan that includes strengthening the students' character. In addition to the character strengthening, the model was an effort to shape *Pancasila* student profile and free learning. The results of this model applied in various junior high schools in Yogyakarta indicated that it assisted the teachers in integrating the key elements of the character value of Pancasila Student Profile in learning during the learning from home process. The teacher's responses implied this process. The three schools agreed that the model developed assisted teachers in conducting learning. The student response also showed positive results seen in their ability to acquire teaching material during online learning. The reason for this success is contextual learning development. It demands critical thinking and gives children the freedom to express an opinion.

The school principals felt this model was advantageous. It helped monitor the strengthening of character in the school by looking at the lesson plan documents, the teaching materials developed, and the reflection sheets written by the students in each lesson. Besides, this model helps school supervisors provide direction related to improving teacher innovation and creativity in developing learning to achieve student competencies in attitude, knowledge, and skills competencies.

Based on the discussion above, all education agents should be aware of the strengthening of character that meets the criteria of the *Pancasila* student profile. The developed model could be one of the implication strategies because it allows students, teachers, parents, principals, and supervisors to synergize to strengthen character education that contains aspects of critical thinking, independence, faithful and godly character, and creative mind could be realized.

CONCLUSION

The need analysis in this study recognized some characters to be strengthened, such as belief in God, independence, creativity, critical thinking, collaboration, and diversity awareness. These characters were developed into the model of character strengthening, which is integrated into the teaching. This model has been validated by the model field testing, which contributes to the policy-making process for the teaching character during learning from home. In regard to character education, the collaboration between students, parents, teachers, principals, and education district supervisors should be maintained and encouraged, especially during learning from home.

ACKNOWLEDGMENT

This work was supported by Kementerian Pendidikan dan Kebudayaan, Badan Penelitian dan Pengembangan dan Perbukuan, Pusat Penelitian Kebijakan, under Grant 0758/H2/PBJ/2020.

REFERENCES

- Asbari, M., Nurhayati, W., & Purwanto, A. (2019). The effect of parenting style and genetic personality on children character development. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 23(2), 206–218. <https://doi.org/10.21831/pep.v23i2.28151>
- Bao, W. (2020). COVID-19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies*, 2(2), 113–115. <https://doi.org/10.1002/hbe2.191>

- Boentaronso, K. B., Dwiarmo, K. P., Suharto, K. R., Iswanto, K. B., Masidi, K., Widodo, K. R. B., & Swasono, K. S.-E. (2016). *Tamansiswa: Badan perjuangan kebudayaan & pengembangan masyarakat* (K. S. Ph (ed.)). Universitas Sarjanawiyata Tamansiswa Press.
- Circular Letter of the Ministry of Education and Culture No. 4 of 2020 on the Implementation of Educational Policy in the Emergency Period of the Spread of Corona Virus Disease-19 (COVID-19), (2020).
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publication.
- Daniel, S. J. (2020). Education and the COVID-19 pandemic. *PROSPECTS*, 49(1–2), 91–96. <https://doi.org/10.1007/s11125-020-09464-3>
- Decree of the Minister of Education and Culture No. 719/P/2020 on the Curriculum Implementation Guidelines in Educational Units during Special Conditions, (2020).
- Dewantara, K. H. (2013). *Karya Ki Hadjar Dewantara bagian pertama: Pendidikan*. Universitas Sarjanawiyata Tamansiswa (UST Press) dan Majelis Luhur Persatuan Tamansiswa.
- Dewia, E. R., & Alam, A. A. (2020). Transformation model for character education of students. *Cypriot Journal of Educational Sciences*, 15(5), 1228–1237. <https://doi.org/10.18844/cjes.v15i5.5155>
- Erol, K., & Danyal, T. (2020). Analysis of distance education activities conducted during COVID-19 pandemic. *Educational Research and Reviews*, 15(9), 536–543. <https://doi.org/10.5897/ERR2020.4033>
- Harun, H., Jaedun, A., Sudaryanti, S., & Manaf, A. (2020). Dimensions of early childhood character education based on multicultural and community local wisdom. *International Journal of Instruction*, 13(2), 365–380. <https://doi.org/10.29333/iji.2020.13225a>
- Hermino, A. (2020). Contextual character education for students in the senior high school. *European Journal of Educational Research*, 9(3), 1009–1023. <https://doi.org/10.12973/eu-jer.9.3.1009>
- Kadek, I. (2020). Development of e-learning oriented inquiry learning based on character education in Multimedia course. *European Journal of Educational Research*, 9(4), 1591–1603. <https://doi.org/10.12973/eu-jer.9.4.1591>
- Lake, R., & Olson, L. (2020). *Learning as we go: Principles for effective assessment during the COVID-19 pandemic*. Center on Reinventing Public Education.
- Law of Republic of Indonesia No. 20 of 2003 on National Education System, (2003).
- Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook* (T. R. Rohidi (trans.); 3rd ed.). UI Press.
- Nucci, L. (2008). *Handbook of moral and character education*. Routledge.
- Parker, D. K. (2006). *Menumbuhkan kemandirian dan harga diri anak* (S. Sunarni (ed.); B. Wibisono (trans.)). Prestasi Pustakarya.
- Regulation of the Minister of Education and Culture No. 20 of 2018 on the Strengthening of Character Education in Formal Education Units, (2018).
- Regulation of the Minister of Education and Culture No. 22 of 2020 on the Strategic Plan of the Ministry of Education and Culture in 2020-2024, (2020).
- Regulation of the Minister of Education and Culture No. 37 of 2018 on the Revision of the Regulation of the Minister of Education and Culture No. 24 of 2016 on the Core

Competence and Basic Competence for Curriculum 2013 Subjects in Primary and Secondary, (2018).

- Romero-Ivanova, C., Shaughnessy, M., Otto, L., Taylor, E., & Watson, E. (2020). Digital practices & applications in a Covid-19 culture. *Higher Education Studies*, 10(3), 80–87. <https://doi.org/10.5539/hes.v10n3p80>
- Rosmiati, R., Mahmud, A., & Talib, S. B. (2016). The effectiveness of learning model of basic education with character-based at Universitas Muslim Indonesia. *International Journal of Environmental & Science Education*, 11(12), 5633–5643.
- Septiani, A. N. S. I. (2020). Development of Interactive Multimedia learning courseware to strengthen students' character. *European Journal of Educational Research*, 9(3), 1267–1279. <https://doi.org/10.12973/eu-jer.9.3.1267>
- Shenoy, V., Mahendra, S., & Vijay, N. (2020). COVID 19 – Lockdown: Technology adaption, teaching, learning, students engagement and faculty experience. *Mukt Shabd Journal*, 9(4), 698–702. <http://shabdbooks.com/gallery/78-april2020.pdf>
- Siron, Y., Wibowo, A., & Narmaditya, B. S. (2020). Factors affecting the adoption of e-learning in Indonesia: Lesson from Covid-19. *Journal of Technology and Science Education*, 10(2), 282–295. <https://doi.org/10.3926/jotse.1025>
- Suyitno, H., Zaenuri, Z., Sugiharti, E., Suyitno, A., & Baba, T. (2019). Integration of character values in teaching-learning process of Mathematics at elementary school of Japan. *International Journal of Instruction*, 12(3), 781–794. <https://doi.org/10.29333/iji.2019.12347a>
- Zhang, W., Wang, Y., Yang, L., & Wang, C. (2020). Suspending classes without stopping learning: China's education emergency management policy in the COVID-19 outbreak. *Journal of Risk and Financial Management*, 13(3), 55. <https://doi.org/10.3390/jrfm13030055>

Developing the flipped learning instrument in an ESL context: The experts' perspective

Wahyu Hidayat^{1*}; Mohammad Musab bin Azmat Ali²; Nur Asmawati Lawahid³; Mujahidah¹

¹Institut Agama Islam Negeri (IAIN) Parepare

Jl. Amal Bhakti No.8, Bukit Harapan, Soreang, Kota Parepare, Sulawesi Selatan 91131, Indonesia.

²Universiti Malaysia Pahang

26600 Pekan, Pahang, Malaysia.

³Institut Agama Islam Negeri (IAIN) Datokarama Palu

Jl. Diponegoro No.23, Lere, Palu Bar., Kota Palu, Sulawesi Tengah 94221, Indonesia.

*Corresponding Author. E-mail: wahyuhidayat@iainpare.ac.id

ARTICLE INFO

ABSTRACT

Article History

Submitted:

20 January 2021

Revised:

17 March 2021

Accepted:

19 March 2021

Keywords

flipped learning approach;
ESL context; Fuzzy
Delphi method

Scan Me:



Numerous studies have accepted the flipped learning approach as an approach in implementing technological-based classroom environments. This article aims to identify the required constructs in developing an instrument for flipped learning in an ESL environment. This study uses the Fuzzy Delphi method to collect and analyze the viewpoints of 18 experts from relevant fields. An online questionnaire was developed to gather the experts' agreement towards seven constructs: flexible environments, the shift in learning culture, intentional content, progressive networking activities, professional educators, engaging & effective learning experiences, and diversified seamless learning platforms, and 68 items gathered from the literature. The Fuzzy Delphi Method (FDM) analysis rejected seven items, finalizing the instrument with seven constructs and 61 items. The instrument is beneficial to teachers and learners of ESL and developers of technology-based learning methods. The implication of the study is the provision of the constructs to help guide and implement the flipped learning approach in educational contexts. Furthermore, these constructs can be used as the basis for further investigations that lead to developing frameworks or models for the flipped learning approach. Future works on the topic may look at a bigger sample for stronger results. Furthermore, the instrument developed can be used on the student population and in other contexts as well.

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



How to cite:

Hidayat, W., Ali, M., Lawahid, N., & Mujahidah, M. (2021). Developing the flipped learning instrument in an ESL context: The experts' perspective. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 35-48.

doi:<https://doi.org/10.21831/pep.v25i1.38060>

INTRODUCTION

Education of the 21st century is radically changing than any preceding decades before. Technology has become a determining factor in helping education and lessons being meaningful and successful to millennial students (Akçayır & Akçayır, 2018; Azman & Dollsaid, 2018; Tsay et al., 2018). These students are more comfortable engaged with technology and learning with it as it gave rise and prominence to tech-based educational approaches such as e-learning, blended learning, and flipped learning (Embi, 2014; Hamdan et al., 2013; Kenna, 2014).

Traditional didactic approaches are becoming more and more inefficient in dealing with 21st-century students (Kenna, 2014). Students today are more sensitive to their divergent abilities and needs in classrooms, which is essential in delivering a meaningful lesson. Teachers must be able to address these divergences in the classroom for having an effective lesson delivered (Lage et al., 2000). This is especially true for tertiary-level education, as global connectivity has seen a rapid rise since the development of digital technology in the last ten years

(Enfield, 2013). The impact of digital technology is that students of this generation are more comfortable interacting and digesting information through many online interactive platforms.

The Flipped Learning Approach

The fact that the new generations prefer digital platforms over conventional ones spurred educators to use the digital technology platforms as an effective medium of teaching students and making learning a meaningful experience through different forms of interactions with the lesson content for different learning styles (Tsay et al., 2018). The flipped learning approach is a teaching approach spurred by digital technology in the classroom. The flexible and independent disposition of the approach jives well with the use of technology in education. Some experts believe that the approach allows for a cornucopia of pedagogical approaches to be implemented in a flip approach classroom, resulting in a flexible range of approaches that is Taylor-suited to each student's learning styles (Baepler et al., 2014). Juhary and Amir (2018) debated that many past studies have proven the ability of the flip learning approach to empower students to be self-dependent learners. Furthermore, the shift of responsibility of learning that befalls on the learner themselves proliferates the usage of learner-centered approaches that, in turn, allows for the individual students of different learning styles and abilities to learn and develop at their own pace (Raihanah, in Ministry of Education Malaysia, 2015).

Several previous studies have shown a positive effect of the flipped learning approach involving students (Chen et al., 2014; Davies et al., 2013; McLaughlin et al., 2013). This proves that the flipped learning approach in the classroom is an alternative in learning approaches and strategies. However, most research on flipped learning is conducted in mathematics and engineering subjects (Baepler et al., 2014; Chen et al., 2014; McLaughlin et al., 2013). Besides, several studies on flipped learning in ESL are only related to the limited readiness of students and lecturers in using the flipped learning approach (Embi, 2014; Jamaludin & Osman, 2014; Osman et al., 2014). Existing studies on model development and guidelines for applying the flipped learning approach are still relatively lacking. Existing studies also focus more on the general student population at universities (Baepler et al., 2014; Embi, 2014), so the flipped learning approach cannot be extended to a broader group. There are gaps in the literature related to research on the flipped learning approach in language learning, especially in English language learning in the ESL program. Thus, this gap in the literature needs to be addressed to examine whether there is consistency in the results and effects of flipped learning in ESL subjects and for engineering and mathematics subjects. The flipped learning approach's effectiveness needs to be studied by developing a model framework in the ESL program at universities.

This study stands on the premise that besides the recorded positive advancement and development, the flipped learning approach has on teaching and learning of the 21st century. There exists minimal proof of a set perimeter to guide the use of the approach in an educational environment effectively (Baepler et al., 2014; Bishop & Verleger, 2013; O'Flaherty & Phillips, 2015). Thus, a conscious effort to establish an instrument for flipped learning has become a primary concern to implement the approach effectively. Experts' perspective of constructs recommended for developing an instrument for flipped learning is invaluable as their professional experience and knowledge on the subject matter should be pivotal in determining that such development is on the right path. The study sees the experts' perspective on the proposed constructs for developing an instrument to implement the flipped learning approach in an ESL environment. Thus, seven constructs were identified from Hamdan et al. (2013) and Chen et al. (2014), that were mapped out into sixty-eight items and used in the form of a questionnaire posted in the form of Google docs and distributed to twenty-two experts of educational technology or ESL and educational technology. The study received eighteen responses, analyzed by the Fuzzy Delphi method. This research aims to see the experts' view of the proposed constructs and quantify these views in the form of Fuzzy Delphi analysis.

RESEARCH METHOD

This study uses a design and developmental research (DDR) approach to develop and verify a flipped learning framework in an ESL context. There are two phases of the study. The first phase focuses on the design of construct for flipped learning framework utilizing literature review concerning the flipped learning approach. The second phase is on the development of the flipped learning construct, starting with the Fuzzy Delphi method.

The samples in this study were selected using the purposive sampling technique. Purposive sampling refers to a sampling procedure in which subjects that have ascertained specific characteristics needed for the research are selected as respondents in the study (Creswell, 2009), and it does not require underlying theories or a certain number of informants (Patton, 2002; Tongco, 2007) as its foundation. It is a “deliberate choice of an informant based on the qualities the informant possesses” (Tongco, 2007, p. 147) and is used in collecting the quantitative (Fuzzy Delphi and survey) data of the study.

There are two groups of samples used in this study. The first group involves a group of 18 experts in the education technology learning field. The experts' responses are gathered during the Delphi technique to obtain the second objective of the study.

Instruments

In the first phase, the researchers used the input or process to design the flipped learning questionnaire for the experts. The input from the literature provided a list of construct, dimensions, and items concerning ESL flipped learning at Universiti Kebangsaan Malaysia. Then, the list was converted into several statements to form an ESL flipped learning questionnaire for the experts to review.

The Fuzzy Delphi Method (FDM) was used to gather the experts' agreement on the dimensions and indicators. FDM aims to solve the problem of traditional Delphi method (Hidayat & Lawahid, 2020; Ishikawa et al., 1993). The method is based on group thinking of qualified experts to ensure the validity of collected data. In FDM, the experts were required to indicate the extent of their agreement with the statements. Also, the experts were encouraged to introduce or recommend any new dimensions or indicators, revise or make an adjustment to the existing statements in the list. Before the actual FDM, the questionnaire went through several face validation processes by the researchers' supervisors and experts identified.

Procedure

The data collection for this study consists of three phases. Each phase is elaborated as follows.

Phase I: Flipped Learning Framework Design in Context

The first phase involved a review of related past studies. The review was done to establish the constructs and items needed for the development of the questionnaire to develop the Flipped learning framework. The researchers conducted this by mapping out the literature to the constructs proposed in the study by Hamdan et al. (2013) and Chen et al. (2014). Once the construct, dimensions, and items were established, the questionnaire was developed and given to two experts to review its content and face validity. After discussions and reviews of the questionnaire, the questionnaire was ready for the next step.

Phase II: Flipped Learning Framework Development

The next step is to identify experts or participants for the Fuzzy Delphi Method (FDM). After the experts expressed their agreement to participate, emails with appointment letters signed by the researchers' supervisor were sent to the experts. The experts were then directed

to a Google form website address to answer the questionnaire and give their opinions on the questionnaire. The findings from FDM were used to help in enhancing and developing the flipped learning questionnaire and framework.

Data Analysis

The first quantitative data of the study were collated from the experts in the second phase using the Fuzzy Delphi Method (FDM). FDM is conducted using the questionnaire instrument, which has a five-point Likert scale of importance ranging from ‘Strongly Important’ (5) to ‘Strongly Unimportant’ (1). FDM is chosen for this analysis as the method has been proven to produce statistically valid constructs, dimensions, and items for many previous kinds of research (Bouzon et al., 2016). Furthermore, the Defuzzification α -cut analysis and the (d) threshold value in accepting and rejecting items and constructs in developing frameworks and models were well established and valid (Hidayat, 2018). In terms of showing a consensus, FDM analysis can demonstrate this effectively (Sanchez-Lezama et al., 2014). This first quantitative analysis establishes the constructs, dimensions, and items of the study.

FINDINGS AND DISCUSSION

What Constructs Should Make Up a Framework for Assessing Flipped Learning Approach Effectiveness in an ESL Context, Based on Literature, in the Design Phase?

The overall findings from the literature analysis done for Phase I points to the importance of the seven constructs (flexible learning environment, shift in learning culture, intentional content, professional educators, progressive networking activities, engaging and effective learning experiences, diversified seamless learning platforms) in discussing what is important for the flipped learning approach implementation by researchers investigating the flipped learning approach in the educational environments. The constructs are essential but not apparent and are usually discussed directly or indirectly, either in the literature discussions or in the conclusions made. This may be due to the approach novelty, as many academics are still investigating what important factors contribute to the effective implementation of the flipped learning approach in educational environments. However, the framework can be divided into two main themes: the educator’s element and the student’s element, which are further divulged in this section.

The researchers have identified and organized several research papers that discuss one or several constructs as important elements to be considered in implementing and administering the flipped learning approach. To note, the importance of the educator’s elements that make up an integral part of the framework in ensuring an effective flipped learning approach classes have been discussed at length (Bishop & Verleger, 2013; Chen et al., 2014; Coufal, 2014; Embi, 2014; Hamdan et al., 2013; Hao, 2016; O’Flaherty & Phillips, 2015). The approach of providing flexible learning modes physical learning space in encouraging students to be engaged in a meaningful lesson cannot be undermined (Balan et al., 2015; Bergmann & Sams, 2012; Yemma, 2015). Hao (2016) stated that the implementation of the flipped learning approach is drawing attention from educators and students alike due to the shift in the learning culture where the main onus to learn is on the students making meaning of the lessons on a personal level, with independence and peer learning. The usage of content developed and customized according to the needs of the flipped environment to ensure meaningful lessons cannot be underestimated (Bergmann & Sams, 2012; Wiginton, 2013). Furthermore, the presence of educators who are aware of the job scope and responsibilities of a flipped learning educator is imperative in determining the success of the implementation (Nederveld & Berge, 2015; O’Flaherty & Phillips, 2015; Yemma, 2015).

The literature evidence pointing towards the importance of the student's element in ensuring a successful flipped learning curriculum has been the focal point in academic discussions (Chen et al., 2014; Embi et al., 2014). The usage of activities that encapsulate networking through the use of technology is integral in ensuring effective lessons that propagate self-discovery learning and peer learning for students of the 21st century (Aw-Yong et al., 2013; Nederveld & Berge, 2015; O'Flaherty & Phillips, 2015; Yemma, 2015). Many research papers have concluded from their analysis that careful planning and thought should be invested in designing engaging and effective learning experiences as it is an integral part of determining the success or failure of the approach in classroom settings (Bergmann & Sams, 2012; Bishop & Verleger, 2013; Soltanpour & Valizadech in Bodomo, 2016; Yemma, 2015). Moreover, the idea of using multiple platforms online for learning is welcomed and seen as an effective and engaging medium to inculcate ideas of discovery learning, peer learning, ubiquitous learning, and many other important facets of 21st-century learning or learning characteristics in the Industrial Revolution 4.0 era. These traits are important characters needed for the students to become successful in the Industrial Revolution 4.0 era (Soltanpour & Valizadech in Bodomo, 2016; Coufal, 2014; Kafi & Motallebzadeh, 2014; Wiginton, 2013; Yemma, 2015).

All these research works have, one way or another, mentioned the importance of the factors proposed as the constructs of the framework suggested. What is missing from the literature so far are investigations that look at the factors together and deliberate the influence and effect these factors have comprehensively on the teaching and learning processes in the Flipped learning approach or environment. Hence, the study investigates the interactions of the factors proposed and their influence in an ESL context in Malaysia. All these researches point to the importance of the constructs proposed by this study in developing and ensuring that any implementation of the Flipped learning approach must have a glance of the factors in molding the lessons or curriculum with Flipped learning in their fore. The researchers also point to the need for a framework to implement the flipped learning approach. Currently, there are only mentions of the important factors to consider. However, the presence of a proper and comprehensive perimeter or framework that guides the implementation of the approach ensures an effective and meaningful learning experience is currently lacking from the literature.

How Effectively and Accurately Can These Constructs Be Determined to Measure the Flipped Learning Framework in an ESL Context in the Development Phase?

Each construct is mapped out to items representing it and put forth in the questionnaire form to the experts identified. The constructs and items are deduced from the literature review and mapped accordingly to form the basis of this study. The seven constructs identified are: (1) flexible environments (FE), (2) shift in learning culture (LC), (3) intentional content (IC), (4) professional educators (PE), (5) progressive networking activities (NA), (6) engaging and effective learning experiences (LE), and (7) diversified seamless learning platforms (LP).

These seven constructs concern the teaching and learning process and the student's experience of the technology-based learning approach. These constructs will be the foundation of the items built and analyzed with the Fuzzy Delphi method. For deliberation and discussion of the findings, the (d) threshold value of the constructs and items and the percentage of experts' agreement are discussed in this section. The bench mark benchmark acceptance of a construct or the items will be ≥ 0.2 for the (d) threshold value and 75% for the percentage of experts' agreement. The results of the analysis are as follows.

Construct of Flexible Environment (FE)

Table 1 deliberates the Fuzzy Delphi calculations of the expert's perspective on "Flexible Environment" construct and its items. The individual item's (d) threshold value is: 0.200

(FE 1), 0.150 (FE 2), 0.186 (FE 3), 0.169 (FE 4), 0.212 (FE 6), 0.167 (FE 7), 0.147 (FE 8), and 0.184 (FE 9). The percentages of experts' agreement of the individual items are: 88.88% (FE 1), 100% (FE 2), 88.88% (FE 3), 94.44% (FE 4), 94.44% (FE 6), 100% (FE 7), and 88.88% (FE 8). Thus, these items have met the benchmark value of the (d) threshold and percentages of experts' agreement mentioned earlier, and are accepted by the experts. Item FE 5 of the construct has been rejected as the (d) threshold value (0.212). The experts' agreement percentage (33.33%) met the benchmark value. Overall, the "Flexible Environment" construct has a (d) threshold value of 0.180 and overall percentage of experts' agreement of 86.10%, which leads to the item being accepted by the experts as congruent and important for this study.

Table 1. Threshold Value (d), Percentage of Experts' Consensus, and Defuzzification of the Flexible Environment Construct (FE)

Experts	Items							
	FE 1	FE 2	FE 3	FE 4	FE 5	FE 6	FE 7	FE 8
1	0.2	0.1	0.2	0.1	0.3	0.2	0.1	0.1
2	0.1	0.2	0.4	0.2	0.3	0.1	0.2	0.1
3	0.1	0.2	0.1	0.2	0.3	0.1	0.1	0.5
4	0.2	0.1	0.1	0.1	0.3	0.1	0.2	0.2
5	0.1	0.1	0.2	0.2	0.3	0.2	0.1	0.1
6	0.2	0.1	0.2	0.1	0.0	0.2	0.2	0.1
7	0.2	0.1	0.4	0.1	0.3	0.1	0.1	0.1
8	0.2	0.1	0.2	0.1	0.0	0.1	0.1	0.1
9	0.2	0.1	0.2	0.1	0.3	0.2	0.1	0.1
10	0.1	0.2	0.1	0.2	0.3	0.4	0.2	0.2
11	0.1	0.2	0.4	0.1	0.3	0.1	0.2	0.5
12	0.1	0.2	0.1	0.5	0.3	0.1	0.2	0.2
13	0.7	0.2	0.1	0.1	0.0	0.2	0.1	0.1
14	0.4	0.2	0.1	0.2	0.0	0.1	0.2	0.2
15	0.1	0.1	0.2	0.1	0.3	0.2	0.1	0.1
16	0.2	0.1	0.2	0.1	0.3	0.2	0.1	0.1
17	0.2	0.1	0.1	0.2	0.0	0.2	0.1	0.1
18	0.2	0.2	0.1	0.2	0.0	0.1	0.1	0.1
d value for each item	0.200	0.150	0.186	0.169	0.212	0.167	0.147	0.184
d value of the construct				0.177				
Number of Item d ≤ 0.2	16	18	16	17	6	17	18	16
Percentage of Item d ≤ 0.2	88.88	100	88.88	94.44	33.33	94.44	100	88.88
Percentage of Construct				86.10				
Fuzzy Evaluation	12.000	12.800	11.600	12.600		12.200	13.000	12.800
Average of Fuzzy Number	0.667	0.711	0.644	0.700	Reject	0.678	0.722	0.711
Rank	6	2	7	4		5	1	3

Construct of Shift in Learning Culture (LC)

Table 2 maps-out the results of the Fuzzy-Delphi analysis for the construct of "Shift in Learning Culture". The individual items' (d) threshold values are: 0.172 (LC 1), 0.129 (LC 2), 0.171 (LC 3), 0.136 (LC 4), 0.200 (LC 5), 0.161 (LC 6), 0.071 (LC 8), and 0.143 (LC 9), while the experts' agreement percentage of each items are 94.44% (LC 1), 100% (LC 2), 94.44% (LC 3), 100% (LC 4), 77.77% (LC 5), 88.88% (LC 6), 83.33% (LC 8) and 100% (LC 9). The experts have approved these items as important for this particular construct. Item LC7, however, is rejected as the (d) threshold value, and is at 0.142. The experts' percentage of agreement is at 66.67%. Therefore, it leads to the rejection of the item, because it fails to achieve the benchmark values of the analysis.

In addition, the overall (d) threshold value and the percentage of experts' agreement of the construct "Shift in Learning Culture" as a whole are at 0.149 and 89.50%, respectively. This results in the construct being acknowledged as compatible and important for developing the instrument for gauging the flipped learning efficiency.

Table 2. Threshold Value (d), Percentage of Experts' Consensus, and Defuzzification of the Shift in Learning Culture (LC)

Experts	Items								
	LC 1	LC 2	LC 3	LC 4	LC 5	LC 6	LC 7	LC 8	LC 9
1	0.2	0.1	0.1	0.2	0.2	0.2	0.4	0.3	0.1
2	0.4	0.2	0.2	0.1	0.1	0.1	0.1	0.0	0.2
3	0.1	0.2	0.1	0.1	0.2	0.1	0.3	0.0	0.1
4	0.2	0.1	0.2	0.1	0.1	0.1	0.1	0.0	0.2
5	0.2	0.1	0.1	0.2	0.2	0.1	0.3	0.0	0.1
6	0.2	0.1	0.1	0.1	0.4	0.1	0.1	0.0	0.1
7	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.0	0.1
8	0.2	0.1	0.1	0.2	0.2	0.2	0.1	0.0	0.1
9	0.2	0.1	0.2	0.1	0.2	0.2	0.4	0.3	0.1
10	0.1	0.2	0.2	0.1	0.4	0.1	0.1	0.0	0.1
11	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.0	0.2
12	0.1	0.2	0.5	0.1	0.1	0.1	0.3	0.0	0.2
13	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.0	0.1
14	0.1	0.2	0.2	0.1	0.4	0.4	0.1	0.0	0.2
15	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.0	0.1
16	0.2	0.1	0.1	0.1	0.4	0.4	0.3	0.3	0.2
17	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1
18	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1
d value for each item	0.172	0.129	0.171	0.136	0.200	0.161	0.142	0.071	0.143
d value of the construct					0.149				
Number of Item $d \leq 0.2$	17	18	17	18	14	16	12	15	18
Percentage of each Item $d \leq 0.2$	94.44	100	94.44	100	77.77	88.88	66.67	83.33	100
Percentage of Construct					89.50				
Fuzzy Evaluation	12.600	13.400	12.800	12.200	11.600	11.600		11.200	13.200
Average of Fuzzy Number	0.700	0.744	0.711	0.678	0.644	0.644	Reject	0.622	0.733
Rank	4	1	3	5	6	7		8	2

Construct of Intentional Content (IC)

Table 3. Threshold Value (d), Percentage of Experts' Consensus, and Defuzzification of Intentional Content (IC)

Experts	Items									
	IC 1	IC 2	IC 3	IC 4	IC 5	IC 6	IC 7	IC 8	IC 9	IC 10
1	0.0	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2
2	0.3	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.1
3	0.0	0.1	0.1	0.2	0.1	0.2	0.2	0.2	0.1	0.2
4	0.0	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.1
5	0.3	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.1
6	0.3	0.1	0.1	0.2	0.2	0.1	0.2	0.2	0.2	0.2
7	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.1
8	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2
9	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2
10	0.0	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.4	0.1
11	0.0	0.1	0.1	0.2	0.4	0.1	0.2	0.1	0.1	0.4
12	0.6	0.1	0.1	0.2	0.7	0.1	0.2	0.1	0.4	0.1
13	0.0	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.2	0.2
14	0.3	0.2	0.2	0.1	0.2	0.4	0.2	0.4	0.1	0.1
15	0.0	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2
16	0.6	0.4	0.4	0.2	0.1	0.1	0.2	0.4	0.4	0.4
17	0.6	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.2
18	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.1
d value for each item	0.224	0.167	0.158	0.151	0.185	0.147	0.153	0.187	0.200	0.181
d value of the construct					0.176					
Number of Item $d \leq 0.2$	8	18	17	18	16	17	18	16	15	16
Percentage of each Item $d \leq 0.2$	44.44	100.00	94.44	100.00	88.89	94.44	100.00	88.89	83.33	88.27
Percentage of Construct					88.27					
Fuzzy Evaluation		12.267	12.000	12.800	11.600	11.800	12.600	12.200	12.200	12.000
Average of Fuzzy Number	Reject	0.681	0.667	0.711	0.644	0.656	0.700	0.678	0.678	0.667
Rank		3	6	1	9	8	2	4	5	7

The “Intentional Content” construct has ten items investigated in this study, and the Fuzzy-Delphi analysis of each item is shown in Table 3. The (d) threshold values of each item accepted by the experts are: 0.167 (IC 2), 0.158 (IC 3), 0.151 (IC 4), 0.185 (IC 5), 0.147 (IC 6), 0.153 (IC 7), 0.187 (IC 8), 0.200 (IC 9), and 0.181 (IC 10), while the experts’ agreement percentages of the accepted items are 100% (IC 2), 94.44% (IC 3), 100% (IC 4), 88.89% (IC 5), 94.44% (IC 6), 100% (IC 7), 88.89% (IC 8), 83.33% (IC 9), and 88.27% (IC 10). The rejected item of IC 1 has a (d) threshold value of 0.244 and experts’ percentage of agreement of 44.44. Therefore, the item is rejected as being representative of the construct in question. The overall (d) threshold value and the experts’ agreement percentage are 0.175 and 88.27%, which lead to the construct acceptance by the experts as a part of the development of a flipped learning instrument.

Construct of Progressive Networking Activities (NA)

Table 4 details the results of Fuzzy-Delphi analysis of “Progressive Networking Activities” construct. The (d) threshold results for each item are: 0.151 (NA 1), 0.185 (NA 2), 0.170 (NA 3), 0.158 (NA 4), 0.132 (NA 5), 0.172 (NA 6), 0.181 (NA 7), and 0.166 (NA 8). The experts’ agreement percentages of each item are: 100% (NA 1), 94.4% (NA 2), 94.4% (NA 3), 88.9% (NA 4), 94.4% (NA 5), 88.9% (NA 6), 88.9% (NA 7), and 94.4% (NA 8). The experts rejected no items for this construct.

The overall (d) threshold value is 0.162, and the percentage of experts’ agreement of the construct is at 90.30%. This shows a strong agreement by the experts in accepting the aforementioned construct as an essential part of the instrument development for analyzing flipped learning.

Table 4. Threshold Value (d), Percentage of Experts’ Consensus, and Defuzzification of Progressive Networking Activities (NA)

Experts	Items							
	NA 1	NA 2	NA 3	NA 4	NA 5	NA 6	NA 7	NA 8
1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2
2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
3	0.1	0.1	0.2	0.1	0.1	0.4	0.1	0.1
4	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2
5	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2
6	0.1	0.2	0.2	0.1	0.1	0.2	0.2	0.2
7	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
8	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2
9	0.1	0.2	0.2	0.1	0.1	0.2	0.2	0.2
10	0.2	0.1	0.1	0.1	0.1	0.1	0.4	0.1
11	0.2	0.4	0.1	0.4	0.1	0.1	0.1	0.1
12	0.2	0.4	0.1	0.1	0.1	0.1	0.1	0.4
13	0.1	0.2	0.2	0.2	0.2	0.1	0.1	0.1
14	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.1
15	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2
16	0.2	0.1	0.2	0.1	0.1	0.1	0.1	0.1
17	0.1	0.2	0.2	0.2	0.2	0.1	0.2	0.2
18	0.1	0.1	0.1	0.2	0.1	0.2	0.2	0.1
d value for each item	0.151	0.185	0.170	0.158	0.132	0.172	0.181	0.166
d value of the construct	0.162							
Number of Item $d \leq 0.2$	18	17	17	16	17	16	16	17
Percentage of each Item $d \leq 0.2$	100.0	94.4	94.4	88.9	94.4	88.9	88.9	94.4
Percentage of Construct	90.3							
Fuzzy Evaluation	12.800	11.600	12.400	11.600	11.600	11.800	12.000	12.200
Average of Fuzzy Number	0.711	0.644	0.689	0.644	0.644	0.656	0.667	0.678
Rank	1	6	2	8	7	5	4	3

Construct of Professional Educators (PE)

Table 5 details the analysis for “Professional Educators” construct. The (d) threshold value for each accepted items are: 0.140 (PE 1), 0.153 (PE 3), 0.165 (PE 4), 0.169 (PE 5), 0.147 (PE 6), 0.195 (PE 9), 0.152 (PE 10), 0.152 (PE 11), 0.171 (PE 12), 0.165 (PE 13), 0.153 (PE 14), and 0.156 (PE 15). Meanwhile, their respective percentages of experts’ agreement are: 100% (PE 1), 100% (PE 3), 94.4% (PE 4), 88.9% (PE 5), 100% (PE 6), 88.9% (PE 9), 100% (PE 10), 100% (PE 11), 94.4% (PE 12), 94.4% (PE 13), 100% (PE 14), and 94.4% (PE 15).

The (d) threshold values of the items rejected are 0.170 (PE 2), 0.154 (PE 7), and 0.112 (PE 8). Their respective percentages of experts’ agreement are 38.9% (PE 2), 55.6% (PE 7), and 66.7%. (PE 8). As such, only three items out of fifteen for this construct are rejected by the experts. Conclusively, the construct's overall construct (d) threshold value is 0.157, and the percentage of experts’ agreement for the construct is 87.8%. This shows the acceptance of the construct by the experts for the development of the flip learning instrument.

Table 5. Threshold Value (d), Percentage of Experts’ Consensus, and Defuzzification of the Construct Professional Educators (PE)

Experts	Items														
	PE1	PE2	PE3	PE4	PE5	PE6	PE7	PE8	PE9	PE10	PE11	PE12	PE13	PE14	PE15
1	0.1	0.0	0.1	0.2	0.2	0.1	0.3	0.3	0.2	0.1	0.1	0.2	0.2	0.2	0.2
2	0.1	0.0	0.1	0.1	0.1	0.2	0.1	0.0	0.1	0.2	0.2	0.1	0.1	0.1	0.1
3	0.1	0.0	0.2	0.4	0.4	0.2	0.3	0.0	0.1	0.1	0.2	0.1	0.1	0.2	0.2
4	0.1	0.0	0.1	0.2	0.1	0.2	0.1	0.0	0.1	0.2	0.2	0.1	0.1	0.1	0.1
5	0.2	0.0	0.2	0.1	0.1	0.1	0.1	0.0	0.2	0.2	0.1	0.1	0.2	0.2	0.1
6	0.2	0.3	0.2	0.2	0.1	0.2	0.1	0.3	0.7	0.2	0.1	0.1	0.2	0.1	0.2
7	0.1	0.0	0.1	0.1	0.2	0.1	0.1	0.0	0.2	0.1	0.2	0.2	0.1	0.1	0.1
8	0.2	0.3	0.2	0.2	0.2	0.1	0.3	0.3	0.2	0.1	0.1	0.2	0.2	0.2	0.2
9	0.2	0.3	0.2	0.2	0.2	0.1	0.3	0.3	0.2	0.1	0.1	0.2	0.2	0.2	0.2
10	0.1	0.3	0.1	0.1	0.1	0.2	0.1	0.3	0.4	0.1	0.1	0.2	0.1	0.1	0.1
11	0.1	0.3	0.1	0.1	0.1	0.2	0.1	0.0	0.1	0.2	0.2	0.1	0.4	0.1	0.4
12	0.1	0.0	0.1	0.1	0.1	0.2	0.1	0.0	0.1	0.2	0.2	0.4	0.1	0.1	0.1
13	0.1	0.0	0.2	0.2	0.2	0.1	0.1	0.0	0.1	0.2	0.1	0.2	0.1	0.1	0.1
14	0.1	0.3	0.1	0.1	0.4	0.1	0.4	0.0	0.1	0.1	0.2	0.1	0.1	0.2	0.1
15	0.2	0.3	0.2	0.2	0.2	0.1	0.3	0.0	0.2	0.1	0.1	0.2	0.2	0.2	0.1
16	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.2	0.2	0.2	0.1	0.1	0.1
17	0.2	0.3	0.2	0.1	0.1	0.1	0.4	0.0	0.2	0.1	0.1	0.2	0.2	0.2	0.2
18	0.2	0.3	0.2	0.2	0.2	0.1	0.3	0.3	0.2	0.1	0.1	0.2	0.2	0.2	0.2
d value for each item	0.140	0.170	0.153	0.165	0.169	0.147	0.154	0.112	0.195	0.152	0.152	0.171	0.165	0.153	0.156
d value of the construct	0.157														
Number of Item d ≤ 0.2	18	7	18	17	16	18	10	12	16	18	18	17	17	18	17
Percentage of each Item d ≤ 0.2	100.0	38.9	100.0	94.4	88.9	100.0	55.6	66.7	88.9	100.0	100.0	94.4	94.4	100.0	94.4
Percentage of Construct	87.8														
Fuzzy Evaluation	12.200		12.600	12.200	11.800	13.000			11.800	12.800	12.800	12.600	12.200	12.600	12.000
Average of Fuzzy Number	0.678	Reject	0.700	0.678	0.656	0.722	Reject	Reject	0.656	0.711	0.711	0.700	0.678	0.700	0.667
Rank	10		5	7	11	1			12	3	2	6	8	4	9

Construct of Engaging and Effective Learning Experiences (LE)

Table 6 entails the (d) threshold values and the percentage of experts’ agreement of each item and the construct of “Engaging and Effective Learning Experiences” as a whole. To begin with, the (d) threshold values of each accepted item is: 0.145 (LE 1), 0.187 (LE 2), 0.181 (LE 3), 0.187 (LE 4), 0.211 (LE 5), 0.196 (LE 6), 0.187 (LE 8), 0.172 (LE 9), 0.181 (LE 10), and 0.196 (LE 11). The percentages of experts’ agreement of the accepted items are: 100.0% (LE 1), 94.4% (LE 2), 88.9% (LE 3), 88.9% (LE 4), 83.3% (LE 5), 88.9% (LE 6), 88.9% (LE 8), 88.9% (LE 9), 88.9% (LE 10), and 88.9% (LE 11).

Item 7 of the construct (LE 7) is rejected because the (d) threshold value is 0.24, more than the 0.2 benchmark value, and 16.7% of the percentage of experts’ agreement, which is below the 75% benchmark. The overall construct (d) threshold value stands at 0.189; meanwhile, the overall percentage of experts’ agreement is at 83.3 %. Hence, the experts accept the construct of Engaging and Effective Learning Experiences as important in developing a flip learning instrument.

Table 6. Threshold Value (d), Percentage of Experts' Consensus, and Defuzzification of the Construct Engaging and Effective Learning Experiences (LE)

Experts	Items										
	LE 1	LE 2	LE 3	LE 4	LE 5	LE 6	LE 7	LE 8	LE 9	LE 10	LE 11
1	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2
2	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1
3	0.1	0.1	0.1	0.1	0.1	0.1	0.3	0.4	0.4	0.1	0.1
4	0.1	0.2	0.1	0.2	0.2	0.2	0.3	0.1	0.1	0.1	0.1
5	0.1	0.2	0.2	0.2	0.2	0.2	0.0	0.2	0.1	0.2	0.2
6	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2
7	0.2	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1
8	0.2	0.2	0.1	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2
9	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2
10	0.1	0.1	0.1	0.4	0.4	0.1	0.3	0.1	0.1	0.1	0.1
11	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1
12	0.1	0.1	0.4	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.1
13	0.2	0.2	0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.7
14	0.1	0.4	0.1	0.1	0.1	0.2	0.0	0.2	0.1	0.4	0.1
15	0.1	0.2	0.2	0.2	0.2	0.1	0.3	0.2	0.2	0.2	0.2
16	0.1	0.4	0.4	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.4
17	0.2	0.2	0.2	0.2	0.7	0.7	0.6	0.2	0.2	0.2	0.2
18	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2
d value for each item	0.145	0.187	0.181	0.187	0.201	0.196	0.241	0.187	0.172	0.181	0.196
d value of the construct						0.189					
Number of Item d ≤ 0.2	18	17	16	16	15	16	3	16	16	16	16
Percentage of each Item d ≤ 0.2	100.0	94.4	88.9	88.9	83.3	88.9	16.7	88.9	88.9	88.9	88.9
Percentage of Construct						83.3					
Fuzzy Evaluation	12.200	12.200	12.000	12.200	11.600	11.800		12.200	11.800	12.000	11.800
Average of Fuzzy Number	0.678	0.678	0.667	0.678	0.644	0.656	Reject	0.678	0.656	0.667	0.656
Rank	1	2	5	4	10	8		3	7	6	9

Construct of Diversified Seamless Learning Platforms (DP)

Table 7 shows the Fuzzy-Delphi analysis of “Diversified Seamless Platform” construct. There are no rejected items, which entails that all seven items are accepted and viewed important by the experts. The (d) threshold value of each item is: 0.193 (DP 1), 0.163 (DP 2), 0.152 (DP 3), 0.147 (DP 4), 0.147 (DP 5), 0.190 (DP 6), and 0.200 (DP 7). Meanwhile, the percentages of experts' agreement of the items are: 83.3% (DP 1), 94.4% (DP 2), 100.0% (DP 3), 100.0% (DP 4), 100.0% (DP 5), 88.9% (DP 6), and 83.3% (DP 7).

The overall (d) threshold value for this construct is 0.171, and the percentage of experts' agreement is 92.9%. This infers the acceptance of the experts and the importance of the construct in the development of the flip learning instrument.

The results of this study point to the acceptance and acknowledgment of the experts of the constructs proposed originally by [Chen et al. \(2014\)](#) and [Hamdan et al. \(2013\)](#) as being important seven constructs to be considered for the development of a flipped learning instrument in an ESL context. These constructs put the idea of the technology, the pedagogies, and the people, educators, and students experience in a continuum of teaching and learning spectrum. These constructs interact to create an instrument that serves as a parameter-gauger for the effective execution of the flipped learning approach in the ESL context.

The study consists of seven constructs that make the basis of the sixty-eight items used to solicit the experts' perspectives on these constructs. Out of the sixty-eight items asked, the experts rejected seven items in relation to the seven constructs. Overall, the experts agreed that the proposed constructs can be used to develop an instrument to check the efficiency of flipped learning in the ESL context. Hence, an instrument to gauge the effectiveness of the flipped learning approach in an ESL context is established as a result of the research.

Table 7. Threshold Value (d), Percentage of Experts' Consensus, and Defuzzification of the Construct Diversified Seamless Learning Platforms (DP)

Experts	Items						
	DP1	DP2	DP3	DP4	DP5	DP6	DP7
1	0.2	0.2	0.1	0.1	0.1	0.2	0.2
2	0.1	0.1	0.2	0.2	0.2	0.1	0.1
3	0.4	0.4	0.2	0.2	0.2	0.1	0.1
4	0.1	0.1	0.2	0.1	0.1	0.2	0.2
5	0.2	0.2	0.1	0.1	0.1	0.2	0.2
6	0.1	0.1	0.1	0.1	0.1	0.2	0.2
7	0.1	0.1	0.1	0.1	0.1	0.2	0.2
8	0.2	0.2	0.1	0.1	0.1	0.2	0.2
9	0.2	0.2	0.1	0.1	0.1	0.2	0.2
10	0.1	0.1	0.2	0.2	0.1	0.1	0.1
11	0.4	0.1	0.2	0.2	0.2	0.1	0.1
12	0.1	0.1	0.2	0.2	0.2	0.4	0.4
13	0.2	0.2	0.1	0.1	0.2	0.1	0.1
14	0.1	0.1	0.2	0.2	0.2	0.1	0.4
15	0.2	0.2	0.1	0.1	0.1	0.2	0.2
16	0.4	0.1	0.2	0.2	0.2	0.4	0.4
17	0.2	0.2	0.1	0.1	0.1	0.2	0.2
18	0.2	0.2	0.1	0.1	0.1	0.2	0.2
d value for each item	0.193	0.163	0.152	0.147	0.147	0.190	0.200
d value of the construct				0.171			
Number of Item $d \leq 0.2$	15	17	18	18	18	16	15
Percentage of each Item $d \leq 0.2$	83.3	94.4	100.0	100.0	100.0	88.9	83.3
Percentage of Construct				92.9			
Fuzzy Evaluation	11.800	12.200	12.800	13.000	13.000	12.400	12.200
Average of Fuzzy Number	0.656	0.678	0.711	0.722	0.722	0.689	0.678
Rank	7	5	3	2	1	4	6

CONCLUSION

Conclusively, the study identified the required constructs in developing an instrument for flipped learning in an ESL environment. Establishing the constructs can trail blaze investigations that lead to developing a framework or model to gatekeep the effective implementation of the flipped learning approach in general or even specific contexts.

The implication of the study can be seen in multiple facets. The first facet is for the policymakers and educational governing bodies. The identification of the constructs means the relevant bodies can now rely on these constructs in guiding and determining parameters needed for effective implementation of the flipped learning approach, especially in the ESL context. Furthermore, curriculum developers and teachers can use these constructs and their items to ensure their flipped learning approach classes are seen as meaningful and relevant by the students in developing their knowledge in a technology-supported environment. Secondly, these constructs can be used as the basis for further research to the development of established frameworks and models for the effective and meaningful flipped learning approach lessons.

REFERENCES

- Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. *Computers & Education*, *126*, 334–345. <https://doi.org/10.1016/j.compedu.2018.07.021>
- Aw-Yong, J., Anderson, N., & Chigeza, P. (2013). Developing culturally-responsive lessons on the iPad for teaching English as Second Language to Chinese learners. *2013 IEEE*

63rd Annual Conference International Council for Education Media (ICEM), 1–11.
<https://doi.org/10.1109/CICEM.2013.6820186>

- Azman, H., & Dollsaid, N. F. (2018). Applying massively multiplayer online games (MMOGs) in EFL teaching. *Arab World English Journal*, 9(4), 3–18.
<https://doi.org/10.24093/awej/vol9no4.1>
- Baepler, P., Walker, J. D., & Driessen, M. (2014). It's not about seat time: Blending, flipping, and efficiency in active learning classrooms. *Computers & Education*, 78, 227–236.
<https://doi.org/10.1016/j.compedu.2014.06.006>
- Balan, P., Clark, M., & Restall, G. (2015). Preparing students for flipped or team-based learning methods. *Education + Training*, 57(6), 639–657. <https://doi.org/10.1108/ET-07-2014-0088>
- Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education.
- Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. *120th ASEE Annual Conference & Exposition*, 6219. <https://peer.asee.org/the-flipped-classroom-a-survey-of-the-research>
- Bodomo, A. (2016). Afriphone literature as a prototypical form of African literature: Insights from prototype theory. *Advances in Language and Literary Studies*, 7(5), 262–267.
<http://www.journals.aiac.org.au/index.php/all/article/view/2748>
- Bouzon, M., Govindan, K., Rodriguez, C. M. T., & Campos, L. M. S. (2016). Identification and analysis of reverse logistics barriers using Fuzzy Delphi method and AHP. *Resources, Conservation and Recycling*, 108, 182–197. <https://doi.org/10.1016/j.resconrec.2015.05.021>
- Chen, Y., Wang, Y., Kinshuk, & Chen, N.-S. (2014). Is FLIP enough? Or should we use the FLIPPED model instead? *Computers & Education*, 79, 16–27.
<https://doi.org/10.1016/j.compedu.2014.07.004>
- Coufal, K. (2014). *Flipped learning instructional model: Perceptions of video delivery to support engagement in eighth grade math*. Doctoral thesis, Lamar University, Beaumont, TX.
- Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publication.
- Davies, R. S., Dean, D. L., & Ball, N. (2013). Flipping the classroom and instructional technology integration in a college-level information systems spreadsheet course. *Educational Technology Research and Development*, 61(4), 563–580.
<https://doi.org/10.1007/s11423-013-9305-6>
- Embi, M. A., Hussin, S., & Panah, E. (2014). Flipped learning readiness amongst graduate and postgraduate students in UKM. In M. A. Embi (Ed.), *Blended and flipped learning: Case studies in Malaysia HEIs* (pp. 209–223). Centre for Teaching and Learning Technologies, Universiti Kebangsaan Malaysia.
- Embi, M. A. (2014). *Blended & flipped learning: Case studies in Malaysian HEIs* (M. A. Embi (ed.); 1st ed.). Universiti Kebangsaan Malaysia.
- Enfield, J. (2013). Looking at the impact of the flipped classroom model of instruction on undergraduate multimedia students at CSUN. *TechTrends*, 57(6), 14–27.
<https://doi.org/10.1007/s11528-013-0698-1>
- Hamdan, N., McKnight, P., McKnight, K., & Arfstrom, K. (2013). *A review of flipped learning: Flipped learning network*. Flipped Learning Network.

https://flippedlearning.org/cms/lib07/VA01923112/Centricity/Domain/41/LitReview_FlippedLearning.pdf

- Hao, Y. (2016). Exploring undergraduates' perspectives and flipped learning readiness in their flipped classrooms. *Computers in Human Behavior*, 59, 82–92. <https://doi.org/10.1016/j.chb.2016.01.032>
- Hidayat, W. (2018). *Developing and validating an inventory of a national character (IKB) for secondary school students*. Doctoral thesis, Universiti Kebangsaan Malaysia, Selangor.
- Hidayat, W., & Lawahid, N. A. (2020). *Metode Fuzzy Delphi untuk penelitian sosial*. Alfabeta.
- Ishikawa, A., Amagasa, M., Shiga, T., Tomizawa, G., Tatsuta, R., & Mieno, H. (1993). The max-min Delphi method and Fuzzy Delphi method via fuzzy integration. *Fuzzy Sets and Systems*, 55(3), 241–253. [https://doi.org/10.1016/0165-0114\(93\)90251-C](https://doi.org/10.1016/0165-0114(93)90251-C)
- Jamaludin, R., & Osman, S. Z. M. (2014). The use of a flipped classroom to enhance engagement and promote active learning. *Journal of Education and Practice*, 5(2), 124–131. <https://iiste.org/Journals/index.php/JEP/article/view/10648>
- Juhary, J., & Amir, A. F. (2018). Flipped classroom at the Defence University. *4th International Conference on Higher Education Advances (HEAD'18)*, 827–835. <https://doi.org/10.4995/HEAD18.2018.8093>
- Kafi, Z., & Motallebzadeh, K. (2014). A flipped classroom: Project-based instruction and 21st century skills. *International Journal of Language Learning and Applied Linguistics World (IJLLALW)*, 6(4), 35–46.
- Kenna, D. C. (2014). *A study of the effect the flipped classroom model on student self-efficacy*. Master thesis, North Dakota State University, Fargo, North Dakota.
- Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education*, 31(1), 30–43. <https://doi.org/10.1080/00220480009596759>
- McLaughlin, J. E., Griffin, L. M., Esserman, D. A., Davidson, C. A., Glatt, D. M., Roth, M. T., Gharkholonarehe, N., & Mumper, R. J. (2013). Pharmacy student engagement, performance, and perception in a flipped satellite classroom. *American Journal of Pharmaceutical Education*, 77(9), 196. <https://doi.org/10.5688/ajpe779196>
- Ministry of Education Malaysia. (2015). *Malaysia education blueprint 2015-2025 (higher education)*. Ministry of Education Malaysia. <https://www.um.edu.my/docs/um-magazine/4-executive-summary-pppm-2015-2025.pdf>
- Nederveld, A., & Berge, Z. L. (2015). Flipped learning in the workplace. *Journal of Workplace Learning*, 27(2), 162–172. <https://doi.org/10.1108/JWL-06-2014-0044>
- O'Flaherty, J., & Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. *The Internet and Higher Education*, 25, 85–95. <https://doi.org/10.1016/j.iheduc.2015.02.002>
- Osman, S. Z. M., Jamaludin, R., & Mokhtar, N. E. (2014). Flipped classroom and traditional classroom: Lecturer and student perceptions between two learning cultures, a case study at Malaysian polytechnic. *International Education Research*, 2(4), 16–25. <https://doi.org/10.12735/ier.v2i4p16>
- Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). SAGE Publication.

- Sanchez-Lezama, A. P., Cavazos-Arroyo, J., & Albavera-Hernandez, C. (2014). Applying the Fuzzy Delphi Method for determining socio-ecological factors that influence adherence to mammography screening in rural areas of Mexico. *Cadernos de Saúde Pública*, 30(2), 245–258. <https://doi.org/10.1590/0102-311X00025113>
- Tongco, M. D. C. (2007). Purposive sampling as a tool for informant selection. *Ethnobotany Research and Applications*, 5, 147–158. <https://ethnobotanyjournal.org/index.php/era/article/view/126>
- Tsay, C. H.-H., Kofinas, A., & Luo, J. (2018). Enhancing student learning experience with technology-mediated gamification: An empirical study. *Computers & Education*, 121, 1–17. <https://doi.org/10.1016/j.compedu.2018.01.009>
- Wiginton, B. L. (2013). *Flipped instruction: An investigation into the effect of learning environment on student self-efficacy, learning style, and academic achievement in an Algebra I classroom*. Doctoral thesis, The University of Alabama, Tuscaloosa, AL.
- Yemma, D. M. (2015). *Impacting learning for 21st century students: A phenomenological study of higher education faculty utilizing a flipped learning approach*. Doctoral thesis, Robert Morris University, Moon Twp, PA.

Developing self and peer assessment to improve student's appreciative critical ability in learning drama appreciation

Khafidatur Rohmah*; Endah Tri Priyatni; Heri Suwignyo

Universitas Negeri Malang

Jl. Semarang No. 5, Sumbersari, Lowokwaru, Kota Malang 65145, Jawa Timur

*Corresponding Author. E-mail: khafidatur.rohmah.1702118@students.um.ac.id

ARTICLE INFO

Article History

Submitted:

1 December 2020

Revised:

23 June 2021

Accepted:

23 June 2021

Keywords

self-assessment; peer assessment; appreciative critical ability; drama appreciation learning

Scan Me:



ABSTRACT

This study aims to develop self and peer assessment instruments to improve students' critical and appreciative abilities in learning drama appreciation. This research was conducted by following the ADDIE model development steps: needs analysis, product design, product development, product implementation, and product evaluation. There are two types of data in this study: qualitative data and quantitative data. Qualitative data are in the form of suggestions and comments from assessment experts, literature experts, and drama appreciation learning experts, as well as students, while quantitative data are in the form of scores obtained from assessment experts, literature experts, and drama appreciation learning experts, as well as students. Both data were obtained through questionnaire guidelines. The data obtained were then analyzed. Qualitative data were analyzed using descriptive analysis techniques, while quantitative data using quantitative descriptive analysis techniques. The analysis technique used shows that the product developed can increase students' critical appreciative abilities by getting an average percentage of 8.3% for the display aspect, 9.7% for the product content aspect, and 90.4% for the language aspect. The three averages were obtained from assessment experts, literary experts, and drama appreciation learning experts. When testing the product, students got an average score of 82.2% on the aspect of students' impressions of the use of self and peer assessment in increasing students' critical appreciation skills in learning drama appreciation and 80.4% on the aspect of practicality and ease of self and peer assessment for improving students' appreciative critical abilities in drama appreciation learning.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



How to cite:

Rohmah, K., Priyatni, E., & Suwignyo, H. (2021). Developing self and peer assessment to improve student's appreciative critical ability in learning drama appreciation. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 49-62. doi:<https://doi.org/10.21831/pep.v25i1.36221>

INTRODUCTION

Critical and appreciative abilities are several abilities that students must improve because with critical abilities, students can observe various problems that occur in daily life (Novtiar & Aripin, 2017, p. 120). In addition, with critical abilities, students' academic abilities will be formed (Pamularsih, 2019). Students will obtain these abilities because the ability to think critically is self-regulation in deciding something that results in interpretation, analysis, evaluation, and inference, as well as exposure using evidence, concept, methodology, criteria, or contextual considerations on which to base the decisions (Facione, 2011). Another opinion says that critical skills include analyzing activities, submitting arguments, providing clarification, evidence, reasons, implications of opinion, and generalizing conclusions based on facts (Ariyatun & Octavianelis, 2020, p. 35).

Appreciation is a characteristic of appreciation, which in the dictionary means appraisal or appreciation for something. The two abilities are related. Someone who will give an assessment must be based on appropriate and reasonable arguments. These arguments are obtained

from critical abilities. In the world of literature, the appreciative ability is an activity in the form of recognizing literary works through feelings or inner sensitivity and understanding and acknowledging the values expressed by the author (Ismawati, 2017).

The level of appreciative critical ability of each individual is different. This can happen due to environmental and age factors. Likewise, junior high school students are likely to have different critical thinking abilities from high school students and students in college. Syahbana (2012, p. 45) says that junior high school students (aged 12-15 years) have not fully thought abstractly. However, they have started to be able to apply thinking patterns that can lead them to understand and solve problems, which is a form of the critical nature of junior high school children.

Seeing that critical and appreciative abilities benefit students, in learning activities, teachers should use models, strategies, learning instruments, or so on that can help students improve these two abilities. Juhji and Suardi (2018, p. 16) said that the task of teachers in the era of globalization must be able to develop students' critical thinking. In the dictionary, intelligence is an experience that one has is ready to be used when faced with new facts or conditions. That way, it is hoped that students can adjust to the globalization era.

There are several ways that teachers can improve these two abilities to students, one of which is by using self-assessment instruments and peer assessment in learning drama appreciation. Both assessment and learning are used to improve students' critical appreciation skills based on the characteristics of appreciative critical abilities. Yunita et al. (2018) said there are five keys related to critical thinking, namely practical, reflective, reasonable, trust and action, while according to Harsiati (2013, p. 128), in a nutshell, receptive appreciation activities are the ability to recognize, understand, analyze, compare, generalize, reflect, and assess the form and content of literary works.

In general, the assessment is carried out by the teacher, and without realizing it, the teacher hones his appreciative critical ability so that the increase in these two traits is dominant to the teacher, not to the student, even though students need these two abilities to support their learning even in everyday life. Therefore, self and peer assessment are used to increase students' appreciative critical abilities. Self and peer assessments are used to improve students' appreciative critical abilities in drama appreciation learning. Both assessments are carried out by students themselves in assessing the learning process, especially in assessing the achievement of drama appreciation learning indicators. Learning indicators in this study were developed based on basic competence 4.15 (interpreting traditional and modern dramas that are read and watched or heard). The indicators developed are learning objectives that will be achieved by students, which implicitly have appreciation activities. Table 1 describes learning indicators based on basic competencies 4.15.

Table 1. Explanation of Basic Competencies and Learning Indicators for Drama Main Materials to Improve Students' Appreciative Critical Thinking Skills

Basic Competencies		Indicator	
4.15	Interpreting the drama (traditional and modern)	4.15.1	Describe the relationship or suitability between the elements of drama that is read and watched or heard.
	that is read and watched or heard	4.15.2	Reflecting the contents of a drama story with real life.
		4.15.3	Responding to the use of language as a medium of expression.

Drama learning is used to increase students' appreciative critical abilities because appreciation is individual. Students from one another are likely to have different appreciating points of view. They have the right to express arguments that they feel are correct. Ghufroni and Dewi (2019) said drama is a portrait of joy and sorrow, bittersweetness, and black and white of human life. Students are familiar with this and are part of the portrait of life; therefore, various types of drama literature were chosen to improve students' appreciative critical abilities.

In this development research, there are a number of things that must be known first: self-assessment and its techniques and peer assessment and its techniques. These two things need to be known to make it easier to understand the concept of this development research.

First, self-assessment is a process assessment carried out by students to discover their specific weaknesses and abilities. According to Basuki and Hariyanto (2014, p. 70), self-assessment is a process that describes how students obtain information and reflect on their own learning. This self-assessment is also used to determine personal progress in knowledge, skills, learning processes, and attitudes. This assessment will guide students towards better awareness and understanding of themselves as learners. Self-assessment has several techniques that can be used to improve appreciative critical abilities, namely (1) self-reflection journals, (2) editor's checklists, and (3) learning journals. The following is an explanation of the three techniques.

The self-reflection journal is an assessment technique used by students to convey their learning experiences in achieving learning indicators. Moon (2013) said that a self-reflection journal leads to experiences. In this case, what is meant is an experience in the form of steps or strategies carried out and experiences of difficulties students face during learning in achieving indicators. There are two types of self-reflection journals developed in this study: self-reflection journals in the form of statements of steps students use in conveying their learning strategies in achieving indicators and self-reflection journals in the form of statements of difficulties experienced by students in achieving indicators.

The editor's checklist is an assessment technique used by students in classifying themselves into the categories of students who are successful or not successful in learning. This editor's checklist contains questions regarding the achievement of indicators with sentences that use the word 'I'. Knowing the classifications of students categorized as successful in doing the assignment provides constructive comments to students who have not been successful in doing the assignment. Wragg (2001) says that the main use of checklists is to stimulate active learning. In this technique, there are three adjectives that describe students' understanding: (1) clear which means I understand as a whole, (2) buggy (blur/foggy) which means that I understand most of the material, but some things are still unclear, and (3) muddy (dusty/dark) that means I do not understand at all.

Self-Reflection Journal (Step Statement)
Basic Competence: interpreting drama (traditional and modern) which is read and watched or heard. (filled in by the teacher)
Indicator: describes whether or not the theme is related to the setting in the drama. (filled in by the teacher)
Name:
Day and date:
<p>The step I take in explaining whether or not a theme is related to the setting of a place is to first find out the drama theme and then identify the setting of the place and the nature of the place. The theme in the drama Sarjana Kambing is the aspirations of the village people (an explanation of how to determine the theme can be seen in the column marked with the lights below this column) and the setting is Irul's simple and clean house (terrace, dining room, living room, and kitchen), goat shed, vast rice fields, and solid Teguh' house.</p> <p>Based on my steps, the theme elements with the place setting in the drama Sarjana Kambing have a relationship or correspondence, with the reason that all the places in the drama are settings that are rarely found in urban areas. Even if there is a setting in this place, it is not like the nature of the place in the village, such as the background of the rice fields which are very broad, the house is simple and clean, and again the goat cage in the city is rarely found, so the setting depicted in the drama is appropriate. with the theme of the ideals of the villagers.</p>

Figure 1. Example of Self-assessment with Self-reflection Journal Techniques

The learning journal is an assessment used at the end of the lesson to find out the learning process that has occurred from the student's point of view. Yusuf (2017) argues that notes in learning journals can be in the form of personal observations, feelings, attitudes, perceptions, impressions, and opinions in response to readings, events, and experiences. Figure 1 shows an example of a self-assessment using a self-reflection journal technique (in the form of a step statement).

Second, peer assessment is a process assessment technique that can also improve students' appreciative critical abilities by involving fellow students to assess their respective work. Rochmiyati (2013) states that peer assessment as an alternative assessment gives students freedom in expressing opinions. This opinion is in the form of constructive comments as an improvement in the work of other students who have not reached the learning indicators. Peer assessment developed in this study is not in the form of numbers, but rather information that students can use to find out and improve their abilities in a certain subject matter. Liu and Carless (2006) say that in peer assessment, students can send more feedback and more quickly than teachers who provide comments. This is due to the teacher's limitations in providing sufficient feedback for the number of students so feedback from friends can be a central part of the learning process. Slamet (2020, p. 41) says that feedback is information communicated to students to modify thoughts or behavior to improve the quality of learning outcomes.

Peer assessment has several techniques that can be used to improve students' appreciative critical abilities: two stars and a wish and warm and cold feedback. Two stars and a wish is one of the process assessment techniques by students to assess their friends' work results by giving two stars for indicators that appear and one sign of hope for indicators that have not appeared. Wiliam (2011) states that two stars and a wish are a structured assessment technique for students to provide feedback on their friends' work results. In giving an assessment, students should not only give praise but also praise it accompanied by mentioning what form the praise is taking. This needs to be considered so each student can get constructive feedback.



Two Stars and a Wish	
Basic Competence: interpreting drama (traditional and modern) which is read and watched or heard. (filled in by the teacher)	
Indicator: explains whether there is a relationship or compatibility between the theme and the place setting in the drama. (filled in by the teacher)	
Name of student:	
Appraiser name:	
	You did a good job, because you concluded that there was a relationship or the appropriateness of the theme with the setting, but unfortunately you did not give vague and detailed reasons.
	I think you need to explain the theme in a concrete way and you can also add the nature of the locations that you mention to strengthen your argument, because the locations you mention also exist in urban areas, but they are different. In the city there may still be rice fields, but not as large as in villages and simple neat houses.

Figure 2. Example of Self-assessment Using the Two Stars and a Wish Technique

Warm and cold feedback is an assessment technique carried out by two students, and each student shares warm and cold feedback on the results of their work. Clark and Duggins (2016) says that warm feedback can include comments about assessments that meet learning objectives. In contrast, cold feedback can include comments on possible non-achievement of learning objectives, gaps, or problems. The way it works is that students in the clear category

assess the results of their work in the dark/dark category for constructive comments. In this case, students can offer ideas or suggestions to strengthen the assessment they provide. Figure 2 shows an example of one of the peer assessments using the two stars and a wish technique.

In implementing self and peer assessment, the teacher must arrange learning steps to realize all the techniques used. The learning steps that can be used in implementing self and peer assessment in learning are as follows. First, the teacher must convey to students when using self and peer assessment in learning so students know the steps or activities carried out during the learning process according to the two assessments characteristics. Second, the teacher provides illustrations, techniques and examples of self and peer assessment in drama appreciation learning (supervised assignments). Third, teacher presents the drama script as an appreciation material. Fourth, students are given the task of reading a drama script with a time limit. Periodically, students are asked to record reading results (results of appreciation) and report them orally. Students who finish appreciating more quickly are asked to share their experiences or appreciation strategies in front of their peers so they can learn from them (self-reflection journal in the form of step statements). Fifth, students who still face difficulties in the process are asked to share their experiences and give feedback (self-reflection journal in the form of difficulty statements). Sixth, students commented on positive aspects of their friends' work in the form of suggestions or strategies for improving their friends (peer assessment). Seventh, students provide constructive comments and feedback on aspects of their friends' work in accordance with the directions and rubrics given by the teacher (peer assessment). During the learning process, the teacher monitors and facilitates student activities.

RESEARCH METHOD

Research Design

The design of this study uses the ADDIE development model with five steps, namely analysis, design, development, implementation, and evaluation (Spector et al., 2014). The five steps can be seen in Figure 3.

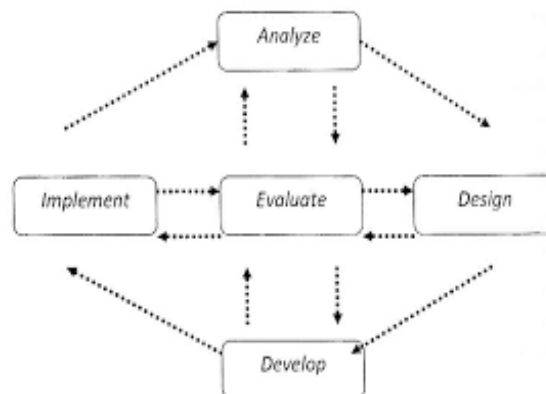


Figure 3. ADDIE Development Model

Development Procedure

At the analysis stage, preliminary research was carried out related to the dominant assessment conducted by the teacher. Besides, classroom observations were conducted to see whether assessment as learning is applied in learning, how it is applied, assessing various assessments used in schools through interview techniques.

The design stage was to design a product tailored to the product user (student), learning objectives, and the instrument's characteristics to be developed. What was done at this stage was making the steps for implementing assessment as learning in drama appreciation learning.

At the development stage, there are four actions carried out. They are (1) preparation of test grids, (2) writing test items, (3) writing instructions and examples of test work, and (4) writing answer keys or scoring signs.

The implementation stage was to develop products in the form of self-assessment and peer-assessment instruments. The two instruments development is adjusted to the respective indicators in the basic competencies used. For self-assessment, there are three techniques developed: self-reflection journals, study journals, and editor's checklists. Two techniques have been developed for peer assessment: two stars and a wish and warm and cold feedback.

The evaluation stage was to assess the products developed based on the data obtained from questionnaires in the form of student and expert responses. This stage was used as an improvement so that the product that was developed is in accordance with the objectives.

The appearance aspect that needs to be improved is the cover image used and the book margins. The suggestions are obtained from learning experts. The content aspect that needs to be improved is the simplification of each step in compiling the instrument developed. These suggestions were from students, while the assessment experts suggested improving the content aspect by reviewing the steps developed so they can be used in other drama scripts.

Data and Data Sources

There are two data in this study: quantitative and qualitative data. Quantitative data in the form of scores from the expert team's validity before the product were implemented and obtained from the results of field trials (students) after using the product. Meanwhile, qualitative data were in the form of responses, suggestions, and criticism of the validity of the contents and constructs of the instruments developed from experts before implementation and from students after using the products developed. The sources of research data were categorized into two, namely experts (drama learning experts, assessment experts, and literature experts) and respondents (11 students from a *madrasah tsanawiyah* or junior high school).

Research Instruments

The instruments used in this development research were interview guidelines and a questionnaire. First, the interview guide was used during preliminary observations to interview practitioners (teachers). The interview guidelines were conducted openly, which only contained an outline of the interview topic. Second, the questionnaire guidelines were used to assess the results of products carried out by experts, practitioners, and respondents (students).

Data Analysis Techniques

Two data analysis techniques were used: qualitative descriptive and quantitative descriptive data analysis techniques. Qualitative descriptive analysis technique described data from the product trial results in the form of suggestions and comments from experts after validated and obtained from respondents (students) after using the products developed. These data were obtained from a questionnaire in the section giving criticism and suggestions for the product developed. Quantitative descriptive analysis technique was used to analyze quantitative data from the validation results of experts and field tests conducted by students after using the product.

Table 2. Table of Eligibility Criteria

Scale	Achievement Level	Qualification	Follow-up
5	86-100	Very worthy	Implementation
4	71-85	Worthy	Implementation
3	56-70	Pretty decent	Implementation
2	41-55	Not feasible	Needs revision
1	≤ 40	Very unworthy	Needs revision

Both data were in the form of scores from filling out the questionnaire. Quantitative data analysis was conducted by collecting and analyzing numerical data from a questionnaire. The numerical data were then processed in percentage form using the percentage formula. After the percentage was known, it was described using the product eligibility criteria to interpret the feasibility of the resulting product by changing the five-scale achievement levels as presented in Table 2.

FINDINGS AND DISCUSSION

Based on the research objectives, namely developing assessment as learning products to improve students' appreciative critical abilities and testing the product's effectiveness, the results of the research are explained as follows.

Product Specifications Developed

The product specifications that were developed are adjusted to the conditions of the research subject. The specifications of the product developed are outlined in Table 3.

Table 3. Developed Product Specifications

Part Name	Contents
Part I Introduction	<ol style="list-style-type: none"> Purpose Scope
Part II Introduction	<ol style="list-style-type: none"> Definition of drama appreciation learning. Definition of assessment as learning. The use of assesment in learning drama appreciation.
Part III Types of Assessment as Learning in Drama Appreciation Learning	<ol style="list-style-type: none"> Self assessment <ol style="list-style-type: none"> Jurnal reflection Editor's checklist Study journal Peer assessment <ol style="list-style-type: none"> Two stars and a wish Warm and cold feedback
Part IV Examples of Assessment as Learning in Drama Appreciation Learning	<ol style="list-style-type: none"> Example of a self-reflection journal Sample checklist editor Journal of learning (learning journal) Two stars and a wish Example of warm and cold feedback
Part V Implementation of Assessment as Learning in Drama Appreciation Learning	Explanation of the implementation of assessment as learning in drama appreciation learning, which contains assessment and appreciation activities based on formulated learning indicators
Part VI Closing	<ol style="list-style-type: none"> Conclusion Suggestion

Table 3 in part V contains appreciation and critical thinking activities. The appreciation activities are described based on the basic competencies and the learning indicators which are defined in this study. In contrast, critical thinking activities are included in the assessment activities. Each appreciation and critical thinking activity is elaborated as follows.

Appreciation Activities

Appreciation activities are based on the basic competencies of learning on drama subject matter. Table 4 can be used as a consideration to find out the basic competencies and also the indicators of drama learning. The first indicator (4.15.1) is then translated into two points: (a)

explaining whether there is a relationship or suitability of the theme with the setting in the drama, and (b) explaining whether there is a relationship or suitability of language (dialogue) with characterizations in drama.

Table 4. Explanation of Basic Competencies and Learning Indicators

Basic Competencies		Indicator
4.15 Interpreting the drama (traditional and modern) that is read and watched or heard.	4.15.1	Describe the relationship or suitability between the elements of the drama that is read and watched or heard.
	4.15.2	Reflecting the contents of a drama story with real life.
	4.15.3	Responding to the use of language as a medium of expression.

The three indicators are translated into several appreciation activities. For the first indicator, point one is translated into four appreciative activities: (1) reading a drama script, (2) identifying the elements of drama, (3) noting the intrinsic elements, and (4) explaining the existence of a relationship or theme to the setting. For the first indicator, the second point is translated into four appreciative activities: (1) reading the drama script, (2) identifying each character and characterization as well as the language (dialogue) in the drama, (3) recording the results of character identification and characterization and the language used, and (4) explaining the relationship or suitability of language (dialogue) with characterizations. The second indicator is translated into four appreciative activities: (1) reading the drama script, (2) identifying each character and characterization and dialogue (language) in the drama, (3) recording the results of character identification and characterization along with the language used in drama, and (4) explaining the relationship or suitability of language (dialogue) with characterizations. The third indicator is translated into four appreciative activities: (1) reading the drama scripts, (2) identifying the language in drama, (3) recording the results of language identification in drama, and (4) responding to the use of language as a medium of expression.

The aforementioned appreciating activities are at the level of being fond of and reacting. Waluyo (2002) says there are four levels of appreciation, namely (1) the level of liking, (2) the level of enjoyment, (3) the level of reaction, and (4) the level of production. The activities of reading drama scripts, identifying the intrinsic elements of drama, and discussing the content or intrinsic elements of drama, then the activities of appreciating those described are categorized at the level of liking and reacting. It is in accordance with Amri and Damaianti (2017, p. 190) that appreciating activities are categorized at the level of liking if students read drama script independently or in groups with teacher directions, while categorized at the level of reaction if students can discuss well and issue an opinion about the content of the drama script.

Critical Thinking Activities

Critical thinking activities are contained in assessment activities described based on the basic competencies of learning on the subject matter of drama as in Table 4. The three indicators listed in Table 4 are translated into four assessment activities for each indicator. The assessment activities are in the form of (1) conducting an assessment using the self-reflection journal technique, (2) conducting an assessment using the checklist editor technique, (3) conducting an assessment using the two stars and a wish technique, and (4) conducting an assessment using the technique of warm and cold feedback.

Based on the characteristics of the four techniques, students are required to think critically when assessing the results of their work or their friends' work because they must provide accurate and easy-to-understand explanations for their assessment results so their friends can accept the explanation. Nuryanti et al. (2018, p. 155) state that critical thinking skills include basic clarification skills, basic decision making, concluding, providing a further explanation, estimation, and integration, as well as additional abilities. Based on this explanation, the assess-

ment activities in this research product include the ability to make decisions, conclusions, provide further explanations, and additional capabilities in the form of giving practical advice. These abilities are realized based on the characteristics of the four assessments used in this study. These abilities will appear in each indicator; if there are five indicators (including the elaboration of the first point indicator), then students will do critical thinking activities five times. By doing the assessment activity five times, it is expected that the students' thinking ability will increase. This is in accordance with the explanation of [Suratno and Kurniati \(2017, p. 2\)](#) that critical thinking skills can be sharpened by accustoming students to be actively involved in solving problems that require critical thinking skills.

Developed Product Effectiveness

The product's effectiveness is obtained from the validation process of the drama appreciation learning experts and the respondents (students) after the product trial. The data from the product trial results are in the form of numerical and verbal data. Numerical data is calculated from the score obtained through a questionnaire, while the verbal data is in the form of criticism and suggestions obtained from experts on the product feasibility test results. Figure 4 presents the instrument used to determine the effectiveness of the product developed.

No	Aspect	Score				
		1	2	3	4	5
1.	View					
	a. The image used on the cover matches the book title.					
	b. The colors used correspond to the user object.					
	c. The symbols used help in finding information.					
2.	Material/Content					
	a. Scope of assessment theory.					
	b. The compatibility of the drama script with the user.					
	c. The scope of drama appreciation theory.					
	d. Suitability of indicators with Basic Competencies.					
	e. Clarity of examples of each assessment.					
	f. Sample coverage included in each assessment technique.					
	g. Clarity of the steps for implementing self and peer assessment in drama appreciation learning.					
	h. The activities described in the implementation steps contain critical thinking activities.					
	i. The activities described in the implementation steps contain appreciative activities.					
	j. The steps in each technique require students to think critically.					
k. Appreciative activities described in the product can improve students' appreciative skills.						
3.	Language					
	a. The language used is in accordance with the target user.					
	b. The language used is effective and does not cause multiple meanings.					
	c. The use of words and punctuation is in accordance with the rules in PUEBI.					
Total score =						
Validator Comments and Suggestions						
<div style="border: 1px solid black; padding: 5px;"> </div>						

Figure 4. Data Collection Instruments by Experts

No.	Criteria	Score				
		5	4	3	2	1
1	Students' impressions of the use of self and peer assessment in drama learning.					
	a. The application of self and peer assessment will make you more motivated in taking drama lessons.					
	b. I find it easier to appreciate using self and peer assessment.					
	c. I feel more critical in appreciating literature, especially drama, using self and peer assessment					
	d. I feel more confident speaking in public after using self and peer assessment in drama appreciation learning.					
	e. Self and peer assessment helped me reveal the difficulties you experienced during the learning process.					
	f. Self dan peer assessment helped me in improving my appreciative skill.					
2	Practicality and ease of self and peer assessment techniques used in drama appreciation learning.					
	a. Practicality and ease of self-reflection journal techniques (in the form of steps) in learning drama appreciation.					
	b. Practicality and ease of self-reflection journal techniques (statement of difficulties) in learning drama appreciation.					
	c. Practicality and ease of editor's checklist technique in learning drama appreciation.					
	d. The practicality and ease of the two stars and a wish in learning drama appreciation.					
	e. Practicality and ease of warm and cold feedback techniques in learning drama appreciation.					
Total score=						
Comments and Suggestions from Students						
<div style="border: 1px solid black; padding: 5px;"> </div>						

Figure 5. Data Collection Instruments by Students

Based on the instrument to determine the effectiveness of the developed product from Figure 4 and Figure 5, the score obtained can be found. Each of it is elaborated as follows.

Acquiring Data from Assessment Experts and Drama Learning Experts

There are three aspects assessed at the validation stage of the assessment expert and drama learning expert: (1) the aspect of book appearance, (2) the material/book content aspect, and (3) the language aspect. The following is an explanation of each of these aspects.

Aspects of Book Display

Three eligibility criteria are assessed in this aspect: (1) the suitability of the image on the cover with the book title, (2) the suitability of the color with the user object, and (3) the suitability of the symbols used in the book. The results of the due diligence are shown in Table 5.

Table 5. Feasibility Test Results from Display Aspects (Numerical Data)

Test Subject	Percentage	Qualification	Follow-up
Assessment expert	60	Pretty decent	Implementation
Drama learning expert	80	Worthy	Implementation
Literary expert	73	Worthy	Implementation
Mean	83	Worthy	Implementation

Aspects of Book Content

Nine eligibility criteria are assessed in this aspect: (1) the scope of the assessment theory, (2) the suitability of the drama script with the user, (3) the coverage of drama appreciation theory, (4) the suitability of indicators with basic competencies, (5) the clarity of examples - each assessment, (6) the scope of examples included in each assessment technique, (7) clarity of the steps for implementing self and peer assessment in drama appreciation learning, (8) the activities described in the implementation steps containing critical thinking activities, and (9) the activities described in the implementation steps containing appreciative activities. The results of the due diligence are presented in Table 6.

Table 6. The Results of the Feasibility Test for the Aspect of Book Contents (Numerical Data)

Test Subject	Percentage	Qualification	Follow-up
Assessment expert	62.23	Pretty decent	Implementation
Drama learning expert	84.4	Worthy	Implementation
Literary expert	100	Very worthy	Implementation
Mean	92.7	Very worthy	Implementation

Language Aspects

Three eligibility criteria are assessed in this aspect: (1) the suitability of the language with target user, (2) the language used is effective and does not cause double meanings, and (3) the suitability of the use of words and punctuation in PUEBI. The results of the due diligence are explained in Table 7.

Table 7. Test Results of the Language Aspect (Numerical Data)

Test Subject	Percentage	Qualification	Follow-up
Assessment expert	60	Pretty decent	Implementation
Drama learning expert	80	Worthy	Implementation
Literary expert	93.3	Very worthy	
Mean	90	Very worthy	Implementation

Field Trial Results (Students)

Two aspects were assessed in the field test: (1) students' impressions regarding the use of assessment as learning to improve their appreciative critical abilities in learning drama appreciation and practicality, (2) the ease of self- and peer-assessment techniques used to improve students' appreciative critical abilities in drama appreciation learning. In the field product trial stage, students also provide suggestions for product improvements developed based on their experiences (Table 8). The explanation on each of these aspects is as follows.

Table 8. Product Improvement Suggestions (Verbal Data)

Test Subject	Product Improvement Suggestions
Assessment expert	<ul style="list-style-type: none"> • The AaL construct needs to be reviewed, so it includes both the process and results. • The giving of examples is too specific, so it cannot be used for assessment of other texts with the same style.
Drama learning expert	Improve product quality in terms of appearance (images and margins).

Aspects of Student Impression of Self- and Peer-Assessment in Increasing Students' Critical Appreciative Ability

Five eligibility criteria are assessed at this stage, namely (1) self- and peer-assessment motivate students to take part in drama appreciation learning, (2) self- and peer-assessment

eases students appreciate drama by using self and peer assessment, (3) self- and peer-assessment increases critical abilities students appreciate drama, (4) self- and peer-assessment increases students' sense of confidence when speaking in public, and (5) self- and peer-assessment helps students express difficulties experienced during learning. The test results are described in Table 9.

Table 9. Feasibility Test Results in the Form of Students' Impressions of Self- and Peer-Assessment in Improving Students' Appreciative Critical Ability

Test Subject	Percentage	Qualification	Follow-up
Students	82	Worthy	Implementation
Mean	82	Worthy	Implementation

Practical Aspects and Ease of Self- and Peer-Assessment in Increasing Students' Critical Appreciative Ability

Five eligibility criteria were assessed at this stage, namely (1) the practicality and ease of the self-reflection journal technique (in the form of steps) in learning drama appreciation, (2) the practicality and convenience of the self-reflection journal technique (statement of difficulty) in learning drama appreciation, (3) the practicality and convenience of the editor's checklist technique in learning drama appreciation, (4) the practicality and convenience of the two stars and a wish technique in learning drama appreciation, and (5) the practicality and convenience of the warm and cold feedback technique in learning drama appreciation. The results of the feasibility test are explained in Table 10, and the explanation on the product improvement suggestions is shown in Table 11.

Table 10. Results of Practicality and Ease of Self-Assessment and Peer-Assessment in Improving Students' Appreciative Critical Ability

Test Subject	Percentage	Qualification	Follow-up
Students	80.4%	Worthy	Implementation
Mean	80.4%	Worthy	Implementation

Table 11. Explanation of Product Improvement Suggestions (Verbal Data)

Test Subject	Product Improvement Suggestions
Students	<ol style="list-style-type: none"> 1. Change the language according to the students. 2. The preparation of self-assessment techniques needs to be simplified for beginners. 3. Provide a self-assessment step related to identifying the intrinsic element in the drama. 4. The technique of self-reflection journals needs to be simplified. 5. Simplify the steps for implementing self and peer assessment in drama appreciation lessons.

Based on the explanation in the introduction and the results of the data obtained, self- and peer-assessment can be categorized as a student-centered learning process. This is because students are active during learning. It is said to be active because students themselves assess the results of their work, and students provide constructive feedback using the developed assessment techniques. In addition, students can also try and find solutions to the problems they face in achieving learning indicators. Cahyadi et al. (2019, p. 207), in their research, said that the method used in learning which can make students active in learning both in attitude, knowledge, and skills, is student-centered learning. Learning that keeps students active is one of the characteristics of innovative learning strategies. Supriyadi (2017, p. 209) states that with innovative learning, students can learn actively, creatively, and innovatively. The discussion also explained that the emergence of active attitudes and creative and innovative abilities is an effort to create student-centered learning so that students have competence.

CONCLUSION

Self- and peer-assessment is one type of assessment carried out by students during learning to identify students' strengths and weaknesses in achieving learning goals. These assessments have their own techniques that can improve students' critical and appreciative abilities through learning drama. For self-assessment, there are three techniques used, namely (1) self-reflection journal technique, (2) checklist editor technique, and (3) learning journal technique. For peer assessment, two techniques are used, namely two stars and a wish and warm and cold feedback. Based on the data obtained, self- and peer-assessment are feasible to increase students' critical-appreciative abilities in learning drama appreciation.

ACKNOWLEDGMENT

The authors thank the Directorate of Research and Community Service, the Director-General of Research and Development Strengthening, and the Ministry of Education and Culture for awarding a research grant in 2020. We hope that the research can contribute to the world of education in Indonesia.

REFERENCES

- Amri, U., & Damaianti, V. S. (2017). Pengaruh penggunaan teknik bermain drama melalui teater tradisional Randai berbasis kepercayaan diri terhadap kemampuan apresiasi drama. *EduHumaniora: Jurnal Pendidikan Dasar Kampus Cibiru*, 8(2), 186–197. <https://doi.org/10.17509/ch.v8i2.5141>
- Ariyatun, A., & Octavianelis, D. F. (2020). Pengaruh model problem based learning terintegrasi stem terhadap kemampuan berpikir kritis siswa. *JEC: Journal of Educational Chemistry*, 2(1), 33–39. <https://doi.org/10.21580/jec.2020.2.1.5434>
- Basuki, I., & Hariyanto, H. (2014). *Asesmen pembelajaran*. Rosdakarya.
- Cahyadi, E., Dwikurnaningsih, Y., & Hidayati, N. (2019). Peningkatan hasil belajar tematik terpadu melalui model project based learning pada siswa sekolah dasar. *Jartika: Jurnal Riset Teknologi Dan Inovasi Pendidikan*, 2(1), 205–218.
- Clark, S., & Duggins, A. S. (2016). *Using quality feedback to guide professional learning: A framework for instructional leaders*. Corwin Press.
- Facione, P. A. (2011). *Critical thinking: What it is and why it counts*. Measured Reasons and the California Academic Press.
- Ghufroni, G., & Dewi, M. R. (2019). Pengembangan bahan ajar bermain drama dengan model pembelajaran SAVI pada siswa SMA. *Jurnal Ilmiah SEMANTIKA*, 1(1), 31–46. <http://jurnal.umus.ac.id/index.php/semantika/article/view/80>
- Harsiati, T. (2013). *Asesmen pembelajaran Bahasa Indonesia*. UM Press.
- Ismawati, E. (2017). Mantra Bumi karya Aprinus Salam sebagai bahan ajar apresiasi sastra. *PIBSI XXXIX*, 671–681. <http://eprints.undip.ac.id/58822/>
- Juhji, J., & Suardi, A. (2018). Profesi guru dalam mengembangkan kemampuan berpikir kritis peserta didik di era globalisasi. *Geneologi PAI: Jurnal Pendidikan Agama Islam*, 5(1), 16–24. <http://jurnal.uinbanten.ac.id/index.php/geneologi/article/view/1043>
- Liu, N.-F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290. <https://doi.org/10.1080/13562510600680582>

- Moon, J. A. (2013). *Reflection in learning and professional development: Theory and practice*. Routledge. <https://doi.org/10.4324/9780203822296>
- Novtiar, C., & Aripin, U. (2017). Meningkatkan kemampuan berpikir kritis matematis dan kepercayaan diri siswa SMP melalui pendekatan Open Ended. *PRISMA*, 6(2), 119–131. <https://doi.org/10.35194/jp.v6i2.122>
- Nuryanti, L., Zubaidah, S., & Diantoro, M. (2018). Analisis kemampuan berpikir kritis siswa SMP. *Jurnal Pendidikan: Teori, Penelitian, Dan Pengembangan*, 3(2), 155–158. <http://journal.um.ac.id/index.php/jptpp/article/view/10490>
- Pamularsih, P. (2019). Pengaruh penggunaan model pembelajaran Cooperative Reading and Composition (CIRC) dan kemampuan berpikir kritis terhadap kemampuan apresiasi cerpen siswa SDN Mrayan Kabupaten Ponorogo. *Linguista: Jurnal Ilmiah Bahasa, Sastra, Dan Pembelajarannya*, 2(2), 106–112. <https://doi.org/10.25273/linguista.v2i2.3699>
- Rochmiyati, R. (2013). Model peer assessment pada pembelajaran kolaboratif elaborasi IPS terpadu di sekolah menengah pertama. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 17(2), 333–346. <https://doi.org/10.21831/pep.v17i2.1704>
- Slamet, S. S. (2020). Hubungan strategi umpan balik (feedback), motivasi berprestasi dan hasil belajar dalam pembelajaran PPKn di SMK. *PINUS: Jurnal Penelitian Inovasi Pembelajaran*, 5(2), 39–56. <https://doi.org/10.29407/pn.v5i2.14539>
- Spector, J. M., Merrill, M. D., Elen, J., & Bishop, M. J. (Eds.). (2014). *Handbook of research on educational communications and technology*. Springer New York. <https://doi.org/10.1007/978-1-4614-3185-5>
- Supriyadi, S. (2017). Pembelajaran bahasa dan sastra Indonesia yang inovatif. *Prosiding Seminar Nasional #3: Bahasa Dan Sastra Indonesia Dalam Konteks Global*, 209–218. <http://jurnal.unej.ac.id/index.php/fkip-eipro/article/view/4871>
- Suratno, S., & Kurniati, D. (2017). Implementasi model pembelajaran math-science berbasis performance assessment untuk meningkatkan kemampuan berpikir kritis siswa di daerah perkebunan kopi Jember. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 21(1), 1–10. <https://doi.org/10.21831/pep.v21i1.11799>
- Syahbana, A. (2012). Peningkatan kemampuan berpikir kritis matematis siswa SMP melalui pendekatan contextual teaching and learning. *Edumatica: Jurnal Pendidikan Matematika*, 2(1), 45–57. <https://online-journal.unja.ac.id/index.php/edumatica/article/view/604>
- Waluyo, H. J. (2002). *Drama: Teori dan pengajarannya* (A. Wulandari (Ed.)). Hanindita Graha Widia.
- Wiliam, D. (2011). *Embedded formative assessment - Practical strategies and tools for K-12 teachers*. Solution Tree Press.
- Wragg, E. C. (2001). *Assessment and learning in the secondary school*. Routledge.
- Yunita, N., Rosyana, T., & Hendriana, H. (2018). Analisis kemampuan berpikir kritis matematis berdasarkan motivasi belajar matematis siswa SMP. *JPMI (Jurnal Pembelajaran Matematika Inovatif)*, 1(3), 325–332. <https://doi.org/10.22460/jpmi.v1i3.p325-332>
- Yusuf, A. M. (2017). *Asesmen dan evaluasi pendidikan*. Kencana.

Evaluation of TOEFL preparation course program to improve students' test score

Mega Selvi Maharani*; Nur Hidayanto Pancoro Setyo Putro

Universitas Negeri Yogyakarta

Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia.

*Corresponding Author. E-mail: megaselvi6@gmail.com

ARTICLE INFO

Article History

Submitted:

12 March 2021

Revised:

10 May 2021

Accepted:

31 May 2021

Keywords

TOEFL preparation;
course program;
improving students' test
score

Scan Me:



How to cite:

Maharani, M., & Putro, N. (2021). Evaluation of TOEFL preparation course program to improve students' test score. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 63-76. doi:<https://doi.org/10.21831/pep.v25i1.39375>

ABSTRACT

This research was conducted to evaluate the management program in English village, maintain the implemented program, and increase program quality. The method used in this study was descriptive qualitative and quantitative research. The model evaluation used was CIRO (Context, Input, Reaction, and Outcome) by War, Bird, and Rackman. The subject of this research were 30 students, four tutors, and a program director. Interview, survey, and observation were used for collecting the data. The analyzed quantitative data used descriptive percentages, and qualitative data used condensation, data exposure, conclusion drawing, and verification. The results of this study show that (1) context in this study considered to participant needed which synchronized to the standard implementation of the TOEFL real test and national standard. (2) input has been provided and is well prepared, including program plan, tutor qualification, for admission of course participants, and facilities. (3) Reaction of the participants' satisfaction was in the satisfactory category, and the participant's assessment reaction of the TOEFL program implementation process was in the very good category. (4) Outcome obtained by course participants in the TOEFL scoring has not increased significantly; it has even decreased during the last test.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



INTRODUCTION

The development of education today has always been a top trending in every discussion; each sector generally contributes to educational progress. Efforts to improve education quality are made by many people using various strategies (Kurniawati, 2017). Educational quality in Indonesia is one of the objectives and integral pieces of developing human resources exhaustively (Mulyasa, 2005). Seeing the importance of improving the quality of education, it takes effort and cooperation that is actualized continuously by foundations or institutions of formal, informal, and non-formal education.

Non-formal education is an effort to improve the education provided by the community and given to the societies according to existing needs (Sudjana, 2004). Most of the non-formal education established in Indonesia is approximately 16,935, including courses and training. Most of them are English courses due to so many requests from the community (Direktorat Pembinaan Kursus dan Pelatihan, 2018). With many courses provided in Indonesia, the quality of English graduates should be better. On the contrary, Indonesia is the lowest rank of mastery of the English language in Asia. Dahuri (2019) states that Indonesia is one of the countries that are still included in the low English Proficiency category. The low level of English proficiency in Indonesia is caused by the poor environment of English (Maruf et al., 2020). It can be seen on the data of EF English Proficiency Index that Indonesian is included on low

rank which has 15th on Asian number and 453 test score (English First, 2020). One of the lousy environmental impacts is students' low motivation in learning English. Thus, in this case, the English village comes to build an environment that supports learning English by developing the communities and raising many institutions to improve students' English skills.

In the English village, many language communities have the same vision to learn English (Nurhayati et al., 2013). English village course institutions provide language services such as basic classes, intermediate classes, and advanced classes. The final program in the English class is the TOEFL preparation program. TOEFL preparation program is a learning program to improve student reading, structure, and listening skills in the academic English language (Sakurai, 2020). TOEFL is one of the world's English proficiency tests (Ismail & Othman, 2020; Syamsuddin & Min, 2014). Furthermore, according to Ali (2012), TOEFL is an accumulation of student learning outcomes and achievements in English. Thus, this clearly shows that the TOEFL test is a measure of students' successfully mastering English, so the TOEFL test preparation program is needed to improve the scores of participants in the test. According to Ma and Cheng (2016), taking a test preparation course is the most time-efficient method of preparing for the TOEFL test. TOEFL preparation has been held out in the Pare English village, but many problems are found in implementing TOEFL preparation courses.

Based on the pre-research in December 2019, the researchers found several problems in some courses in the Pare English village, one of which is the TOEFL program. The problem is that students' TOEFL scores had not reach the national standard yet. Students' TOEFL score in the final test was less than 450 score. However, the average Indonesian minimum standard score is 450 as the requirements of the college. In order to improve students' scores, the course plan must be well designed to help increase the scores of course participants (Pre-research data, 2019). Further, another study found that the problem that is often mistaken at holding in TOEFL courses is that the tutor spends much time providing strategies for answering questions to increase student scores, but they are ignoring participant comfort in the class so that participants feel that the TOEFL class is boring (Wang, 2019). The organizer must handle these problems so it can give an attractive impression to the students in the learning process without neglecting the actual course objectives. To overcome this case, it is necessary to have good management in a course institution. As such, for controlling the management program, an evaluation is a need when the program has been running. An evaluation is needed to see the effectiveness of a program since it is an activity to improve quality, performance, and productivity in implementing a program (Mardapi, 2017). Problems in a program can be overcome with an evaluation which is the basis for decision making (Adib et al., 2019). Therefore, the acceleration of improvement in improving the quality of course institutions must be carried out by evaluating the program being implemented in-depth.

Generally, evaluation is the comparison between the goal of the program and objectives that have been achieved in the program (Topno, 2012). Thus, the goal is the important component that should be appropriately planned. Different from the previous one, according to Sahayu and Friyanto (2019) believe that solving a problem that occurs in the program is not only enough by designing goals and needs analysis; it is also necessary to see the level of satisfaction of the course participants. Course participant satisfaction is the most important aspect that will influence the motivation and success rate of the program (Dewi & Kartowagiran, 2018). Correspondingly, Choudhury and Sharma (2019) state that in implementing an evaluation at the course and training institution, it is necessary to see the program's effectiveness and benefits for the company and participants. The benefits felt by course participants can be seen through the participants' goal achievement and the satisfaction of services provided by the course institution. Thus, to have good training and courses program, the goals of the program and the participants' marking about the program should be prepared. To broadly and deeply see an institution's management, such as goals, service satisfaction, and outcome of the program, the appropriate evaluation model that can be used is CIRO (Context, Input, Reaction,

and Outcome) evaluation model. The CIRO evaluation model of course and training emphasizes the problem domain and program performances as a step to design the objective, improve performance, and gain recognition (Sutton, 2006). Thus, the CIRO evaluation model is suitable to be used in this research, so it underlines the importance of the study on evaluating the management of TOEFL preparation course at one of the English village courses in East Java. The components of the evaluation model are context, input, reaction, and outcome of the program. Those components are chosen based on the importance of the institution's preparation and participants' response to the implementation of the TOEFL preparation program.

RESEARCH METHOD

This study used qualitative and quantitative research methods. These methods were chosen because this study used the CIRO evaluation model, which required detailed data and measurement percentages. A qualitative method was used to examine the context and input aspect, while a quantitative method was used to measure the percentage of the course participants' reactions and calculate the results of their achievements. This research was conducted from February to June 2020 in the English village of Pare, Kediri, East Java.

The subject of this study was the program director, four tutors, and 30 course participants at the famous course in Pare English village. The informants were selected using the purposive sampling technique, which allowed researchers to get detailed information. The data collecting technique in this research were observation, interviews, questionnaires, and documentation. In the qualitative research, the data collected from observation and interviews were validated by triangulation. In the quantitative research, the questionnaire was validated by content validity utilizing Gregori formula by the result of 0.96. The construct validity was analyzed using SPSS utilizing the *Kaiser Mayer Olkin Measure of Sampling Adequacy (KMO)* by the result of 0.97, followed by the reliability testing analyzed using SPSS employing Alpha Cronbach's formula by the result of 0.70. It proved that the instrument is proper to use in this research.

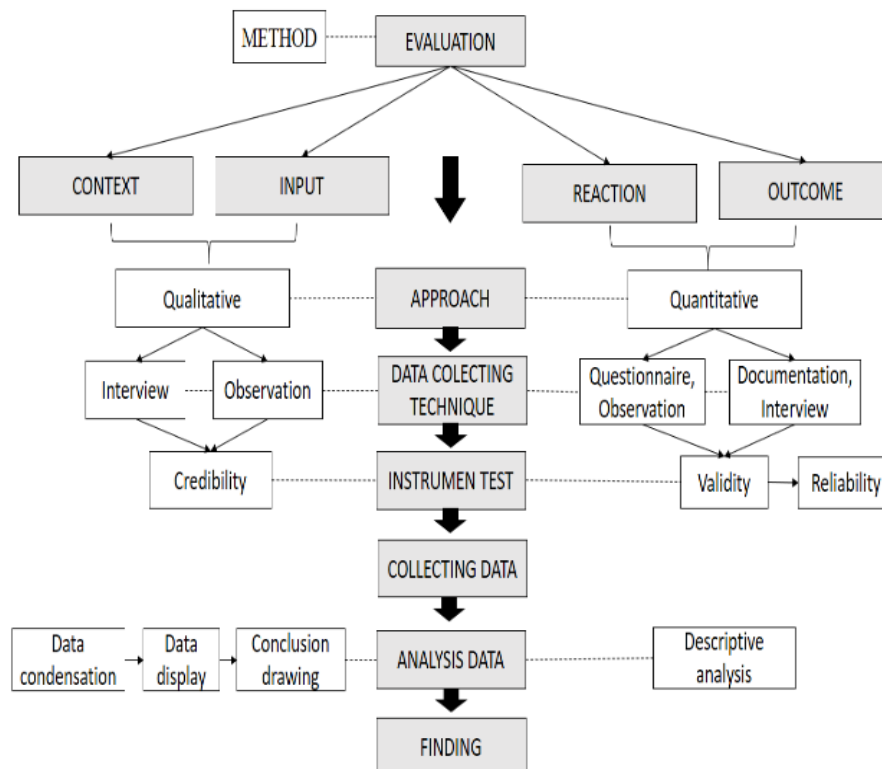


Figure 1. Method Schemes

Data analysis techniques were divided into each method used. Qualitative methods were analyzed through three stages, namely data condensation, data display, and drawing conclusions. In addition, the quantitative method uses descriptive analysis techniques based on normal distribution using Ms. Excel 2016 program to determine the percentage of responses and the learning outcomes of course participants. The details of the schemes used in the study are shown in Figure 1.

FINDINGS AND DISCUSSION

The evaluation model used in this study is the four aspects of CIRO, which have different data collection methods. Therefore, it is explained step by step based on these components, namely context, input, reaction, and outcome, elaborated as follows.

Context

The first stage of the evaluation is the evaluation of the context, which is divided into two aspects, namely the ultimate objective and the immediate objective in the management of the TOEFL preparation program. The program's ultimate objective is the final objectives designed by the course institute to achieve progress in the implementation of the program. The decision evaluation making in the aspect of context evaluation in the TOEFL preparation program is then compared to the standard procurement of courses that apply in Indonesia.

The research results by collecting data using interviews with the institution's director and confirmed through the tutor who taught at the institution have been confirmed verbally. The objective of the TOEFL preparation courses program has been designed without a written document. It can be interpreted that there is no official document regarding the purpose of establishing the TOEFL preparation program. It can be seen by the director's statement, "The program objectives were designed a long time before the program was implemented, but they were not written or recorded about the objective of the program." The director maintains that the objective was constructed by coordinators of the program. It is confirmed to the tutor that the program's objective is a collision between the program coordinator and the research and development department at the institution.

Based on the research interview, the ultimate objective of the TOEFL preparation program according to the tutor of the program is:

"The final goal of the program? Yes, the course has several goals, such as helping students to achieve the score they want, it is like to make students pass in real test because the final result is not here but the real test".

Additionally, the program director and student both argued, "The ultimate objective of the TOEFL preparation program is to pass the TOEFL real test to get a certificate which can be used for looking for job or college approval requirements". Based on the statement, the ultimate objective of the TOEFL preparation program is to help the participants pass the real test and get the score they expect.

The TOEFL minimum score of the course is 450, considering the average standard of Indonesian universities and job requirement. However, this is not an absolute value since many participants have different targets according to their needs. Another ultimate objective of the TOEFL preparation program is to help participants answer similar questions even though the questions are not the same. Based on an interview with a tutor, the TOEFL standard is as follows. "It is based on the agreement of the student's own needs". It means that participants and institutions are interrelated and support each other in increasing students' achievement.

TOEFL preparation program at these institutions is suitable to the [Regulation of the Minister of National Education No. 49 of 2007](#) concerning educational management standard in nonformal education units, in which such formulation of objectives must be planned before

the program is started. Discussions about program procurement objectives and program planning are better held regularly, once or twice a year (Gonçalves & Chauma, 2020). Such activities have been done to improve the quality of the program implementation. The final objective setting has been carried out to determine the goals that will be achieved after implementing the program, even though it is not written in an official document. Correspondingly, the lack of documents in determining the objectives is similar to the English language program, such as the AADU program, which does not have clear documents or program objectives; this has resulted in a lack of communication between stakeholders (Aktaş & Gündoğdu, 2020). The result of the research states that there is a good communication between the coordinator and program implementer about program planning, but the preparation of the documents is one of the crucial aspects in program planning, which aims to anticipate problems.

TOEFL planning program of English village course institution is adapted to the course participants needed as having a TOEFL score that can be used according to the personal target. It can be concluded that there is no score determination standard in improving learning because of the differences in the needs and basic abilities of the course participants. Similar to this study, Zhao (2020) states that attention to the needs of the course participant is the most important thing, which is useful for helping the course participants achieve the targeted competencies (Syakur et al., 2020). Further, Liu (2020) elaborates that omitting the needs of participants can affect the quality of learning and participants' satisfaction with the program. Emphasizing the course participants' needs can be considered the right step to be implemented to form the final objective so that program planning has been implemented properly. It has been implemented in the courses that have been researched.

The immediate objectives are skills that the courses provide to the participants to support the final objective. Based on the interview, the special skills students want to achieve in the TOEFL preparation course are structure skills. However, other interview results show that the participants are not all weak in this structure field. The program director states:

“The specific objective leads to the goal of how members can reach the TOEFL score according to the required standards, but the structure is more difficult from the others, so two meetings are given a day, and listening and reading are given one time in one day”.

As seen in the excerpt, structure skills have special attention due to the amount of material that must be discussed more because of the difficulties and demands of the participants. It was confirmed to participants of the TOEFL program who attended the courses. From the whole of the result, it can be concluded that the institution's immediate objectives are giving the skills to improve the participants' scores in TOEFL. The skills are given, including listening, reading, and structure. Based on the participants' needs, the structure class is held two classes per day while listening and reading are held one class for each per day. The meeting frequencies are four times of discussion or theory in one week, then one scoring a week, so that the total meetings in the course become 68 meetings in one-course period.

The interview results conclude that research and review have been carried out prior to the design of program objectives. The skills presented in the course are planned based on the needs of the community who would attend the course. This is parallel to the theory that paying attention to skills according to the needs of participants is important in improving the learning outcomes of the course participants (Silva & Tosqui-Lucks, 2020). As such, the steps that have been carried out by the institution are the right steps in increasing the achievement of the course participants' scores.

One of the ways is that the institution provides intensive classes that have four meetings in one day. It would give effective class, but intensive classes risk fatigue in learning course participants, so that appropriate strategies and arrangements are needed in the learning process. According to Wenjie (2020), organizing learning slots every day will be a challenge faced in order to avoid physical, mental fatigue that can interfere with the learning effectiveness. The

dividing of the learning schedule is something that becomes the attention of the program. In these courses, the timing of the TOEFL program has been declared in a good way, but it has to be improved in regulating class distances because of the long distances between each class so that it will give the teacher enough preparation time.

Inputs

Inputs are needed to support program implementation. Common inputs that are important to maintain are school resources, teacher quality, facilities, and learning outcomes (Britton & Vignoles, 2020). Input is an important factor in the management of a course institution. The results of the interview show that the preparation of the institution towards implementing learning activities is as follows.

First, creating a program plan: the program team prepared a program planning by referring to the students' needed and TOEFL real test. Program planning was written on the syllabus, which was outlined in the learning module. However, the teacher should shape the class conditions in the learning process. The syllabus was used to synchronize the methods and strategies the tutor would use in class. Several aspects were covered in the program plan at the institution, such as learning schedule, learning model, learning assessment, and learning regulation. The examples of data obtained through analysis and documents checking are as follows.

"Program planning refers to a learning method that is possible to use in the class that can be controlled by the teacher in the class. Then the strategies were prepared by the teacher before. All things have been exhaustive in the program or lesson planning. The mentor allows the participants to follow the syllabus by seeing the class condition. The tutors can improve the learning strategies based on the possibilities."

Based on the interview, it can be seen that the learning plan has been determined, but it has to follow and adapt to the actual conditions in the classroom. Furthermore, all tutors and program director have the same opinion about the schedule of the program that:

"For the admission program, those were 10 and 25 periods, so period 10 is the entry period approaching the 10th and 25th period which approaches the 25th, it is not always 10th, but sometimes we will start in the date that is approaching the 10th. For example, the 10th is on Saturday, so we will start on Monday, or if the 10th is on Tuesday, we start on Monday also".

Learning program, one of the aspects that are considered to start learning, provides the determination of the schedule and period parallel to the learning period agreed upon in the English village. The learning model, learning assessment, and learning regulation have been well prepared and are suitable for the procurement of the course programs in Indonesia. The interview and document about the aspects of the planning program can be summed up in Table 1.

Table 1. Contents of the Planning Program

Aspect	Planning
Learning schedule	Classes would start on the 10th and 25th of every month.
Learning model	The learning model has been based on the applicable operational standards but can be modified according to class conditions.
Learning assessment	The assessment was given to participants, and evaluation was given to the tutors. Participants took the test once a week on Friday, called Scoring. Evaluation for tutors is held every week and every month during the briefing process.
Learning regulations	There were two learning regulations given in the TOEFL program: rules for tutors and the course participants. The rules for participants were the rules for the course environment that must be obeyed by course participants. Meanwhile, the regulation for tutor emphasizes the activities, duties, and responsibilities of the tutor in the learning process.

The result of this study is similar to previous research, which states that learning planning must match the students' needs and the syllabus (Perez & Mardapi, 2015). To avoid misunderstanding and misuse of instructions, themes and lesson plans are required to be formed by the program organizer (Liu, 2020). It can be concluded that the institution has designed the planning program well by looking at the aspects included in the learning plan and those data.

Second, tutor qualification: teaching tutors must match the qualifications given by the institution or national regulation. Based on the interview results, two aspects are considered in the preparation of tutors to teach in the TOEFL class: the qualifications of tutor candidates and the development of tutor candidates' skills. In the TOEFL program, the educational background is not the main problem of being a tutor. The qualification of tutors who teach in the TOEFL program is to have a minimum score of 550 in the TOEFL real test, as cited in the director program's statement:

"The quality standard for educators in the TOEFL program is to have TOEFL score (the TOEFL standard of above 550). Some tutors do the Real Test within a minimum score of 550; then they can directly teach in the class (approximately 60% tutors in the institution have taken the real test)".

Despite the tutor saying something conversely, the tutor revealed that his TOEFL score was 500 the first time he joined the institution. Tutors who had a score of 500 were allowed to teach, but they had to attend a similar course and were willing to participate in skills development. The tutor put in a statement, "The teachers must have studied TOEFL before and got a TOEFL test with a minimum score standard of 500". It can be concluded that tutors can be accepted in the institution even though the initial test results are insufficient because there is a development skills program that the tutor must master before starting to teach. Development is a process to improve the quality of tutors in teaching; it is called Briefings. "... it is called Briefing, which is a kind of training to improve the tutors' mastered skills". Correspondingly, another tutor explained that:

"General development was holding through briefing, tutors in basic skill are developed to be able to teach at the second level such intermediate levels, tutors in intermediate levels are developed to the upper intermediate or to the advanced level, they are developed slowly and gradually, although when they become seniors, they will be developed more until they become briefers or trainer."

It is concluded that the skills development process at the institution for teachers is that basic tutors would have their teaching quality improved to an intermediate level, then they would be upgraded to upper mediate or advanced. Skills development efforts at the institution are also carried out by conducting comprehensive training every year.

Third, preparation for admission of course participants: the criteria for students allowed to take the TOEFL course have been found by interviewing the director of the TOEFL program. Based on the research results at this institution, the institution did not limit course participants to a certain age or ability level. The program director said:

"We accept whoever wants to join the TOEFL program then we would hold a pre-test before the class is started so that the classes are divided according to the results of the pre-test. For example, participants who got 400 score test and below of are treated in one class: (they will get more theory or basic English will be deepened), participants who got 400-450 score is gathered into one class, then participants who got 450 score and above will be in one class".

It can be concluded that there are no specific criteria that participants must fulfill in order to register as prospective course participants. Nevertheless, it does not mean that the institution is unaware of the circumstances and participants' abilities in class. To reduce gaps in learning of TOEFL preparation class, the institutions conduct a pre-test prior to the class. The aim of the test is to group the course participants according to the basic abilities that they have had before taking the course.

There are three class categories provided in the implementation of the TOEFL preparation class: firstly, a class devoted to participants who have a test score of less than 400; secondly, the class for participants who already have basics and have a test score of 400-450; thirdly, the class for participants who already have scored above average skills, such having a minimum of 450 scores on the pre-test. Each class receives special treatment, methods, and material according to their basic skills.

Another aspect considered in preparation for admission of the course participants by the course provider is transparency and suitability of admission of the course participants to the infrastructure provided in the course. The implementation of the pre-test is carried out by transparency of scores, thereby creating trust between course participants and course providers is essential. Admission of prospective course participants is not given specific qualifications, but it is adjusted to the facilities available at the institution. The facilities that have been given include classrooms that are comfortable for use. The course has provided 20 classrooms filled by 15-20 participants in one class—supported by available instructors such as 37 teachers each month. The program implemented at the time of the study has three classes provided, with each of them having ten participants per class. This clearly shows that the organizer in the institution has done good preparations.

Fourth, the preparation of facilities that would be used in learning: infrastructure is evaluated by looking directly at the quality and quantity of the infrastructure available at the institution. The planning and control of infrastructure facilities are carried out by the operational division and the program team. Meanwhile, the evaluation is carried out by the management team supported by another team. Evaluation has been done every week and every month for seeing the things that need improvement quickly. The director of the program argued that "Maintenance of the infrastructure is carried out by the management team, namely the program director, operational director, and marketing director. It is the operational division which has arranged the existing facilities professionally." The statement suggests that every director in the program has their respective roles in managing existing facilities. In maintaining the available facilities, socialization is given to tutors as the users of class facilities. Socialization is held once a month. Then, the socialization of the use of the facilities should carry forward to course participants; it is used to protect the institution's facilities.

The classroom and learning environment infrastructure are as follows. One classroom has been provided for a maximum capacity of 20 people, which is equipped with a noise reducer that supports listening learning. The institution provided 20 active speakers that are used in class for listening lessons. Based on the interview results, the sound system's quality is good, while the facilities available in the classroom have been assessed based on the standard of infrastructure for managing the courses as written in the [Regulation of the Minister of National Education No. 49 of 2007](#), presented in Table 2.

Table 2. Infrastructure Quantity and Quality

Quantity		Quality	
Total Score	54	Total Score	47
Maximum Score	75	Maximum Score	75
Percentage	72%	Percentage	62%

Based on these results, the percentage of the quantity of the course infrastructure is declared good enough and meets the applicable standards. The quality of the facilities and infrastructure is not the same as the result of the quantity calculation. However, the quality of the facilities can be said in a good line.

It can be concluded that the input provided by the course has been well prepared and complies with the national government standard of non-formal education management. Thus, [Sahayu and Friyanto \(2019\)](#) state that the classroom arrangement, tutor proficiency training,

the participant needs analysis, and material designs are the things that must be given more attention in managing an institution. Fulfillment of the participants' expectations and needs can be done by dividing participants into classes according to their abilities, giving a class for students of less than 30 course participants, and setting flexible schedules that allow participants to choose according to their needs (Wenjie, 2020). In the course management, the supporting components of the implementation of training and courses must be adjusted and well prepared as the situation that institutions have carried out in this research. This is carried out in accordance with applicable standards, and it is well accomplished by regular maintenance of the facilities.

Reaction

The assessment of participant reactions is divided into two aspects: participant satisfaction reactions to the TOEFL program and participant assessment reactions to the implementation of TOEFL learning in class. First, participants were allowed to fill out participant satisfaction questionnaires to assess the infrastructure provided, participant satisfaction reactions to the tutoring service, participant satisfaction reactions to the material presented, participant satisfaction reactions to the methods used in class, participant satisfaction reactions to the time management provided, and participant satisfaction reactions to training activities of the TOEFL questions implemented. The total number of questions on the satisfaction questionnaire is 36 statement items. The result of the interview to the participant will confirm the data; the number of questions is six questions. The results of the questionnaire obtained are depicted in a pie diagram in Figure 2.

**Participant Satisfaction Level
of the TOEFL Program**

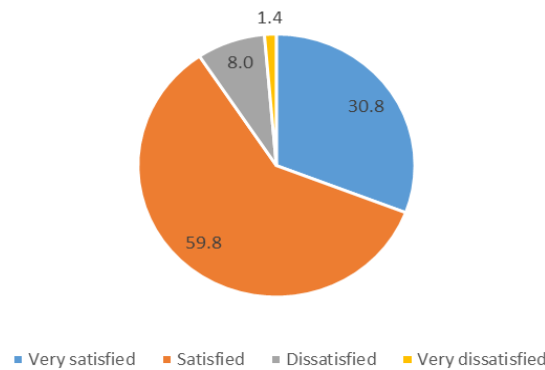


Figure 2. Participant Satisfaction Level

The data were generated through data analysis using a flowchart in the Microsoft Excel 2016. The data are the analysis result of course participants' satisfaction with the TOEFL program at a course institution that has been conducted for a month. The data show that the participants are very satisfied 30.8%, 59.8% satisfied, 8.0% dissatisfied, and 1.4% very dissatisfied. Thus, most of the course participants are satisfied with the implementation of the TOEFL program in English Village.

Based on the participants' interview, they are satisfied with the infrastructure provided because the class is comfortable, the facilities are adequate, modules have been fulfilled, media and teaching materials have been adjusted. In contrast, the participants are not satisfied with the existing facilities because it is sometimes hot in class during the day. It could be seen based on participant's statement, "Classroom conditions are quiet, comfortable, and have adequate

facilities”. Participants are satisfied with the tutor because the tutor provides the opportunity to ask questions for the participants, but one of the tutors sometimes gave the material fastly. Satisfaction with the material in the TOEFL program is caused by the material that is appropriate to the students’ level, even though some of them feel that the material is difficult. One of the participants stated that, ”The tutor teaches well using a good method, but sometimes the material is too hard for me”. The method used in the material teaching makes students understand it well. The schedule is prepared well based on the students’ needs, but they are disappointed because they should stop the program before one month because of the Covid-19. The participants are satisfied with the practice activities held in class since the tutor gives the tips and tricks to answer the TOEFL test.

Second, participants’ reaction in the implementation of learning is taken to see the participants’ assessment of the activities carried out in class. Assessment of the learning process is differentiated according to the class that has been carried out, which has different assessment results. Four classes are assessed: listening class, reading class, structure 1 class, and structure 2 class. The number of questions that course participants must answer is 17 questions in each session. The score for the course participant’s assessment of the TOEFL learning process is categorized based on Table 3.

Table 3. Assessment Score Category

Category	Interval
Very good	$X \geq 54.74$
Good	$54.74 > X \geq 42.50$
Bad	$42.50 > X > 30.26$
Very bad	$X < 30.26$

Based on calculations using Microsoft Excel, the score of the course participants’ assessment of the learning process is 55.56. It indicates that the results of data analysis of the four subjects that the participants have followed have an average score of 55.56 in the very good category. The research findings support this during observation; the result of the observation value is 9.70, with a very good category. Thus, it can be concluded that the learning process has gone very well.

It can be summed up that participants of the TOEFL preparation course program are included in the satisfactory category. Most of the participants are satisfied with the services provided by the institution. All learning sessions are in good categories. These results are different from previous research on the evaluation of the English for a specific exposure program. Participants gave a negative view of learning that did not improve the participants’ skills. The problems found were generalized classes, inaccurate methods, inappropriate materials, and also inappropriate syllabus designs to the needs of the participants (Alemi & Pazooki, 2020).

It can be said that the reaction results depend on the responses and ratings that are given by the participants. Thus, to avoid future problems, quality improvement must be continued. In line with this, Nazri et al. (2020) believe that participants must have good strategies and techniques in answering the TOEFL questions. However, most of the problems that arise are that teachers provide strategies to answer questions but they ignore the atmosphere and comfort built in the classroom. Further, Barnes (2016) explains that the TOEFL preparation class is very structured and goal-oriented. Therefore, teachers often limit the teaching styles and methods that are used. The results of this study show that the data are different from the aforementioned previous studies. The teachers in these institutions have taught well, and they can adjust the class to the condition so that the course participants do not get bored in learning. This can be reflected in the results of the learning process questionnaire, which achieve very good grades.

Outcome

Scoring is an assessment carried out once a week. Scoring is done to see the learning progress of the course participants. There are three skills included in the TOEFL test: listening, structure, and reading. These three skills are combined in one TOEFL test. The TOEFL test results that are calculated are the pre-test, scoring 1, scoring 2, and scoring 3, as described in Table 4.

Table 4. The TOEFL Test Results

Time	Comparison Score	Average Value	Score Increasing
Pre-Test	365	40	Initial score
Scoring 1	378	44	Increase
Scoring 2	391	48	Increase
Scoring 3	372	42	Decreased

Based on Table 4, the pre-test score or initial test was 365 with the correct number of 40 items. In the first scoring, it increased to 378 with the correct number of 44 items. The second scoring increased from the previous average score to 391 with 48 correct points, and the third scoring decreased to 372 with 12 correct points. Thus, there was a decrease in the score at the end of the course period.

The results of interview with course participants state that the improvement in learning is most felt in learning structures, even though some course participants felt that their TOEFL scores increased more in listening and reading skills. Some participants state they have an increase in learning even though the scores they get do not increase at the time of the TOEFL test. The problems faced during the exam are the speaker voice that is not heard, the processing time is too fast, and some participants come late, which disturb concentration.

Several problems can occur during the test, such as the results of this study. According to [Gür and Eriçok \(2020\)](#), understanding the material's content is the problem that can inhibit increasing the TOEFL test score. However, the institution in this study has anticipated this problem by dividing the participants based on the basic skills seen in the pre-test before the TOEFL class was held. The problems also occur during the test. [Yogawati and Widihastuti \(2019\)](#) state that students often face problems during exams, such as nervousness, low motivation, pessimism, mastery of vocabulary, low self-confidence. Another theory problem that often arises in exams is anxiety and stress fatigue, which can make it difficult for participants to answer questions ([Nikolaieva, 2016](#)). Problems often occur during the TOEFL test implementation, so it must be considered to reduce problems that will arise. Generally, it can be stated that the planning and implementation of the TOEFL test preparation program have been carried out properly by the institution.

CONCLUSION

Based on the research findings and discussion, the evaluation of the TOEFL program in the English village can be concluded as follows. The evaluation context is divided into the final objective and the program's ultimate objective, which are the ultimate objective. The context of the program has been fulfilled in accordance with the applicable national and real test standards. The input of the evaluation is well prepared, including a program plan that provides some aspects. Tutor qualification improvement is carried out every week and month, which is called the Briefing program. The admission of the course participants is based on a pre-test to see the participants' initial abilities. The facilities of the TOEFL program are in a good result; the quantity assessment score is 72%, and the quality is 62%. It can be said that the infrastructure has met the existing standards. However, it needs to be further improved and repaired in terms of its quality and quantity.

The reaction of participant satisfaction in the satisfied category is 90.6% and is 9.4% dissatisfied. It states that the TOEFL program is irresistible because many course participants are satisfied. The reaction of the participants to the learning process is in the good and very good categories. Generally, learning is carried out according to the learning procedures, so it is declared that the learning goes well. Learning outcomes increased in the first and second scoring but decreased in the third scoring with the same results as the pre-test. However, the decision in this evaluation is seen as a whole. It is concluded that the institution has implemented the program in accordance with the applicable standards.

REFERENCES

- Adib, H. S., Mardapi, D., Zamroni, Z., & Jait, A. (2019). Evaluation of Islam education teachers training implementation. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 23(2), 106–116. <https://doi.org/10.21831/pep.v23i2.20986>
- Aktaş, C. K., & Gündoğdu, K. (2020). An extensive evaluation study of the English preparatory curriculum of a foreign language school. *Pegem Eğitim ve Öğretim Dergisi*, 10(1), 169–214. <https://doi.org/10.14527/pegegog.2020.007>
- Alemi, M., & Pazooki, S. J. (2020). *A stakeholder-based evaluation of engineering ESP courses' effectiveness*. <https://doi.org/10.21203/rs.3.rs-58769/v1>
- Ali, H. (2012). The use of silent reading in improving students' reading comprehension and their achievement in TOEFL score at a private English course. *International Journal of Basic and Applied Science*, 1(1), 47–52. <https://doi.org/10.17142/ijbas-2012.1.1.8>
- Barnes, M. (2016). The washback of the TOEFL iBT in Vietnam. *Australian Journal of Teacher Education*, 41(7), 158–174. <https://doi.org/10.14221/ajte.2016v41n7.10>
- Britton, J., & Vignoles, A. (2020). Education production functions. In G. Johnes, J. Johnes, T. Agasisti, & L. López-Torres (Eds.), *Handbook of contemporary education economics* (pp. 246–271). Edward Elgar Publishing. <https://doi.org/10.4337/9781785369070.00016>
- Choudhury, G. B., & Sharma, V. (2019). Review and comparison of various training effectiveness evaluation models for R & D organization performance. *PM World Journal*, 8(2), 1–13.
- Dahuri, D. (2019, December 12). Indeks kemampuan bahasa Inggris orang Indonesia nomor 61. *Media Indonesia*. <https://mediaindonesia.com/humaniora/277217/indeks-kemampuan-bahasa-inggris-orang-indonesia-nomor-61>
- Dewi, L. R., & Kartowagiran, B. (2018). An evaluation of internship program by using Kirkpatrick evaluation model. *Research and Evaluation in Education*, 4(2), 155–163. <https://doi.org/10.21831/reid.v4i2.22495>
- Direktorat Pembinaan Kursus dan Pelatihan. (2018). *Data dan informasi kursus dan pelatihan 2018*. Direktorat Jenderal Pendidikan Anak Usia Dini dan Pendidikan Masyarakat.
- English First. (2020). *The world's largest ranking of countries and regions by English skills*. English First - English Proficiency Index. <https://www.ef.co.id/epi/>
- Gonçalves, G. J., & Chauma, A. M. (2020). Challenges for foreign English language teacher education programme in Mozambique with focus on Zambézia Teacher Training Colleges. *London Journal of Research in Humanities and Social Sciences*, 20(1), 23–34. <https://research.journalspress.com/index.php/socialscience/article/view/585>

- Gür, R., & Eriçok, B. (2020). The relationship among academic success scores of graded foreign language courses. *Dil ve Dilbilimi Çalışmaları Dergisi*, 16(2), 809–821. <https://doi.org/10.17263/jlls.759309>
- Ismail, I., & Othman, R. (2020). A review of literature on the English language entry requirement for international students into postgraduate programs in Universiti Teknologi Malaysia. *Journal of Critical Reviews*, 7(11), 543–549. <http://www.jcreview.com/?mno=117854>
- Kurniawati, E. (2017). Manajemen strategik lembaga pendidikan Islam dalam meningkatkan mutu pendidikan: Studi kasus di Madrasah Aliyah Nahdlatul Ulama Gondang Sragen. *At-Taqaddum*, 9(1), 113–132. <https://doi.org/10.21580/at.v9i1.1784>
- Liu, F. (2020). Failure of humanities-based instruction to achieve students' language goal in college English courses. *Journal of Language Testing & Assessment*, 3(1), 1–4. <https://doi.org/10.23977/langta.2020.030101>
- Ma, J., & Cheng, L. (2016). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth, and significance. *TESL Canada Journal*, 33(1), 58–79. <https://doi.org/10.18806/tesl.v33i1.1227>
- Mardapi, D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan* (Revised ed). Parama Publishing.
- Maruf, Z., Rahmawati, A. S., Siswantara, E., & Murwantono, D. (2020). Long walk to quality improvement: Investigating factors causing low English proficiency among Indonesian EFL students. *International Journal of Scientific & Technology Research*, 9(3), 7260–7265. <http://www.ijstr.org/final-print/mar2020/Long-Walk-To-Quality-Improvement-Investigating-Factors-Causing-Low-English-Proficiency-Among-Indonesian-Efl-Students.pdf>
- Mulyasa, E. (2005). *Menjadi guru profesional*. Remaja Rosda Karya.
- Nazri, M. A., Wijaya, H., & Zainurrahman, Z. (2020). EFL students' ability in answering TOEFL reading comprehension section. *Journal of Physics: Conference Series*, 1539, 012044. <https://doi.org/10.1088/1742-6596/1539/1/012044>
- Nikolaieva, O. (2016). *A qualitative study on preparing EFL students to take the TOEFL internet-based (iBT) test in the Ukrainian context* [Master thesis, University of Stavanger]. <https://uis.brage.unit.no/uis-xmlui/handle/11250/2400189>
- Nurhayati, N., Hendrawaty, N., & Angkarini, T. (2013). The acquisition of English as a foreign language in Pare East Java (Kampung Inggris) (A case study of what and how the acquisition of English in Pare). *Deiksis*, 5(2), 81–88. <https://journal.lppmunindra.ac.id/index.php/Deiksis/article/view/462>
- Perez, B. E. O., & Mardapi, D. (2015). Evaluation of the bridging course offered at a university to foreign students: Batches of 2012 and 2013. *Research and Evaluation in Education*, 1(2), 146–157. <https://doi.org/10.21831/reid.v1i2.6667>
- Regulation of the Minister of National Education No. 49 of 2007 on the Standard of Educational Management by Nonformal Educational Units, (2007).
- Sahayu, W., & Friyanto, F. (2019). The effect of Youtube on high school students' second language acquisition. *International Journal of Linguistics, Literature and Translation (IJLLT)*, 2(6), 38–44. <https://doi.org/10.32996/ijllt.2019.2.6.5>

- Sakurai, N. (2020). Exploring a placement test for extensive reading programs. *Humanities Series*, 2(3), 53–70. https://ksu.repo.nii.ac.jp/?action=repository_action_common_download&item_id=10468&item_no=1&attribute_id=22&file_no=1
- Silva, A. L. B. de C. e, & Tosqui-Lucks, P. (2020). Around the world in Aeronautical and Aviation English courses. *Revista CBTeLE*, 2(1), 418–440. <https://revista.cbtecle.com.br/index.php/CBTeLE/article/view/274>
- Sudjana, S. (2004). *Pendidikan nonformal*. Falah Production.
- Sutton, B. (2006). *Adopting a holistic approach to the valuation of learning programmes deployed in corporate environments* [Thesis, Middlesex University, London]. <https://eprints.mdx.ac.uk/2667/>
- Syakur, A., Zainuddin, H. ., & Hasan, M. A. (2020). Needs analysis English for specific purposes (ESP) for vocational pharmacy students. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*, 3(2), 724–733. <https://doi.org/10.33258/birle.v3i2.901>
- Syamsuddin, I., & Min, A. (2014). Assessing moodle as learning management system platform for English course based TOEFL. *International Journal of Computer Trends and Technology*, 18(6), 276–279. <https://doi.org/10.14445/22312803/IJCTT-V18P158>
- Topno, H. (2012). Evaluation of training and development: An analysis of various models. *IOSR Journal of Business and Management*, 5(2), 16–22. <https://doi.org/10.9790/487X-0521622>
- Wang, Y. (Tina). (2019). The impact of TOEFL on instructors' course content and teaching methods. *The Electronic Journal for English as a Second Language*, 23(3), 1–18.
- Wenjie, S. (2020). Evaluating an English course for master students in China: A case of business English for accounting program. *International Journal of English Language Teaching*, 7(1), 31–40. <https://doi.org/10.5430/ijelt.v7n1p31>
- Yogawati, N. D., & Widihastuti, W. (2019). Evaluating the implementation of English communication therapy (ECT): An objective structured clinical assessment (OSCA) approach. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 23(1), 87–94. <https://doi.org/10.21831/pep.v23i1.22449>
- Zhao, L. (2020). A study on the feedback of college English dynamic classified teaching effect. *Theory and Practice in Language Studies*, 10(6), 678–684. <https://doi.org/10.17507/tpls.1006.08>

A psychometric evaluation of the career decision making self-efficacy scale

Chandra Yudistira Purnama*; Linda Ernawati

Universitas Jenderal Achmad Yani

Jl. Terusan Jenderal Sudirman, Cibeber, Cimahi Selatan, Kota Cimahi, Jawa Barat 40531, Indonesia

*Corresponding Author. E-mail: chandra.yudistira@lecture.unjani.ac.id

ARTICLE INFO

Article History

Submitted:

8 April 2021

Revised:

25 June 2021

Accepted:

29 June 2021

Keywords

career; CDMSE (career decision making self efficacy); CFA

Scan Me:



ABSTRACT

The assessment tool for a career currently being developed requires special treatment from a psychologist/psychometrist. The measurements are conducted when students are confused about career options. However, for students who have decided, it is uncommon for them to seek professional help. Psychological tools that focus on capturing information about students' maturity in relation to their ability to make career decisions can help them choose a major that is suitable for their career. This study concerns adapting the career decision-making self-efficacy (CDMSE) that can predict one's confidence in his/her ability to make career choices. The adaptation of this instrument went through several stages such as translation, back translation, testing the reliability, and testing the validity evidence of content and internal structure using confirmatory factor analysis (CFA). This study used a sample of 539 high school students in Bandung and Cimahi. The construct reliability (CR) of the instrument was $\alpha=0.929$. The evidence for internal structure using CFA showed that the CDMSE scale has an acceptable goodness of fit index. The standardized loading factor item is in the range 0.710-0.998. It can be concluded that the Bahasa Indonesia version of the CDMSE scale has good psychometric properties and can be used for research or assessment to measure a person's degree of confidence about his/her ability to make career choices.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



How to cite:

Purnama, C., & Ernawati, L. (2021). A psychometric evaluation of the career decision making self-efficacy scale. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 77-87. doi:<https://doi.org/10.21831/jpep.v25i1.39960>

INTRODUCTION

Senior high school students have an age range between 15-19 years so that with this age range, they can be categorized into adolescence. Adolescence is a period where a person begins to be faced with a lot of situations where they have to choose. In this period, one of the development tasks was to choose and prepare to carry out a job (Hurlock, 1972). According to Seligman (1994), at approximately 17 years old, teens realized that they are responsible for his career planning. In adolescence, career development runs along with getting older and experiencing dynamics that are important in the school (Seligman, 1994). Senior high school students are in the period to determine a major in higher education to achieve the desired career. Adolescents are in the phase of self-determination. They are required to develop their self-identity and make decisions in accordance with environmental demands (Lewis, 1981). During this period, adolescents are expected to know and realize the need to make career decisions, understand their potential, be aware of their interests and talents, measure their own abilities, and identify suitable job opportunities (Walsh et al., 2000).

The selection of fields of work is closely related to the selection of educational programs to be pursued. The chosen area of educational programs can support their success when starting and pursuing their future careers. Thus, one must understand the demands of their choice of work (Islamadina & Yulianti, 2017). A certain job will require special abilities, expertise, or

skills that can be obtained through formal or informal learning. In their career development, adolescence is filled with exploration (Suryanti et al., 2011). Adolescents try to explore all the possibilities. They try to understand everything that will become their needs in facing future challenges (Super & Super, 2001).

The current phenomenon among adolescents is that many students experience confusion when they have to choose a major and their future careers (Creed et al., 2005). There are still many cases of adolescents who choose a major in university without considering their abilities, skills, interests, talents, or personality. Most of them tend to choose majors for prestige, following peers, the trends of professions, job popularity, income, certain figures, and even their parents, and some are because of their parent's wishes.

Thus, determining the majors in higher education and future careers at the end of high school becomes an important moment. The decision to make and plan career choices supports the individual's success in the future life. Their understanding of the type of career, the field of work, interests, and talents is important before deciding which career to pursue. Students who have a strong career maturity can make decisions steadily, while those with low career maturity will experience confusion and have no clear career plan. Regarding this, many students seek help from school counselors, teachers, career consultants, and even educational psychologists to help them and provide insight into their potential to avoid mistakes in choosing a major and establishing a career. One method to see an individual's maturity degree in determining the career is using psychological assessment tools. Currently, many psychological assessment tools can measure this to help students be optimal in determining careers. One such measure is Career Decision Making Self Efficacy (henceforth CDMSE) (Betz et al., 1996).

CDMSE scale was developed by Taylor and Betz in 1983 consisting of five dimensions, namely, career choice competencies in the areas of goal setting (GS), gathering occupational information (GI), problem-solving (PS), planning (PL), and also self-appraisal (SA) (Betz et al., 1996). From the five dimensions, 50 items were compiled. Each dimension is represented by ten statements. In 1996, the instrument was revised by Betz, Klein, and Taylor to only 25 items. The statement items are selected from the best five statements from each dimension (Betz et al., 1996). CDMSE are self-report using Likert scale with five options, namely strongly agree, agree, neutral, disagree, and strongly disagree.

Observing the CDMSE aspects, this measurement tool will greatly assist education counselors, counseling teachers, career consultants, and educational psychologists in assessing and predicting the right career prospect for students in universities, helping plan, and choosing a career in the future. Measuring instruments can be used to capture information about students, especially the suitability between individual characteristics and the desired career choice.

One of the challenges that sometimes surge is when there is a mismatch of the items used in the instruments made internationally when it is used in Indonesia. The incompatibility of the context of language, culture, or even the meaning of the term can be misinterpreted, so it is necessary to re-research and adapt the assessment tool to Indonesia's cultures. The purpose of adapting this measuring instrument is to obtain a CDMSE measuring instrument in accordance with Indonesian culture and the provisions of psychometric rules. The adaptation focuses on language, terms, diction, wording, scaling, and norm. Besides, by re-conducting research and adapting the instruments, it is hoped that items or statements obtained are in line with the culture of students in Indonesia and the validity, reliability, and norms to interpret the results of the scores of these students.

RESEARCH METHOD

Data collection for testing the CDMSE instrument was carried out online using the Google Form involving 539 high school students in Bandung and Cimahi. In detail, a description of the respondents participating in this CDMSE instrument testing is shown in Table 1.

Table 1. The Description of the Respondents

Category	N = 539
Gender	
Male	219 students
Female	320 students
Age	
17 years old	489 students
18 years old	50 students

Based on the demographic description in Table 1, respondents who participated in the testing of the CDMSE instrument were dominated by female participants, with a total of 320 respondents (59%) and 219 men (41%). The age of respondents was dominated by students aged 17 years old with a total of 489 respondents (90.7%), and aged 18 years old, with total of 50 respondents (9.3%), and the mean age of the respondents was 17.09 years old (SD=0.29).

Procedure

The procedure of adaptation CDMSE scale begins with the process of translating a foreign language (English) into Bahasa and back translating it from Indonesian to a foreign language (English) (explained at the translation stage). After completing the translation process and being approved to proceed to the next stage, the next step is testing instruments CDMSE using statistical methods. The process carried out to obtain the statistical analysis results is by distributing online questionnaires to respondent high school students in Bandung and Cimahi. After the data was obtained, and then continued by analyzing the reliability coefficient of the measuring instrument and evidence of validity based on the internal structure using confirmatory factor analysis. The stages of adaptation to the CDMSE scale and its data processing are elaborated as follows.

Stage 1: The Career Decision Making Self Efficacy Scale Translation Process

In the first stage, the researchers' initial step was to translate the instrument from English to Bahasa Indonesia. This stage refers to the process of adapting assessment tools based on the guidelines from the International Test Committee (ITC) guidelines for translating and adapting tests (International Test Commission, 2016). The process of translating CDMSE in English into Bahasa Indonesia was carried out by four people separately. The first and second translators are professional translators who have a bachelor's degree in English literature and work as English teachers. The third and fourth translators are psychologists who have experience constructing psychological assessment tools for both academic and practical needs.

The second step was done after obtaining the translation from the four translators. The translators and researchers discussed, reviewed, and made revisions to the translation results. The final result in the second stage was to obtain a Bahasa Indonesia version of the CDMSE draft. In the third step, the initial manuscript of the Bahasa Indonesia version of the CDMSE was re-translated into English by three professional translators. Two translators work as English teachers in high school, and as professional translators, another person works as a teacher and translator at the English Language Course Institute. The results of the re-translation into English were checked for their suitability in meaning by comparing the CDMSE translation results from Bahasa Indonesia to English with the original English version of the CDMSE assessment tool. The wording of the questionnaire sentences in Bahasa Indonesia that did not match or have different meanings from the English version was corrected and revised again to get the appropriate and relevant words. Next, in the fourth stage, the Bahasa Indonesia version of the CDMSE manuscript, which was revised and adjusted based on the input from the re-translation process, was submitted to four experts to get a review of the clarity and appro-

priateness of conceptualization the aspects being measured. The three experts involved in reviewing the Bahasa Indonesia version of the CDMSE manuscript were lecturers at the Faculty of Psychology, Universitas Achmad Yani (henceforth UNJANI), who had experience compiling measurement tools in psychology and teaching career development courses. Meanwhile, one other person was a psychology doctoral student at the Faculty of Psychology, Universitas Padjajaran Bandung. The four experts reviewed, provided input, and corrected the wording of the translated items on the Bahasa Indonesia version of the CDMSE assessment tool. The experts were given attachments of the original English version of the CDMSE instrument, the translation result from English to Bahasa Indonesia, the CDMSE manuscript agreed to be re-translated into English, the results of the re-translation from Bahasa Indonesia to English, and the Bahasa Indonesia version of the CDMSE manuscript that has been adjusted. After the experts gave suggestions, comments, input, and corrections to the less relevant or inappropriate items, the researchers made improvements to the wording of the items. These improvements were discussed again and were followed up to get the final manuscript of the Bahasa Indonesia version CDMSE instrument.

In the fifth stage, the Bahasa Indonesia version of the CDMSE final manuscript was distributed to 20 students of the Faculty of Psychology UNJANI to be tested for its readability. This stage is to get valid evidence based on test responses from the subject. The process done with students was to read together with the Bahasa Indonesia version of the CDMSE final script with a loud speaks, then asked for their explanation and confirmed the understanding of each student on each item in the Bahasa Indonesia version of the CDMSE manuscript. After confirmation of the students' readability and comprehensiveness, the Bahasa Indonesia version of the CDMSE final manuscript was converted into a digital/online version using the Google Form. Then, the Bahasa Indonesia version of the CDMSE scale that has been converted into an online version was distributed to students in Bandung and Cimahi through the counseling teachers in each school. During the one week of data collection, 539 high school students filled the Google Form.

Stage 2: Reliability Testing and Model Testing of Carrer Decision Making Self Efficacy Scale

The second stage was done to gain the reliability coefficient of the assessment tool using construct reliability (CR), average variance extracted (AVE), and testing the internal structure of CDMSE using confirmatory factor analysis (CFA). Data processing for testing the measurement tools (reliability test and validity evidence) was done using JASP 0.14.1 and Lisrel. This CDMSE scale consists of five dimensions (GI, SA, GS, PL, and PS), and each dimension consists of five items. Thus, the total number of items in this instrument is 25 items.

The data collection to obtain reliability values and validity evidence of the internal structure involved 539 high school students in Bandung and Cimahi. The CDMSE instrument that had been translated and approved through an expert review was compiled into an online questionnaire using the Google Form. The link to the questionnaire was sent to the counseling teachers in each school and distributed to students. Within one week, 539 respondents submitted the answers to be used in the data processing to test reliability and obtain valid internal structure evidence.

Data Analysis

The researchers used the JASP version 0.14, Lisrel program, and Statcal to process the data and get a composite reliability score and validity evidence from the CDMSE scale. Composite Reliability is a reliability coefficient that can be used for multidimensional measures (Heise & Bohrnstedt, 1970), and confirmatory factor analysis is used to confirm whether the indicator variables can be used to confirm a factor (Ferdinand, 2011).

FINDINGS AND DISCUSSION

Reliability Test and Item Analysis of Carrer Decision Making Self Efficacy Scale

Data processing for the reliability test of the CDMSE instrument was carried out using Microsoft Excel 2019, JASP version 0.14, and Statcal. Reliability can be expressed as the internal consistency of an instrument that can be measured based on the level of item homogeneity. [Hair et al. \(2010\)](#) explain that the reliability test in the CFA analysis includes the construct reliability (CR) and variance extracted (AVE). [Hair et al. \(2010\)](#) state that the CR value ≥ 0.7 is good reliability, while the CR value between 0.6 and 0.7 is acceptable reliability, with a note that the indicator has a factor load that matches the criteria. Internal consistency can also be measured using the Average Variance Extracted (AVE) estimate. The recommended AVE value is > 0.5 ([Hair et al., 2010](#)).

Table 2. Construct Reliability (CR) and Average Variance Extracted (AVE)

Variable	CR	AVE
Gathering of Information	0.84	0.51
Self Appraisal	0.83	0.51
Goal Selection	0.86	0.55
Planning	0.83	0.53
Problem Solving	0.89	0.63
CDMSE	0.97	0.88

Source: Statcal

Table 3. Frequentist Individual Item Reliability Statistics

If Item Dropped			If Item Dropped		
Item	McDonald's ω	Item-Rest Correlation	Item	McDonald's ω	Item-Rest Correlation
GI 1	0.957	0.575	GS 4	0.956	0.670
GI 2	0.956	0.644	GS 5	0.955	0.745
GI 3	0.955	0.747	PL 1	0.956	0.640
GI 4	0.955	0.708	PL 2	0.956	0.652
GI 5	0.956	0.682	PL 3	0.956	0.622
SA 1	0.956	0.686	PL 4	0.957	0.602
SA 2	0.956	0.632	PL 5	0.956	0.697
SA 3	0.955	0.700	PS 1	0.955	0.730
SA 4	0.956	0.608	PS 2	0.955	0.777
SA 5	0.956	0.632	PS 3	0.955	0.757
GS 1	0.956	0.658	PS 4	0.956	0.672
GS 2	0.955	0.746	PS 5	0.956	0.685
GS 3	0.955	0.711			

Source: [Goss-Sampson \(2018\)](#). JASP (Version 0.14.1) [Computer software]

Table 2 shows the construct reliability (CR) and average extracted variance (AVE), while Table 3 informs about item-rest correlation. Table 2 depicts the results of data processing for the reliability test CDMSE. It shows that the CR value for the CDMSE instrument is 0.97. An assessment instrument is said to be reliable if the CR coefficient value is greater than 0.7 ($\alpha \geq 0.7$) ([Hair et al., 2019](#)), so the CDMSE instrument adapted into Bahasa Indonesia is reliable. Table 3 provides information about item analysis to see the quality of items on the CDMSE scale. The data shows the quality of items on the CDMSE scale, whether they have good items or not, by looking at the scores in the item-rest correlation column. An item is good if it has an item-rest correlation value greater than 0.3 ([Pallant, 2011](#)). The test results on the item quality obtained the item-rest correlation coefficient value in the range of 0.575 to 0.777.

Validity Evidence

The examination of the validity evidence on the CDMSE scale used guidelines from AERA, that is, evidence based on test content, evidence based on test responses, evidence based on internal structure, evidence based on relation to other variables, and evidence based on consequences of the testing (American Educational Research Association (AERA) et al., 2014). Of the five pieces of evidence validity, the researchers only carried out three pieces of evidence. First, it was evidence-based on test content obtained from the subject matter expert (henceforth SME). Second, it was based on evidence-based test responses obtained from the readability of students and lecturers. Lastly, the third was evidence based on the internal structure using confirmatory factor analysis.

In obtaining valid evidence based on test content, the researchers asked the SME to assess whether the items in the CDMSE scale were relevant and in accordance with the construct. The SMEs gave a score of 1 for items that were considered strongly irrelevant, a score of 2 for items considered irrelevant, a score of 3 for items considered fair enough relevant, a score of 4 for items considered relevant, and a score of 5 for items that were considered strongly relevant. The context of the assessment is viewed from the suitability of the language, clarity of the wording, and the suitability of the meaning of the translated sentence. The results of the SME assessment were processed with Aiken's V formula (Aiken, 1985). The Aiken's V value for all CDMSE items ranged from 0.83 to 0.92. Hence, the CDMSE scale can be used to measure confidence in making career decisions.

In the process of obtaining valid evidence in the form of evidence based on test responses, it is obtained through the readability of 20 respondents. Based on the responses from students, the items in the CDMSE scale can be understood well.

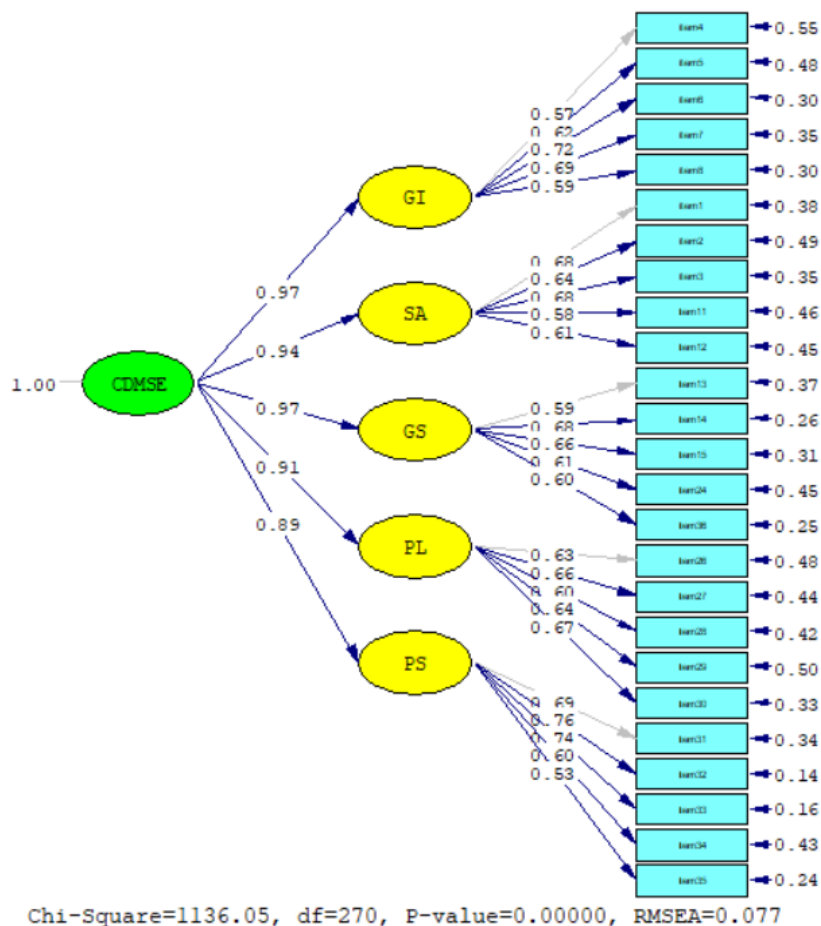


Figure 1. The Results of the Model Fit Test

The next evidence is evidence based on the internal structure using confirmatory factor analysis (CFA). The benchmarks used to interpret the suitability of the model in this study refers to Hu and Bentler who recommended four parameters, namely, the Chi-Square Test p -value ≥ 0.05 ; the Root Mean Square Error of Approximation (RMSEA) ≤ 0.08 ; Comparative Fit Index (CFI) ≥ 0.95 ; and Standardized Root Mean Square Residual (SRMR) ≤ 0.08 (Hu & Bentler, 1999). Meeting these criteria means that the instrument meets the appropriate model criteria. Figure 1 shows the results of the model fit test using CFA.

Table 4. The Goodness of Fit Indices from the CDMSE Model

Category	Parameter Fit	Output	Cut Off	Criteria	Information
Absolute Fit	Chi square p-value	< 0.000	> 0.05	Good Fit	Not Fit
	GFI	0.86	≥ 0.90 0.80 - 0.89	Good Fit Marginal Fit	Marginal Fit
	RMSEA	0.07	≤ 0.08 0.08 - 0.10	Good Fit Marginal Fit	Good Fit
	SRMR	0.048	< 0.08 0.08 - 0.10	Good Fit Marginal Fit	Good Fit
	Incremental Fit	AGFI	0.83	≥ 0.90 0.80 - 0.89	Good Fit Marginal Fit
	NFI	0.97	≥ 0.90 0.80 - 0.89	Good Fit Marginal Fit	Good Fit
	IFI	0.98	≥ 0.90 0.80 - 0.89	Good Fit Marginal Fit	Good Fit
	CFI	0.98	≥ 0.90 0.80 - 0.89	Good Fit Marginal Fit	Good Fit
	Parsimonious Fit	PNFI	0.87	> 0.5	Good Fit

Source: Ghozali & Fuad (2005)

Based on the goodness of fit indices test presented in Table 4, there are six that meet the criteria, namely, RMSEA of $0.07 \leq 0.08$, SRMR of $0.048 \leq 0.08$, NFI of $0.97 \geq 0.90$, IFI of $0.98 \geq 0.90$, CFI of $0.98 \geq 0.90$ and PNFI $0.87 \geq 0.05$. There are two criteria that meet marginal fit criteria: GFI of $0.85 \leq 0.90$, AGFI of $0.83 \leq 0.90$, and one criterion that cannot be fulfilled is the chi-squared test because the p -value obtained < 0.001 . The Chi-Square value is the traditional measure to evaluate the suitability of the overall model (Hu & Bentler, 1999). A good fit model will give insignificant results at the 0.05 threshold (Barrett, 2007). The Chi-square index is the most used index to check the accuracy of the model. However, this index is strongly affected by the sample size (Bergh, 2015). If the sample is too small, the trend will be insignificant, while the trend will be significant if the sample is too large. Thus, the Chi-square almost certainly rejects the model if a large number of samples are used. In this study, the sample size used was 539.

Based on the test results of the fit model, from nine proposed parameters, there are eight that meet the criteria, six criteria (RMSEA, SRMR, NFI, IFI, CFI, and PNFI) for a good fit, and two criteria (GFI and AGFI) for marginal fit, while one criterion (Chi-square) does not meet the criteria. Therefore, the researchers tried to modify the model to get a better model. Attempts to modify the model are by deleting items. The results of model testing after modification are shown in Figure 2.

Based on the modification results, ten items were obtained that matched the model, including GI3, GI4, SA1, SA2, GS2, GS3, PL4, PL5, PS2, and PS5. Then, the results of the goodness of fit of indices test are shown in Table 5. From Table 5, the model is fit. There is no significant difference between the model ideal with the proposed model based on measurements. All parameters of goodness of fit indices have also been according to the criteria set for

obtaining a model fit. Thus, the final model is already fit, which means the model proposed fits the empirical data. The complete model and loading factor of each item in the final model can be seen in Figure 2.

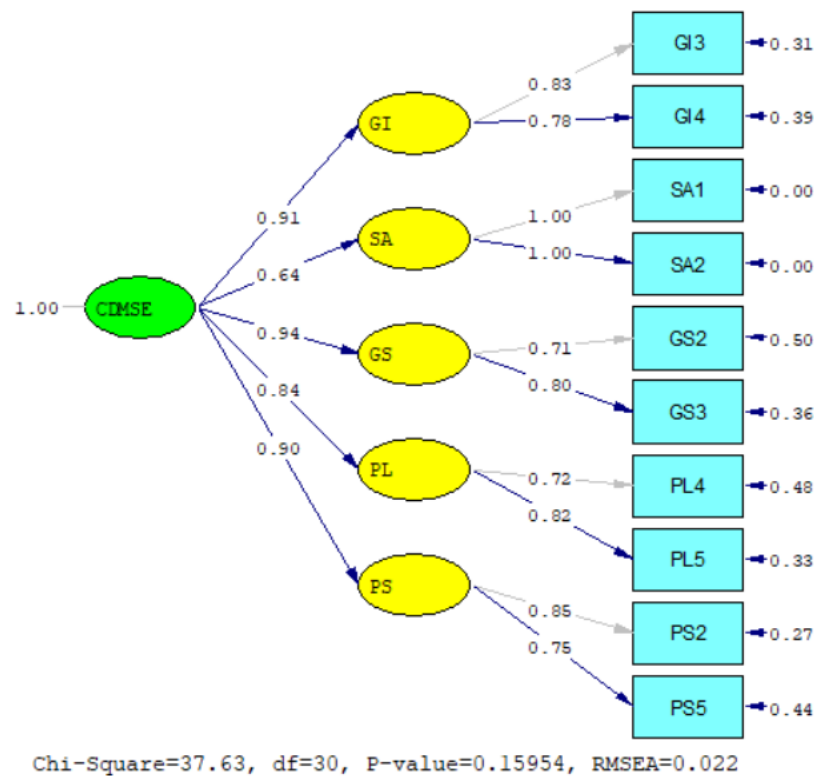


Figure 2. The Results of the Model Fit Test

Table 5. The Goodness of Fit Indices from the CDMSE Model

Category	Parameter Fit	Output	Cut Off	Criteria	Information
Absolute Fit	Chi square p-value	0.159	> 0.05	Good Fit	Good Fit
	GFI	0.99	≥ 0.90	Good Fit	Good Fit
			0.80 - 0.89	Marginal Fit	
	RMSEA	0.022	≤ 0.08	Good Fit	Good Fit
			0.08 - 0.10	Marginal Fit	
Incremental Fit	SRMR	0.020	< 0.08	Good Fit	Good Fit
			0.08 - 0.10	Marginal Fit	
	AGFI	0.97	≥ 0.90	Good Fit	Good Fit
			0.80 - 0.89	Marginal Fit	
Parsimonious Fit	NFI	0.99	≥ 0.90	Good Fit	Good Fit
			0.80 - 0.89	Marginal Fit	
	IFI	0.99	≥ 0.90	Good Fit	Good Fit
			0.80 - 0.89	Marginal Fit	
	CFI	0.99	≥ 0.90	Good Fit	Good Fit
			0.80 - 0.89	Marginal Fit	
Parsimonious Fit	PNFI	0.66	> 0.5	Good Fit	Good Fit

Source: Ghozali & Fuad (2005)

The next part examines the results of the factor loading analysis of the CDMSE assessment tool. The test results can be seen in Table 6. The criteria for an item to be said to have a good factor loading is when the factor loading value is ≥ 0.5 (Hair et al., 2019).

Table 6. Factor Loading of CDMSE

Factor	Indicator	Std. Est. (all)
Gathering of Information	GI 3	0.827
	GI 4	0.782
Self Appraisal	SA 1	0.999
	SA 2	0.998
Goal Selection	GS 2	0.710
	GS 3	0.794
Planning	PL 4	0.720
	PL 5	0.823
Problem Solving	PS 2	0.856
	PS 5	0.750

Based on the results of the standardized loading factors, the CDMSE items and each dimension have a factor loading value above 0.5 with a range between 0.710 - 0.998. This shows that the quality of the items is classified as good.

After testing the model and standardized loading factors, a reliability test was then carried out. Reliability testing in CFA includes the construct reliability (CR) and variance extracted (AVE). The results of the reliability test are shown in Table 7 and Table 8.

Table 7. Construct Reliability (CR) and Average Variance Extracted (AVE)

Variable	CR	AVE
Gathering of Information	0.786	0.648
Self Appraisal	0.998	0.997
Goal Selection	0.723	0.568
Planning	0.747	0.597
Problem Solving	0.786	0.648
CDMSE	0.929	0.726

Source: Statcal

Table 8. Frequentist Individual Item Reliability Statistics

Item	If Item Dropped	
	McDonald's ω	Item-Rest Correlation
GI3	0.896	0.714
GI4	0.898	0.683
SA1	0.898	0.688
SA2	0.898	0.687
GS2	0.902	0.605
GS3	0.897	0.701
PL4	0.906	0.576
PL5	0.900	0.663
PS2	0.828	0.729
PS5	0.901	0.640

Based on Table 7, the CR value for the CDMSE instrument after modification is 0.929. Thus, based on the test results, the CDMSE instrument adapted into Bahasa Indonesia is reliable. Table 8 shows that the test results on the item quality obtained the item-rest correlation coefficient value from 0.576 to 0.729.

The results of this study were conducted on a sample of students from Bandung and Cimahi. This is one of the limitations of this study because it does not involve many students from other cities in Indonesia. Therefore, for further research, it is recommended to expand the coverage of student participants from various cities and regions in Indonesia and increase the sample size to obtain a more fit model with a larger scale of participants.

CONCLUSION

Based on the findings of the research regarding the psychometric analysis of the adaptation of the CDMSE scale into Bahasa Indonesia, some conclusions can be drawn, as follows. (1) The CDMSE scale test results can be said reliable. It means that it can consistently measure an individual's level of self-confidence regarding his or her ability to make career choices. (2) The CDMSE scale with five dimensions has a model that fits the original construct based on Goodness of Fit Indices. (3) The test results showed good item quality and can accurately measure the dimensions of one's CDMSE. Thus, these items can measure the level of an individual's self-confidence regarding his or her ability to make career choices. (4) The CDMSE scale can be used to measure a person's CDMSE scale and as an additional variety of instruments related to career assessment tools.

ACKNOWLEDGMENT

The authors express their heartfelt gratitude to the Institute of Research and Community Service (LPPM) of Universitas Jenderal Achmad Yani for supporting and funding this research.

REFERENCES

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME).
- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5), 815–824. <https://doi.org/10.1016/j.paid.2006.09.018>
- Bergh, D. (2015). Chi-squared test of fit and sample size - A comparison between a random sample approach and a Chi-square value adjustment method. *Journal of Applied Measurement*, 16(2), 204–217.
- Betz, N. E., Klein, K. L., & Taylor, K. M. (1996). Evaluation of a short form of the career decision-making self-efficacy scale. *Journal of Career Assessment*, 4(1), 47–57. <https://doi.org/10.1177/106907279600400103>
- Creed, P., Prideaux, L.-A., & Patton, W. (2005). Antecedents and consequences of career decisional states in adolescence. *Journal of Vocational Behavior*, 67(3), 397–412. <https://doi.org/10.1016/j.jvb.2004.08.008>
- Ferdinand, A. T. (2011). *Metode penelitian manajemen: Pedoman penelitian untuk penulisan skripsi, thesis, dan disertasi Ilmu Manajemen*. Universitas Diponegoro.
- Ghozali, I., & Fuad, F. (2005). *Structural equation modeling: Teori, konsep, & aplikasi dengan program Lisrel 8.54*. Universitas Diponegoro.
- Goss-Sampson, M. A. (2018). *Statistical analysis in JASP: A guide for students*. <https://static.jasp-stats.org/Statistical Analysis in JASP - A Students Guide v2.pdf>
- Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Pearson.
- Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, 2, 104–129. <https://doi.org/10.2307/270785>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Hurlock, E. B. (1972). *Child development (McGraw-Hill series in psychology)*. McGraw-Hill.
- International Test Commission. (2016). *The ITC guidelines for translating and adapting tests* (2nd ed.). International Test Commission (ITC). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf
- Islamadina, E. F., & Yulianti, A. (2017). Persepsi terhadap dukungan orangtua dan kesulitan pengambilan keputusan karir pada remaja. *Jurnal Psikologi*, 12(1), 33–38. <https://doi.org/10.24014/jp.v12i1.3006>
- Lewis, C. C. (1981). How adolescents approach decisions: Changes over grades seven to twelve and policy implications. *Child Development*, 52(2), 538–544. <https://doi.org/10.2307/1129172>
- Pallant, J. (2011). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Open University Press.
- Seligman, L. (1994). *Developmental career counseling and assessment* (2nd ed.). SAGE Publication.
- Super, C. M., & Super, D. E. (2001). *Opportunities in psychology careers*. McGraw-Hill.
- Suryanti, R., Yusuf, M., & Priyatama, A. N. (2011). Hubungan antara Locus of Control internal dan konsep diri dengan kematangan karir pada siswa kelas XI SMK Negeri 2 Surakarta. *Wacana: Jurnal Psikologi*, 3(1). <https://jurnalwacana.psikologi.fk.uns.ac.id/index.php/wacana/article/view/46>
- Walsh, W. B., Bingham, R. P., Brown, M. T., Ward, C. M., & Osipow, S. H. (Eds.). (2000). *Career counseling for African Americans*. Lawrence Erlbaum Associates.

Determinant factors affecting the improvement of education index

Jalil Setiawan Jamal*; Muslim Salam; A. Nixia Tenriawaru; Didi Rukmana; Muhammad Hatta Jamil; Saadah

Universitas Hasanuddin

Jl. Perintis Kemerdekaan Km. 10, Tamalanrea Indah, Tamalanrea, Kota Makassar, Sulawesi Selatan
90245, Indonesia

*Corresponding Author. E-mail: jalilsetiawan357@gmail.com

ARTICLE INFO

ABSTRACT

Article History

Submitted:

17 April 2021

Revised:

24 June 2021

Accepted:

30 June 2021

Keywords

education index; teacher to student ratio; school to student ratio; class to student ratio

Scan Me:



How to cite:

Jamal, J., Salam, M., Tenriawaru, A., Rukmana, D., Jamil, M., & Saadah, S. (2021). Determinant factors affecting the improvement of education index. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 88-96. doi:<https://doi.org/10.21831/pep.v25i1.40160>

The Human Development Index (HDI) of the Selayar Islands Regency experienced an insignificant improvement. The low education index causes the low HDI achievement of the Selayar Islands Regency because the achievement of the education index is lower than the health index and the expenditure index. Therefore, it is essential to improve the education index. This study aims to analyze the factors that influence the education index. This study uses secondary data in panel data, a combination of time-series data from 2014 to 2019, and cross-section data from 11 sub-districts. Panel data to measure the factors that affect the Education Index were analyzed using regression analysis. The results show that the teacher to student ratio at elementary school has a negative effect on the education index, the class to student ratio at elementary school has a positive effect on the education index, while the school to student ratio at elementary school, school to student ratio at junior high school, class to student ratio at junior high school and teacher to student ratio at junior high school do not affect the education index.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

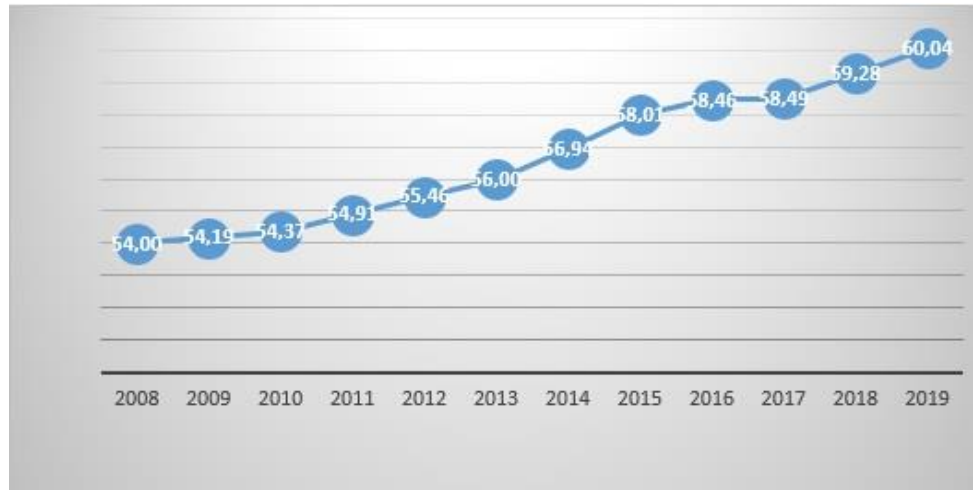


INTRODUCTION

Education, according to Ki Hajar Dewantara, is a way of life that can guide a child in utilizing all natural strengths and potentials that will bring maximum safety and happiness as a human being or as a member of society (Sugiarta et al., 2019). Regional development performance can be measured through the performance of the education sector in the region. Thus, according to Chamadi (Sukarsa, 2012), several indicators to measure the quality of education in an area include (1) the teacher to student ratio, which is the ratio between the number of teachers by the number of students in certain education level; (2) the school to student ratio, which is the ratio between the number of schools and the number of students at a certain level of education; and (3) the class to student ratio is the ratio between the number of classrooms and the number of students at a certain level of education.

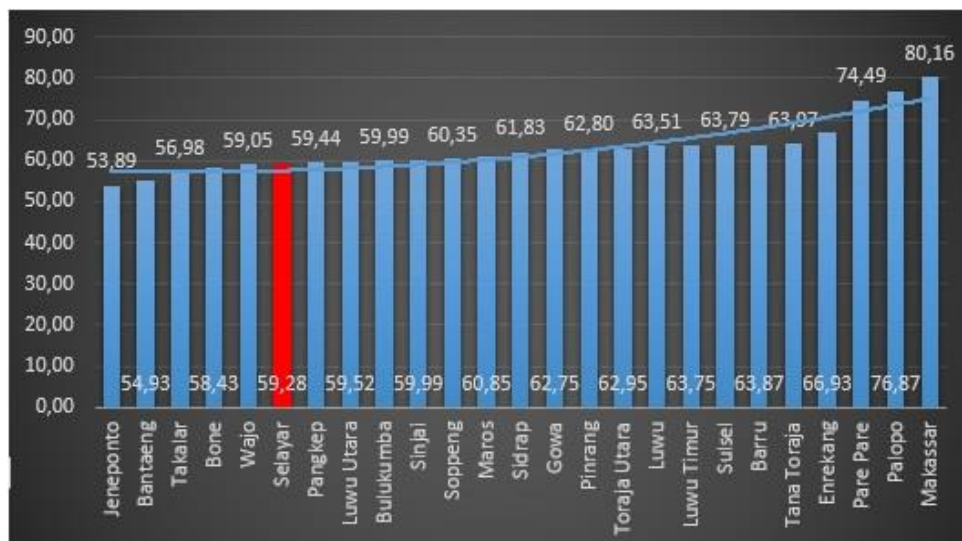
One of the indicators to measure the success of education development is the education index (Mahendra et al., 2016). The Central Bureau of Statistic (*Badan Pusat Statistik* or BPS) explains that Education Index is a combination of two indicators of education in which the average length of schooling is described as the number of years needed by the population in formal education and expectancy of schooling that is described as a school which the child expects at a certain age in the future (Mahendra et al., 2016).

Education Index on Selayar Islands Regency has experienced an insignificant improvement over the past decade. In line with this, the Education Index achievements of Selayar Islands Regency in 2018, when compared to 24 Regencies/Cities in South Sulawesi Province, are in the 19th or sixth lowest rank above Jenepono, Bantaeng, Takalar, Bone, and Wajo Regency. This achievement can be seen in Figure 1 and Figure 2.



Source: Adapted from data of the Central Bureau of Statistics of Selayar Island Regency

Figure 1. The Education Index of Selayar Island Regency



Source: Adapted from data of the Central Bureau of Statistics of South Sulawesi Province

Figure 2. The Education Index of Cities/Regencies in South Sulawesi Province on 2018

The education index is one of three composite indexes of the Human Development Index (HDI) besides the health index and expenditure index (Mirza, 2011). When viewed from the achievement of these three composite indexes, it can be concluded that the achievement of the education index is lower than the achievement of the Health Index and the Expenditure Index, as shown in Table 1. The low achievement of the education index causes the low HDI achievement of the Selayar Islands Regency. It happens because, as one of the composite indexes of the HDI, the Education Index is very influential on the HDI itself. The improvement of the Education Index has a positive and significant correlation with the improvement of HDI (Cahill, 2005). The Education Index also has a significant and positive effect on HDI (Lestari & Sanar, 2018).

Table 1. Results of the Composite HDI Index for Selayar Islands Regency in 2010-2019

Year	Education Index	Health Index	Expenditure Index	HDI
2008	54.00	70.62	60.20	61.23
2009	54.19	70.65	61.83	61.86
2010	54.37	70.67	62.49	62.15
2011	54.91	70.72	62.96	62.53
2012	55.46	70.78	63.31	62.87
2013	56.00	70.82	63.53	63.16
2014	56.94	70.83	63.96	63.66
2015	58.01	71.17	64.46	64.32
2016	58.46	71.27	65.77	64.95
2017	58.49	71.37	66.98	65.39
2018	59.28	71.72	67.75	66.04
2019	60.04	72.23	69.07	66.91

Source: Adapted from data of the Central Bureau of Statistics of Selayar Islands Regency

Based on those explanations, this study aims to analyze the factors that affect the improvement of the education index. To answer the research objectives, it is necessary to formulate a research hypothesis. The formulation of the hypothesis in this study is based on the problems described previously and is based on the study and analysis of several previous studies related to this research. The research hypothesis that is formulated can be explained as follows: "The variables of the teacher to student ratio at elementary school, the teacher to student ratio at junior high school, the school to student ratio at elementary school, the school to student ratio at junior high school, the class to student ratio at elementary school, the class to student ratio at junior high school respectively and or simultaneously have a significant influence to the Education Index improvement in the Selayar Islands Regency".

RESEARCH METHOD

This research employed a quantitative method approach. In order to analyze the factors affecting the Education Index, the research variables are the teacher to student ratio at elementary school (X1), the teacher to student ratio at junior high school (X2), the school to student ratio at elementary school (X3), the school to student ratio at junior high school (X4), the class to student ratio at elementary school (X5), the class to student ratio at junior high school (X6) as the independent variable, and the Education Index as the dependent variable. The type of data required is secondary data in the form of panel data formed by time-series data from 2014 until 2019 with cross-section data in 11 sub-districts.

Regression analysis technique was used to analyze the factors affecting the Education Index using panel data. Panel data regression analysis was started by selecting the best panel data model in the study. Widarjono (Sunarya, 2016) says that the estimation model with panel data uses three approaches: *Common Effect Model* (CEM), *Fixed Effect Model* (FEM), and *Random Effect Model* (REM). The best model of the three approaches was estimated by the Chow test to choose between the CEM and FEM models and the Hausman test to choose between the FEM and REM models. After the panel data model was determined, a regression equation was generated, followed by a hypothesis test in the form of the F statistical test, R² statistical test, and t statistical test. The regression equation is shown in Formula (1), where IP_{it} = Education Index for the i-subdistrict in the t-year, X_{1it} = the teacher to student ratio at elementary school for i-subdistrict in t-year, X_{2it} = the teacher to student ratio at junior high school for i-subdistrict in t-year, X_{3it} = the school to student ratio at elementary school for i-subdistrict in t-year, X_{4it} = the school to student ratio at junior high school for i-subdistrict in t-year, X_{5it} = the class to student ratio at elementary school for i-subdistrict in t-year, X_{6it} = the class to student ratio at junior high school for i-subdistrict in t-year, β_0 = intercept coefficient, $\beta_n; n = 1, 2, \dots, 6$ = regression parameters, and ε_t = error term. In addition, Ajja (2011), Aulia (2004), Gujarati

(2003), Verbeek (2000), and Wibisono (2005) state that in regression analysis with panel data, it is unnecessary to test classical assumptions as an implication of the various advantages possessed by panel data compared to time series data or cross-section data.

$$IP_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \beta_4 X_{4it} + \beta_5 X_{5it} + \beta_6 X_{6it} + \varepsilon_t \dots\dots\dots (1)$$

FINDINGS AND DISCUSSION

The Achievement of Education Sector Development in Selayar Islands Regency

The achievement of the Education Index in a region is determined by its educational development performance. Educational development aims to ensure the availability of education services that include all components related to education sector, including human resources, namely students and teachers, educational infrastructures such as schools, learning infrastructure, and others. Several indicators related to the availability of education services include the teacher to student ratio, the school to student ratio, and the class to student ratio, which are the independent variable in this study. Thus, the education indicators in Selayar Islands Regency, especially the variables in this study, generally have fluctuating achievements from year to year, as illustrated in Table 2.

Table 2. Achievements of Education Indicators for Selayar Islands Regency

Year	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆
2008	0.122	0.122	0.009	0.008	0.053	0.047
2009	0.115	0.177	0.010	0.014	0.060	0.044
2010	0.110	0.150	0.009	0.009	0.049	0.040
2011	0.088	0.152	0.009	0.009	0.051	0.046
2012	0.109	0.116	0.009	0.010	0.049	0.051
2013	0.116	0.121	0.008	0.008	0.049	0.043
2014	0.115	0.112	0.008	0.010	0.048	0.048
2015	0.072	0.105	0.008	0.010	0.047	0.047
2016	0.114	0.133	0.009	0.009	0.052	0.042
2017	0.116	0.131	0.009	0.009	0.055	0.043
2018	0.109	0.115	0.009	0.008	0.057	0.041
2019	0.112	0.082	0.010	0.008	0.060	0.037

Analysis of Factors that Affect the Education Index

Estimation of Panel Data Regression Model

The panel data regression model consists of three types of models: the Common Effect Model (CEM), Fixed Effect Model (FEM), and Random Effect Model (REM). The estimation results of the three models can be seen in Table 3.

Table 3. Estimation Result of CEM, FEM, and REM

No	Variable	Regression Model					
		CEM		FEM		REM	
		Statistics	Prob.	Statistics	Prob.	Statistics	Prob.
1	X ₁	2.046691	0.0452	-2.19797	0.0327	-1.976374	0.0528
2	X ₂	0.409423	0.6837	0.648490	0.5197	0.348200	0.7289
3	X ₃	-0.54278	0.5893	1.856542	0.0694	1.387706	0.1704
4	X ₄	-2.75153	0.0004	-1.42239	0.1612	-2.821329	0.0065
5	X ₅	1.055067	0.2957	3.094697	0.0033	2.994592	0.0040
6	X ₆	1.084531	0.2825	-0.33288	0.7406	0.022577	0.9821
R-squared		0.477352		0.926743		0.462315	
Prob(F-Statistic)		0.000001		0.000000		0.000001	

From Table 3, it can be seen that the CEM estimation produces an R2 value of 0.477 and a prob (F-statistic) value of 0.000001, the FEM estimation produces an R2 value of 0.927 and a prob (F-statistic) value of 0.000000, and the REM estimation model produces an R2 value of 0.462 and a prob (F-statistic) value of 0.000001. Therefore, it can be said that the entire regression model produces two independent variables with a prob (t-statistic) value of less than the 5% significance level.

Determination of the Best Panel Data Regression Model

The best panel data regression model is determined using two tests: the Chow test and the Hausmann test. The Chow and Hausman test results in this study are through the Eviews 10 application, which can be seen in Table 4.

Table 4. The Results of the Chow Test and the Hausman Test of the Panel Data Regression Model

Test	Statistics	Probability
Chow Test	129.685418	0.0000
Hausmann Test	13.893708	0.0308

Based on the results of the Chow test as presented in Table 4, the resulting probability value is 0.0000. It is less than the 5% significance level. Based on the results of the Chow test, it can be said that FEM is a better panel data regression model than the CEM model. Furthermore, in order to compare the FEM and REM, the Hausmann test was performed. Based on Table 3, the probability generated in the Hausman test is less than the 5% significance level. Therefore, it indicates that the FEM model is better than the REM model. From this explanation, it can be concluded that the Fixed Effect Model (FEM) panel data regression model is the best model that can be used to estimate the factors that affect the Education Index in Selayar Islands Regency.

The Result of the F and R² Statistical Test

The F statistical test is used to prove the hypothesis that all independent variables affect the dependent variable simultaneously. Meanwhile, the R2 test is used to determine how much influence all independent variables simultaneously have on the dependent variable. The results of the F test and R2 test in this study can be seen in Table 5.

Table 5. The Results of the F Test and the R² Factors that Affect the Education Index

Model	F-statistic	Prob(F-statistic)	R Square
Fixed Effect Model	38.74223	0.000000	0.926743

Based on Table 5, the statistical F value is 38.742, and the F table value is 2.26. Therefore, the statistical F value is more than the F table. In addition, the F statistical probability value of 0.000 is less than the significance level α (0.05). Thus, the previous research hypothesis is accepted, which means the variable of teacher to student ratio at elementary school, the ratio of teachers to students at junior high school, the ratio of schools to students at elementary school, the ratio of schools to students at junior high school, and the ratio of class to students at elementary school, the ratio of class to students at junior high school affect the Education Index simultaneously.

Furthermore, the R2 value presented in Table 5 is 0.927. This shows that all of the independent variables have an effect on the Education Index simultaneously by 92.7%. In comparison, the remaining 7.3% is influenced by other variables besides the variables contained in this study.

The Results of Panel Data Regression Analysis and Statistical T Test

Panel data regression analysis is used to test the effect of the variables expressed in the form of equations, and the t statistical test is used to prove the hypothesis that all independent variables partially affect the dependent variable. The results of regression analysis with panel data and the t statistical test in this study are presented in Table 6.

Table 6. The Results of Panel Data Regression Analysis and T Statistical Test

Variable	T-statistic	Prob.	Coefficient	Std. Error
C			53.47487	2.204696
X ₁	-2.197974	0.0327	-0.160864	0.073188
X ₂	0.648490	0.5197	0.028767	0.044360
X ₃	1.856542	0.0694	3.112364	1.676431
X ₄	-1.422394	0.1612	-1.215816	0.854767
X ₅	3.094697	0.0033	0.862743	0.278781
X ₆	-0.332877	0.7406	-0.032869	0.098742

After conducting panel data regression analysis in this study, the regression equation is obtained as in Formula (2). In addition, based on the regression equation and model estimation shown in Table 5, it can be seen and explained as follows.

$$IP = 53.475 - 0.160X_1 + 0.029X_2 + 3.112X_3 - 1.216X_4 + 0.863X_5 - 0.033X_6 \dots\dots\dots (2)$$

The variable of the teacher to student ratio at elementary school (X₁) has a coefficient of -0.161. It shows that an increase of one percent of the teachers to students ratio at the elementary school will decrease the Education Index by 0.161 percent, assuming the other variables are constant. The X₁ variable has a t count of 2.198. The value of the t count is more than the t table, which is 2.001, and it is negative. Furthermore, the probability value of this variable is less than the 5% significance level. Therefore, the teacher to student ratio variable at the elementary school has a negative and significant effect on the education index variable.

The variable of the teacher to student ratio at junior high school (X₂) has a coefficient of 0.029. It shows that an increase of one percent of the teachers to students ratio at the junior high school will increase the Education Index by 0.029 percent, assuming the other variables are constant. However, the X₂ variable does not affect the Education Index because the probability value more than the 5% significance level.

The variable of the school to student ratio at elementary school (X₃) has a coefficient of 3.112. It shows that an increase of one percent of the schools to students ratio at the elementary school will increase the Education Index by 3.112 percent, assuming the other variables are constant. Furthermore, the X₃ variable does not affect the Education Index because the probability value is more than the 5% significance level.

The variable of the school to student ratio at junior high school (X₄) has a coefficient of -1.216. It shows that an increase of one percent of the schools to students ratio at the junior high school level will reduce the Education Index by -1.216 percent, assuming the other variables are constant. Furthermore, the X₄ variable does not affect the Education Index because the probability value is more than the 5% significance level.

The variable of the class to student ratio at elementary school (X₅) has a coefficient of 0.863. It shows that an increase of one percent of the class to students ratio at the elementary school level will reduce the Education Index by 0.863 percent, assuming the other variables are constant. Furthermore, The X₅ has a t count of 3.094. The value of the t count is more than the t table, which is equal to 2.001, and it is positive. Furthermore, the probability value of this variable is less than the 5% significance level. Therefore, it can be concluded that the Class to Student Ratio variable at the elementary school has a positive and significant effect on the Education Index variable.

The variable of the class to student ratio at the junior high school (X6) has a coefficient of -0.033. This shows that an increase of one percent of the class to students ratio at junior high school level will reduce the Education Index by -0.033 percent, assuming the other variables are constant. However, the X6 variable does not affect the Education Index because the probability value more than the 5% significance level.

Factors that Affect the Education Index Improvement

Based on the previous explanation and analysis, there are some factors that influence the Education Index that contribute to the HDI improvement. Each factor is elaborated as follows.

The Ratio of Teacher to Student at Elementary School

Based on the previous analysis, the teacher to student ratio at the elementary school has a negative and significant effect on the Education Index. It means that increasing the ratio of teachers to students at elementary school will decrease the Education Index in Selayar Islands Regency.

The results of this study are not in accordance with the results of the previous study of [Sapaat et al. \(2020\)](#), which state that the ratio of teachers to students has a positive effect on HDI in North Sulawesi Province, and the result of a study by [Mahendra et al. \(2016\)](#), that the ratio of teachers to students at the elementary school has a positive and significant effect on the Education Index in East Java Province. This situation is caused by the number of teachers at the elementary school is overload compared to the number of students in Selayar Islands Regency.

As it is well known, the ratio of teachers to students at elementary school in Selayar Islands Regency during the period 2008 to 2019 exceeds the National Education Standard that has been established by the Ministry of Education and Culture. In 2019, the ratio of teachers to students at elementary school in Selayar Islands Regency reached 1:9, which means that one teacher supports nine students. It does not meet the standard that is established by the Ministry of Education and Culture, in which it is regulated that each elementary school has one teacher supporting 32 students. Based on the data obtained from the Ministry of Education and Culture, in 2019, Selayar Islands Regency has an excess of 289 teachers at the elementary school.

The problem of the excessive ratio of teachers to students at elementary school is inversely proportional to the problem that is experienced by the Selayar Islands Regency for several years, namely the shortage of civil servant teachers, especially in the island region. It is because the Selayar Islands Regency has an excess of non-civil servant teachers. This excess number of the non-civil servant teachers, followed by the lack of civil servant teachers, will certainly contribute to the less than optimal quality of the teachers. As it is commonly known, the quality and also capability of the civil servant teachers will be higher when compared to non-civil servant teachers. In addition, the large number of non-civil servant teachers will cause civil servant teachers not to fulfill their duties optimally, and they sometimes delegate their responsibilities to non-civil servant teachers. It certainly will affect the decline in the quality of education, which will lead to the failure in the achievement of the Education Index improvement.

Therefore, based on the results of this study, in order to improve the education index, which will contribute to HDI improvement, the teacher to student ratio at elementary school must be reduced by reducing the number of non-civil servant teachers in elementary school. In addition, it also will streamline and make effective use of the budget for the payment of the non-civil servant teachers' salaries.

The Ratio of Class to Student at Elementary School

The class to student ratio variable at elementary school is related to the availability of elementary school classrooms in a region compared to the number of elementary school students in that region. Based on the previous analysis, the class to student ratio at elementary school has a positive and significant effect on the Education Index. This means that the improvement of the class to student ratio at elementary school is directly proportional to the improvement of the Education Index in Selayar Islands Regency. The results of this study are in line with the result of the study by Cahyadi (Syamsuri & Bandiyono, 2018), which states that the ratio of the number of students at elementary school to the number of classrooms at elementary school has a positive and significant effect on HDI.

Selayar Islands Regency has launched a compulsory education program for at least nine years, in which there are activities such as building new classrooms and rehabilitating classrooms to support the learning process. Compulsory education of at least nine years will be realized if only educational facilities such as classrooms as a place for learning and interaction between teachers and students are available and in good quality. In other words, the role of the classroom is crucial in the educational process because the availability of a good and comfortable classroom will affect the continuity of the educational process (Hambali, 2016).

When the view from the number of classrooms at elementary school is in good condition, the performance increased from 160 classrooms in 2015 to 745 classrooms in 2019. The availability of educational facilities and infrastructure, especially adequate classrooms, will motivate the community to take advantage of these educational facilities and participate in the learning process (Syamsuri & Bandiyono, 2018). The increase of community participation in the education process in schools is very influential in improving the average length of schooling and the number of years of schooling, which are components of the calculation of the Education Index (Syamsuri & Bandiyono, 2018).

CONCLUSION

From this research, it can be concluded that the factors that affect the education index improvement are as follows. The ratio of teachers to students at the elementary school has a negative and significant effect on the education index improvement. Besides, the class ratio to students at the elementary school has a positive and significant effect on the education index improvement in the Selayar Islands Regency.

REFERENCES

- Ajija, S. R. (2011). *Cara cerdas menguasai EVIEWS*. Salemba Empat.
- Aulia, T. (2004). *Modul pelatihan Ekonometrika*. Fakultas Ekonomi dan Bisnis Universitas Airlangga.
- Cahill, M. B. (2005). Is the human development index redundant? *Eastern Economic Journal*, 31(1), 1–5. <http://www.jstor.org/stable/40326318>
- Gujarati, D. (2003). *Ekonometri dasar* (S. Zain (trans.)). Erlangga.
- Hambali, H. (2016). Pembangunan gedung sekolah dan ruang kelas baru di Kabupaten Seluma pasca pemekaran. *Manajer Pendidikan*, 10(1), 20–28. <https://ejournal.unib.ac.id/index.php/manajerpendidikan/article/view/1229>
- Lestari, W. W., & Sanar, V. E. (2018). Analysis indicator of factors affecting human development index (IPM). *Geosfera Indonesia*, 2(1), 11–18. <https://doi.org/10.19184/geosi.v2i1.7333>

- Mahendra, R., Fariyanti, A., & Falatehan, A. F. (2016). Strategi peningkatan indeks pendidikan melalui alokasi belanja pemerintah daerah bidang pendidikan di Provinsi Jawa Timur. *Jurnal Manajemen Pembangunan Daerah*, 8(2), 1–19. https://doi.org/10.29244/jurnal_mpd.v8i2.24823
- Mirza, D. S. (2011). Pengaruh kemiskinan, pertumbuhan ekonomi, dan belanja modal terhadap IPM Jawa Tengah. *JEJAK: Jurnal Ekonomi Dan Kebijakan*, 4(2), 102–113. <https://journal.unnes.ac.id/nju/index.php/jejak/article/view/4645>
- Sapaat, T. M., Lopian, A. L. C. P., & Tumangkeng, S. Y. L. (2020). Analisis faktor-faktor yang mempengaruhi indeks pembangunan manusia di Provinsi Sulawesi Utara tahun (2005-2019). *Jurnal Berkala Ilmiah Efisiensi*, 20(03), 45–56. <https://ejournal.unsrat.ac.id/index.php/jbie/article/view/30641>
- Sugiarta, I. M., Mardana, I. B. P., Adiarta, A., & Artanayasa, W. (2019). Filsafat pendidikan Ki Hajar Dewantara (Tokoh timur). *Jurnal Filsafat Indonesia*, 2(3), 124–136. <https://doi.org/10.23887/jfi.v2i3.22187>
- Sukarsa, I. M. (2012). Pemetaan kualitas pendidikan di Propinsi Bali berbasis spasial. *Majalah Ilmiah Teknologi Elektro*, 8(1), 6–11. <https://ojs.unud.ac.id/index.php/JTE/article/view/1570>
- Sunarya, I. W. (2016). Analisis pembangunan sumber daya manusia di Provinsi Bali tahun 2011-2014. *Jurnal Aplikasi Manajemen*, 14(3), 577–584. <https://doi.org/10.18202/jam23026332.14.3.18>
- Syamsuri, M. R., & Bandiyono, A. (2018). Pengaruh belanja pemerintah daerah berdasarkan fungsi terhadap peningkatan IPM dan pengentasan kemiskinan (Studi pada kabupaten/kota di Provinsi Aceh). *Info Artha*, 2(1), 11–28.
- Verbeek, M. (2000). *A guide to Modern Econometrics*. John Willey & Sons.
- Wibisono, D. (2005). *Metode penelitian & analisis data*. Salemba Medika.

Applying Item Response Theory model for evaluating item and test properties of academic potential test for students with disability

Sukaesi Marianti*; Dian Putri Permatasari; Unita Werdi Rahajeng

Universitas Brawijaya

Jl. Veteran, Ketawanggede, Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia

*Corresponding Author. E-mail: s.marianti@ub.ac.id

ARTICLE INFO

Article History

Submitted:

16 February 2021

Revised:

28 June 2021

Accepted:

14 July 2021

Keywords

admission selection;
disability; computer-based
academic potential test;
item response theory

Scan Me:



How to cite:

Marianti, S., Permatasari, D., & Rahajeng, U. (2021). Applying Item Response Theory model for evaluating item and test properties of academic potential test for students with disability. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 97-107. doi:<https://doi.org/10.21831/pep.v25i1.38808>

ABSTRACT

Universitas Brawijaya (UB) is one of the pioneers of inclusive education in higher education in Indonesia. One of the innovations in the policies related to inclusive education is affirmative action admissions special for students with disabilities, namely *Seleksi Mandiri Penyandang Disabilitas* (Independent Selection for Person with Disabilities), which focuses on accommodating admissions selection for students with disabilities who want to enroll in bachelors or vocational programs. A part of this admission selection is the test called the Computer-Based Academic Potential Test. This study aims to evaluate, from a psychometric perspective, the psychometric properties of the potential academic test. The approach used in this study is the item response theory (IRT) framework, which is mostly used for evaluating psychometric quality at both item-level and test levels. This study's IRT model is a two-parameter logistic model that includes difficulty parameter and discrimination parameter. The result of this study exhibited that the three subtests of the Computer-Based Academic Potential Test, in general, have satisfying results from the 2PL model estimation. The result also showed that most of the item difficulties ranged from medium to very difficult.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



INTRODUCTION

Since 2012, Universitas Brawijaya (UB) has developed a special admission system for students with disabilities: *Sistem Penerimaan Khusus Penyandang Disabilitas* or SPKPD (Special Admission System for Person with Disabilities). The system is part of affirmative action to address a problem where the education level attainment of people with disabilities is still much lower than those without disabilities. Initially, the system provided for people with disabilities to enter the university was still limited. Even if they want to participate in national selection into public universities, they must face less adaptive and less accommodating selection models for people with special needs. Palombi (2000) states that standardized test models often do not consider the special needs of people with disabilities, so there will be unfairness if regular test scores are used for the decision making in the admission of students with disabilities.

SPKPD is designed as a system that considers the interests of people with disabilities, which will then be adjusted to the departments and programs in UB. The SPKPD implementation is fully submitted to *Pusat Studi dan Layanan Disabilitas* (PSLD), a service center for students with disabilities in UB. Initially, PSLD developed SPKPD using only interview and observation methods. Administratively-qualified prospective students with disabilities are invited to participate in interviews and observations in several activity settings (Pratiwi et al., 2018).

Since 2018, SPKPD has been modified by adding a Computer-Based Academic Potential Test. Some of the underlying considerations are (1) advice from program managers related to general standards that prospective students with disabilities must own, (2) the increasing number of prospective students with disabilities participating in the SPKPD makes the implementation of interviews or observations inefficient and impractical. Besides, the Computer-Based Academic Potential Test results are used as the basis for selecting prospective students to be interviewed by the manager of the selected program.

In 2019, UB Rector Regulation No. 33 of 2019 changed the term of SPKPD to SMPD, which stands for *Seleksi Mandiri Penyandang Disabilitas* (Independent Admission for Person with Disabilities). In general, the admission selection process does not change from the previous selection system, which aims to select and obtain information about the academic potential of prospective students. The underlying reason in the selection system is to provide opportunities for prospective students with disabilities by going through the selection process. Although prospective students with disabilities are not given the same selection test as regular prospective students, ideally, the quality of tests for both groups is psychometrically acceptable.

The computer-Based Academic Potential Test, as a part of the admission selection, is expected to represent the academic potential of prospective students with disabilities, characterized by mastery of basic academic abilities, including language and numeric. Therefore, the Computer-Based Academic Potential Test sub-test consists of Bahasa Indonesia, English, and Mathematics. A selection system involving a Computer-Based Academic Potential Test that is psychometrically feasible is one way to ensure the quality and readiness of prospective students with disabilities to study in college. Basically, to study in college, one must have a minimum requirement. Without the selection process, the university does not have enough information to know which students are ready and not academically ready to attend college education. In addition, if there are academically not ready students, it will be difficult to attend courses in college because they do not have the adequate basic academic ability.

Wolanin and Steele (2004) explained that in terms of admission of students with disabilities, each course must still consider the minimum academic requirements of prospective students. In general, without a well-designed selection model, prospective students with disabilities in universities are vulnerable to getting caught up in the charity model paradigm. In the charity model paradigm, prospective students with disabilities are the parties entitled to mercy (Rukmantara & Lesmana, 2018). Of course, the spirit is not in line with UB policy that opens the opportunity to study in universities for prospective students with disabilities as a form of social reconstruction and fulfillment of human rights equality.

A good quality test is a test that has good psychometric characteristics through a series of psychometric analyses to obtain evidence that it is feasible to use. One indication of psychometric feasibility is that the items function accurately and fairly to all test takers. A test that psychometrically functions optimally is a test that produces a score that truly represents the test taker's ability so the scores obtained from the test results can be used for decision making.

In order to create a well-design test, psychometric evaluation is inevitable. Evaluation of the psychometric characteristics of a test involves psychometric analysis to prove that a test is not very easy and not too difficult and can distinguish participants with high and low abilities. A very popular approach used to evaluate psychometric characteristics is IRT, also known as latent traits theory or modern theory. The advantage of IRT is this theory's ability to describe the relationship between the ability, difficulty of the item, and the probability of answering correctly on a particular item (Zoghi & Valipour, 2014).

Item Characteristic Curve (ICC)

An ICC is a curve used to describe the relationship between the ability or characteristics of a test taker, the characteristics of an item, and the probability of answering correctly on the

item. There are item parameters used in ICC, namely, item difficulty (b), item discrimination (a), and guessing (c) (Schmidt & Embretson, 2012). The number of parameters used as fixed parameters depends on the selected model. This research uses the 2PL model, which involves two parameters, a and b .

Information Functions

The information function is intended to demonstrate the ability of an item and/or a test in providing precise information at a certain level of ability (θ). A high information value represents higher precision in providing information about test takers at a certain level of ability. IRT has information functions at the item level (item information) and the test level (test information) (Baker, 2001; Hambleton et al., 1991).

Two Parameter Logistic Model (2PL Model)

Birnbaum developed the 2PL model in 1968, where the logistics of the model were easier to work with than normal. The probability of the test taker answering correctly on an item based on the 2PL model is written on Formula (1), where $P_i(\theta)$ is the probability is the probability of test taker at a certain level of ability to answer the question correctly, θ is the ability of the test taker, b is the difficulty level of the item, a is the discrimination power of the item, and e is the constant value, 2.718.

$$P_i(\theta) = \left[\frac{1}{1+e^{-a(\theta-b)}} \right] \dots\dots\dots (1)$$

The use of the 2PL model in this study is based on the comprehensiveness of the 2PL model compared to the 1PL model since the 2PL model includes item discrimination parameters. Compared to the 3PL model, the 2PL model has fewer parameters. However, in the calibration process, the 2PL model is easier to achieve convergence. In the 3PL model, the difficulty of achieving convergence often occurs because the scale of the guessing parameter is different from the other two parameters.

Therefore, this study has several important points, including (1) evaluating the psychometric characteristics of the Computer-Based Academic Potential Test used for the admission selection for prospective students with disabilities during the period 2018 to 2019, (2) evaluating the characteristics of items and the amount of information based on the IRT framework, and (3) evaluating the characteristics of the test and the amount of information that the test can provide. Furthermore, important findings in this study can be useful to obtain a scientific basis in deciding whether it is necessary to reconstruct new assessments in the future, as a basis for deciding whether the test can be used to determine the score of prospective students.

RESEARCH METHOD

This research is quantitative psychometric research that aims to evaluate the psychometric characteristics of the Computer-Based Academic Potential Test. This research was conducted in four stages: test review, data collection, data analysis, and interpretation and decision making, as described in Figure 1.

The first stage was a test review. This stage involved studying the test equipment used to select prospective students with disabilities in-depth, such as examining the basis of the theory used and the construction study used, considering the number of dimensions or structural factors. It also involved the study of test quality evaluation techniques that have been done and studying the techniques of estimating the test taker's scores by considering the type of construction and psychometric quality.

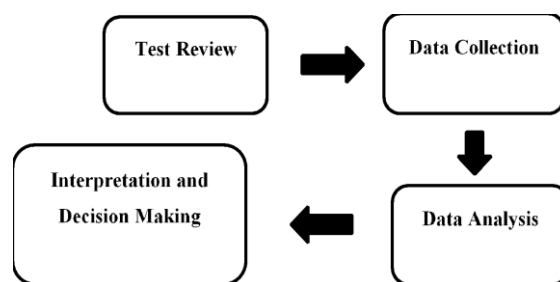


Figure 1. Four-stage Flowchart of Research in Psychometric Analysis

The second stage was data collection. It was done by collecting all the test taker's answers for the last five years (the test taker's identity was confidential). The data was only an answer response for all items in the test. The response had been coded into quantitative data, i.e., scores of 1 and 0, based on actual or false answers.

After collecting quantitative data, the next stage was analyzing the data based on the type of the previously-described construct. The analysis technique used is Item Response Theory (IRT) for the type of dichotomy response. The IRT model used is a logistic model with two parameters (2PL). The 2 PL Model is useful for estimating test items' characteristics based on item difficulty parameters (b) and item discrimination (a). The researchers also estimated the function of item information and tests to find out how informative an item and test in providing information about the test taker's ability.

The fourth stage, as the last stage, was interpretation and decision-making based on the results of data analysis, which will then lead to a decision on whether an item is feasible for university entrance selection for UB prospective students with disabilities. After the items were examined psychometrically, the interpretation was carried out at the test level. If most of the items are of good quality, then a group of such items can be concluded as a set of psychometrically qualified tests for use.

This research was conducted in *Pusat Studi Layanan Disabilitas* (PSLD) of Universitas Brawijaya, using test result data for the last two years. Samples in this study are 60 prospective students with disabilities who participated in UB entrance selection in the range of 2018 to 2019. All personal identities of the study sample were not used as research data, so that the test taker's identity was not listed in the research report.

The Computer-Based Academic Potential Test is a selection test for UB students with disabilities. The test is administered in groups and with a limited time of 90 minutes to work on the whole question. This test aims to measure academic capability consisting of three subtests: Bahasa Indonesia, English, and Mathematics. Each subtest consists of 15, ten, and five items, respectively. Each item is given a score of 1 or 0, based on a true or false answer.

FINDINGS AND DISCUSSION

A crucial IRT assumption test was conducted before analyzing the data using the item response theory (IRT) technique, known as unidimensionality. Unidimensionality testing aims to determine whether a subtest measures only one latent trait (Zanon et al., 2016).

In this study, a dimensionality test was conducted using the confirmatory factor analysis (CFA) technique. The results of the CFA can be seen in Table 1. Each subtest is modeled as unidimensional. Table 1 shows that all three subtests have a unidimensional model that fits the data. This conclusion is based on the cut-off point. The CFI fit index is said to be a reasonable fit with a minimum value of 0.90 (Wang & Wang, 2019). A value of SRMR less than 0.08 is considered a good fit (Hu & Bentler, 1999), and it is acceptable when the value is less than 0.10 (Kline, 2016). A value of RMSEA is said to be a fair fit if it falls between 0.05-0.08 and is said to be a close fit if it is less than 0.05 (Byrne, 1998).

Table 1. Fit Indices of the Three Subtests

Subtest	Index		
	CFI	SRMR	RMSEA
Bahasa Indonesia	0.921	0.088	0.037
English	0.987	0.065	0.031
Mathematics	0.922	0.070	0.063

After the assumption test is completed, further analysis is performed using the item response theory (IRT) technique. Based on data analysis, information about item characteristics and test characteristics are obtained. Item characteristics are represented by discrimination power, item difficulty level, item characteristic curve (ICC), item information function (IIF). Test characteristics are indicated by the test characteristic curve (TCC) and test information function (TIF). Both the item characteristics and the test characteristics are very important information as the basis for deciding whether the test used for selection is a psychometrically feasible test.

Item Characteristics

The item characteristics were estimated for each sub-test, and the results of the estimation of all three sub-tests are shown in Table 2. The analysis results show that most items from all sub-tests tend to have moderate difficulty levels, and some items have difficulty levels in the range of difficult to very difficult, such as item 13 in the Bahasa Indonesia sub-test and item 5 in the Mathematics sub-test. The item discriminations fall in the range of 0.301-2.267, which is low-very high (Baker, 2001). This shows quite good results, especially since there is no negative discriminatory power.

Table 2. Item Parameter Estimation of the Three Subtests

Item	Item Parameter		Item	Item Parameter	
	α	β		α	β
Bahasa Indonesia	1	1.170	English	1	1.193
	2	0.502		2	1.349
	3	0.876		3	0.665
	4	0.301		4	2.267
	5	0.768		5	1.080
	6	1.570		6	2.165
	7	0.489		7	0.774
	8	1.901		8	0.634
	9	1.819		9	1.663
	10	1.053		10	0.764
Mathematics	11	1.495	Mathematics	1	0.751
	12	0.437		2	0.921
	13	0.523		3	1.089
	14	0.783		4	0.820
	15	0.453		5	0.471

The parameters in Table 2 are more clearly illustrated in the ICC, as shown in Figure 2. The ICC illustrates how item parameters and person parameters interact with each other in a single frame. Based on the ICC on each sub-test, most items show a fairly sharp shape resembling the letter S and no ICC that has reversed direction. This is in line with Table 2, which indicates the item parameters tend to be good, and there is no negative item discrimination (Wu, 2017). However, some items look flatter, that is, item 13 in the Bahasa Indonesia sub-test. The ICC form of item 13 is flat. The item parameter ($\alpha=0.523$, $\beta=4.071$) indicates that the item cannot distinguish variations in test taker's ability at a low-moderate level of ability, but rather can optimally distinguish variations in test taker's ability at a very high level of ability.

After ICC is obtained, the Item information function (IIF) is obtained. Shown in Figure 3, IIF generally affirms the quality of items illustrated by ICC (Figure 2). Items that have a good ICC shape (with a sharp S shape) tend to show a high IIF curve. Items that indicate the peak of a high curve spread at a moderate-high range. Some items have a flat ICC followed by a flat IIF curve, that is, item 13 in the Bahasa Indonesia sub-test has a very flat curve because basically, the item is too difficult so almost no test taker can answer correctly. Besides, Figure 3 also shows that no item has an IIF with a peak that is at a low level of ability. This indicates that all three sub-tests did not have items that functioned optimally at the low ability level.

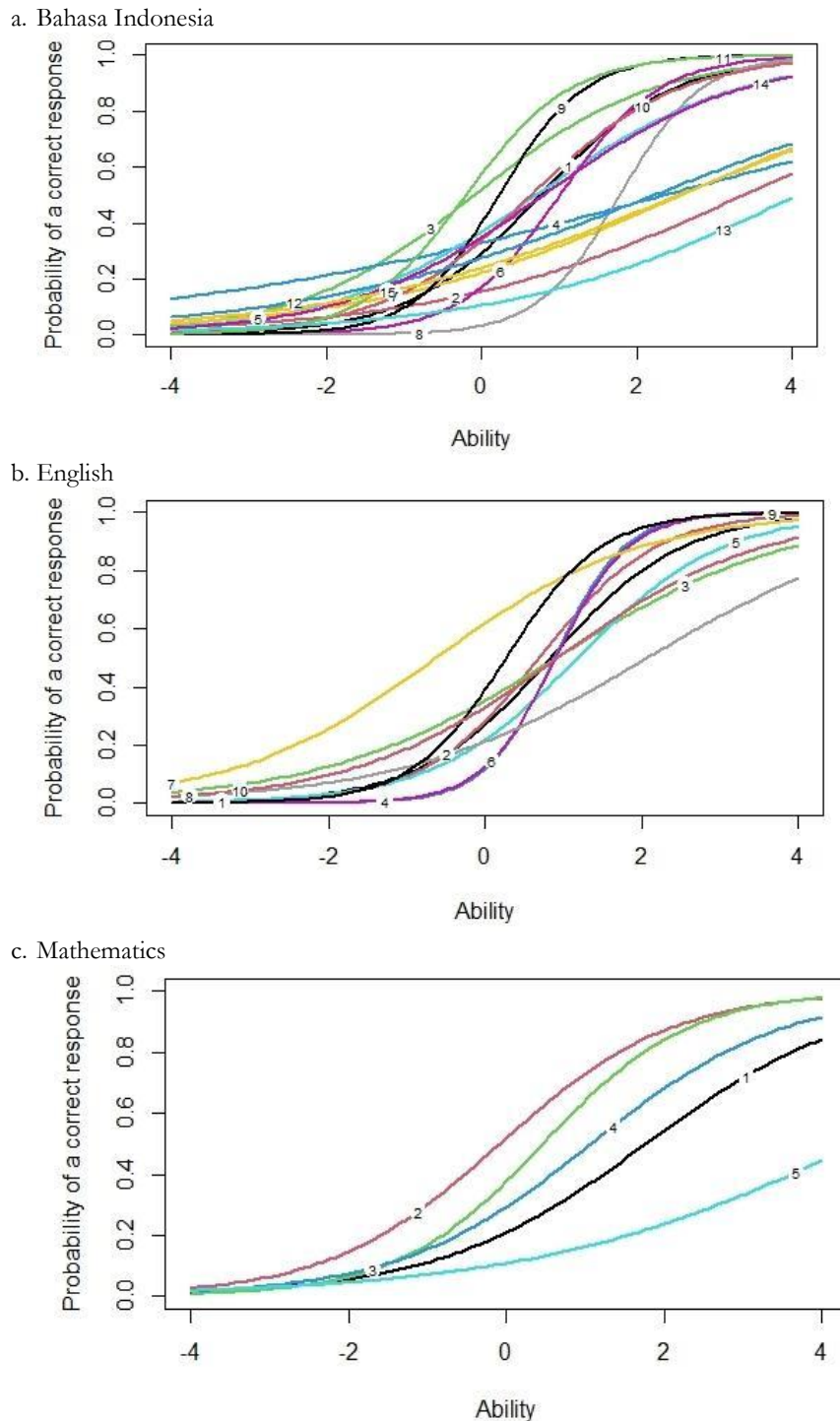
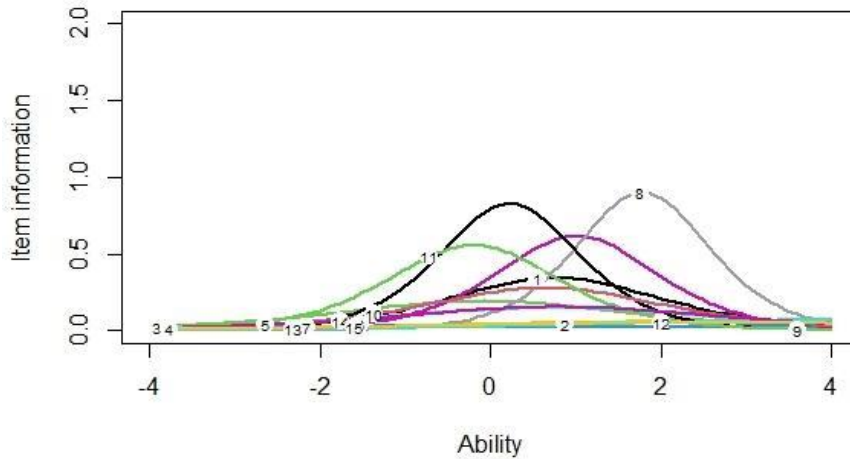
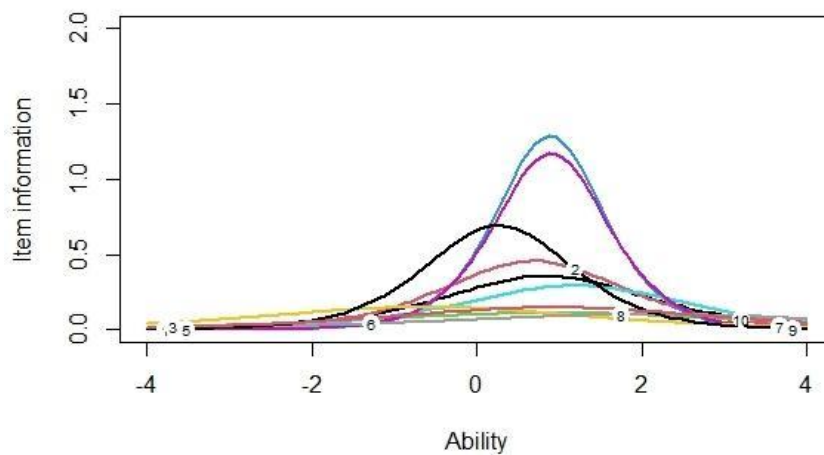


Figure 2. Item Characteristic Curve (ICC) of the Three Subtests

a. Bahasa Indonesia



b. English



c. Mathematics

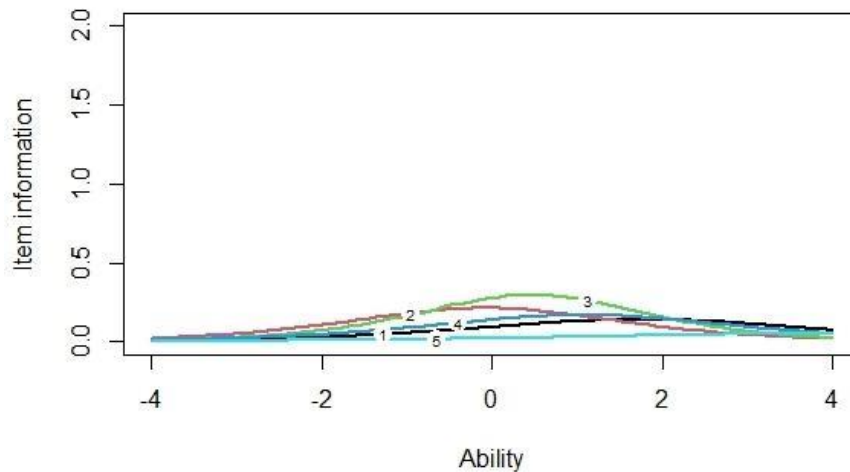


Figure 3. Item Information Function (IIF) of the Three Subtests

Test Characteristics

Beside the item's characteristics, it is also necessary to provide test characteristics to present the psychometric qualities of a set of items. Some important test characteristics are TCC and TIF. Both characteristics are obtained from the accumulation of ICCs and IIFs. TCC and TIF are obtained for each subtest (Bahasa Indonesia, English, and Mathematics) in this study. TCC for all three sub-tests is shown in Figure 4, and TIF for the three sub-tests is in Figure 5.

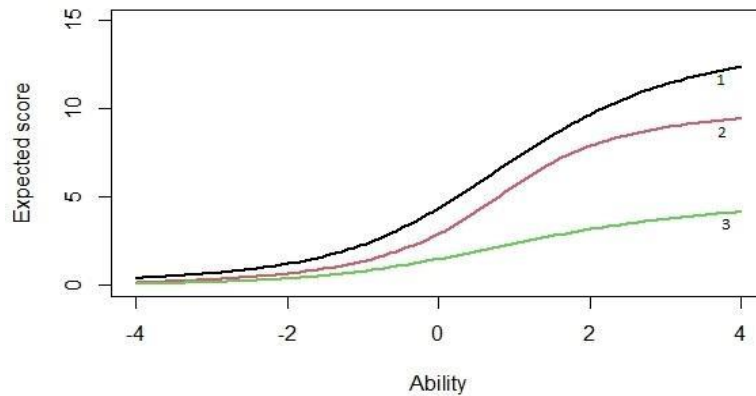


Figure 4. Test Characteristic Curve (TCC) of the Three Subtests. 1 = Bahasa Indonesia, 2 = English, 3 = Mathematics

Figure 4, where the Y-axis is the expected score/true score, and the X-axis is the ability, illustrates the relationship between true score and ability. In practical situations, TCC has an important role in transforming ability into a true score. This makes it very easy for test participants who are not familiar with scaling used in IRT (Baker & Kim, 2017). Figure 4 shows that Bahasa Indonesia is a subtest that has the best psychometric quality among the three sub-tests, while the Math sub-test tends to have a flat curve compared to the other two sub-tests.

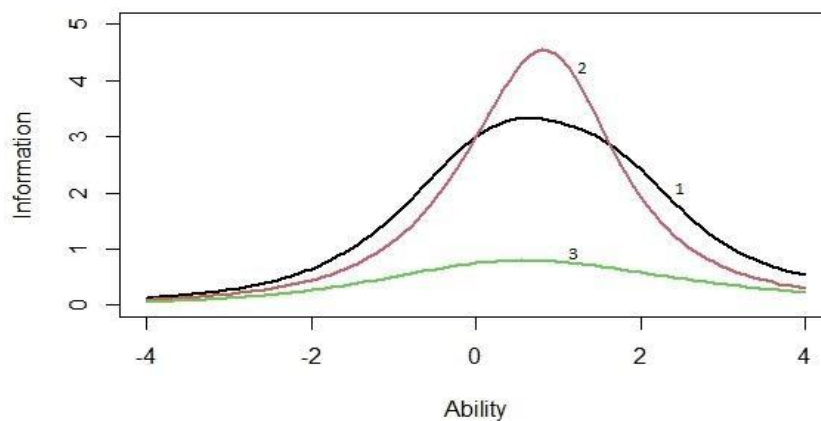


Figure 5. Test Information Function (TIF) from Three Subtests. 1 = Bahasa Indonesia, 2 = English, 3 = Mathematics

Along with IIF, TIF demonstrates the information function at the test level. Figure 5 shows all three sub-tests have the maximum information function at the moderate level, which tends to the high level of ability. This is in line with the IIF showing that the three sub-tests do not have items that can provide maximum information at low ability levels. Besides, the Mathematics subtest has the lowest maximum score (curve peak) compared to other sub-tests.

This research was done to analyze the feasibility of the Computer-Based Academic Potential Test used in admission selection for students with disabilities from a psychometric perspective, both at the item level and test level. IRT is used in this research because it has been proven that IRT has a strong performance in showing the quality of items and tests. Based on the IRT framework, item characteristics and personal characteristics are estimated separately (invariant property of IRT). Moreover, the information function at the item level (IIF) and the test level (TIF) is used to illustrate the test results' precision (An & Yung, 2014; Fan, 1998).

The results of the data analysis showed that all three sub-tests had a fairly good test quality shown from TCC and TIF. According to Baker and Kim (2017), TCC and TIF are the

accumulation of ICC and IIF. It is said to be satisfactory when TCC exhibits a curve that is not flat or a sharp S shape. TIF is informative when it has high information, followed by a low standard error. In this study, both TCC and TIF demonstrated that the Bahasa Indonesia sub-test performed the best, followed by English and Mathematics.

Basically, TIF is intended to show how precise a test estimates ability. The more items in a test, the more informative a test is in describing the characteristics of the test taker, and it is characterized by the high maximum value of information (curve peak) (Baker & Kim, 2017). That is why in Figure 5, it appears that the Mathematics sub-test has the flattest curve compared to the other sub-tests. This is because the number of items in the Math sub-test is the least. In practical situations, TIF is generally more attractive than IIF. For example, in selecting prospective students with disabilities, ideally, the cut-off score of ability is determined in advance to determine the candidates who are accepted or who are not accepted as students. In this case, the curve peak of the TIF of the selection test should be at the cut-off score.

For more details, in each sub-test, items in the Bahasa Indonesia sub-test tend to exhibit the difficulty level in the medium to the difficult range. However, two items (2 and 13) are too difficult with β values that are extremely high (3.374 and 4.071), as shown in Table 2, so are the items in the Math sub-test, which are at moderate to difficult levels, and there is one very difficult item, that is, item 5. Unlike the others, the items in the English sub-test exhibit moderate difficulty levels and no extreme values.

The good news from the results of this study is that no ICC reverses direction (reversed S), since there are no negative discrimination parameters with the range between low to very high. Along with the item difficulty, the item discrimination of all three sub-tests also exhibited pretty good results. However, test developers need to be more aware of some items regarding the difficulty level of the item and the item discrimination. The weak discrimination power of items will affect the ICC's slope to be flatter (Zanon et al., 2016). This indicates that the item cannot detect differences among levels of ability. It could be because the item is ambiguous, the item is too difficult, or the item is too easy. When the item is too difficult, then almost all test takers will answer the test incorrectly. If the incorrect answer is coded 0, almost all item responses are 0, and the variation becomes very small. Likewise, most test-takers will respond to the item correctly on items that are too easy. If most of the test takers gave the same response, then the item will not be able to distinguish different levels of ability.

The aforementioned explanation also shows that there is a relationship between item discrimination and item difficulty. A study conducted by Sim and Rasiyah (2006) found a non-linear relationship between the power of discrimination and the item's difficulty level. The curve shape depicting the relationship is curved downwards (vault), which means that the item discrimination is low when the item difficulty level is low, and increases as the item difficulty level increases, but it begins to go lower when the item difficulty is too difficult. This can be seen in this study (Table 2), which has an extremely high item difficulty value, followed by lower discrimination.

The problem to note is that there are no easy items in the three sub-tests that can function optimally at low levels of ability. Ideally, in a test, it is necessary to have an IIF peak that spreads from low to high ability levels. Thus, it is very necessary to have items that function well and are informative in providing information about the ability of test-takers at all levels of abilities. The Computer-Based Academic Potential Tests used in SMPD cannot provide reliable information for test-takers with a low level of ability, so that the test still needs to be improved, taking into account the informative items which spread evenly at all levels of ability. In short, the test needs to have items of all difficulty levels from easy to difficult.

In developing measuring instruments, things that can be an obstacle for people with disabilities need to be considered. For example, as for the blind, problems in the form of images or maps will be difficult to interpret by screen reader applications. Therefore, if the blind students get problems with many pictures, then he/she will not accurately capture the informa-

tion about it. In general, people with hearing impairment/deaf are vulnerable to language comprehension deprivation, considering that deaf education in Indonesia has not sided with the natural language learning model of the deaf. This leads to a relatively slower and more profound knowledge gain, similarly for people with intellectual disabilities who can process information slowly. For both types of disabilities, special norms are needed to be developed.

Administrative challenges are another obstacle that occurs when administering a test to persons with disabilities, since people with disabilities have different characteristics and special needs. To be fair, it is necessary to note the administration of tests that accommodate people with disabilities' specific needs, for examples, they are elaborated as follows. (1) For the blind test takers, this computer-based test should be readable by the application layer reader precisely following the writing pronunciation. In some cases, screen readers often do not have an accurate ability to read Indonesian text. It is better to be concerned about the time provided for the test takers with disabilities, mostly blind test takers. (2) For the physically impaired, using the computer used for the test's work should not prevent them from working using their limbs, especially for students who have difficulty moving their hands and upper limbs. There need to be special adjustments for people with disabilities who have a less adaptive physical posture with the size of most computer and keyboard products (e.g., dwarfs). (3) For the deaf, there is still a need for a Sign Language Interpreter to explain the procedure of conducting the test at once to facilitate if the test taker has difficulty doing the test.

These specific needs also implement tests that should be done in different rooms for people with disabilities. For example, deaf participants cannot be placed in a room with mental disabilities such as autism or ADHD, because the deaf will do a lot of movement as a form of communication for those who will become a significant distractor for autism or ADHD.

In the context of this study, the limited number of participants in each testing period was also an obstacle in the development of measuring instruments, given that large and representative samples are indispensable in the analysis of psychometric characteristics. Therefore, it is necessary to do continuous data banks so that later psychometric analysis for accommodating measuring instruments for people with disabilities can be better. Furthermore, the items in the subtests have been administered only to the prospective students with disabilities, so there is no evidence that the quality of these items is different between the two populations, which are prospective students with disabilities and prospective students without disabilities. In the future, it would be more interesting if the subtest were administered to the two populations so that the item characteristics of the two populations can be compared.

CONCLUSION

The study found that computer-based tests used in SMPD UB have three sub-tests containing items with satisfying performance. However, some items are still too difficult for the test takers. Besides, the three sub-tests also do not have easy items, so it is very difficult to get information about a test taker with low ability levels. In general, the test has good items with moderate to difficult difficulty levels. It is very effective for measuring the ability of test-takers ranging from moderate to high level.

REFERENCES

- An, X., & Yung, Y.-F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS364*, 1–14. <https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf>
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

- Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Springer.
- Byrne, B. M. (1998). *Structural equation modeling with Lisrel, Prelis, and Simplis*. Psychology Press. <https://doi.org/10.4324/9780203774762>
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357–381. <https://doi.org/10.1177/0013164498058003001>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (Volume 2)*. SAGE Publications.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.
- Palombi, B. J. (2000). Recruitment and admission of students with disabilities. *New Directions for Student Services*, 2000(91), 31–39. <https://doi.org/10.1002/ss.9103>
- Pratiwi, A., Lintang Sari, A. P., Rizky, U. F., & Rahajeng, U. W. (2018). *Disabilitas dan pendidikan inklusif di perguruan tinggi*. Universitas Brawijaya Press.
- Rukmantara, A., & Lesmana, B. (2018). Inclusive education and SDGs: Snapshots from the field. In M. Anwar (Ed.), *International Conference on Sustainability Development Goals for Disabilities (ICSDDGD)* (pp. 1–17). Asosiasi Profesi Pendidikan Khusus Indonesia (APPKHI). <http://appkhi.or.id/Proceedings ICSDDGD.pdf>
- Schmidt, K. M., & Embretson, S. E. (2012). Item response theory and measuring abilities. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology - Volume 2: Research methods in psychology* (2nd ed., pp. 451–473). John Wiley & Sons.
- Sim, S. M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals of the Academy of Medicine, Singapore*, 35(2), 67–71.
- Wang, J., & Wang, X. (2019). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.
- Wolanin, T. R., & Steele, P. (2004). *Higher education opportunities for students with disabilities: A primer for policymakers*. The Institute for Higher Education Policy.
- Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, 59(4), 453–470. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/04_Wu.pdf
- Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29(1), 18. <https://doi.org/10.1186/s41155-016-0040-x>
- Zoghi, M., & Valipour, V. (2014). A comparative study of Classical Test Theory and Item Response Theory in estimating test item parameters in a linguistics test. *Indian Journal of Fundamental and Applied Life Sciences*, 4(s4), 424–435. <https://www.cibtech.org/sp.ed/jls/2014/04/JLS-051-S4-052-VALIPOUR-COMPARATIVE.pdf>

Analysis of mathematics test items quality for high school

Budi Manfaat; Ayu Nurazizah*; Muhamad Ali Misri

Institut Agama Islam Negeri Syekh Nurjati Cirebon

Jl. Perjuangan, Sunyaragi, Kesambi, Kota Cirebon, Jawa Barat 45132, Indonesia

*Corresponding Author. E-mail: ayunurazizah@mail.syekhnurjati.ac.id

ARTICLE INFO

Article History

Submitted:

4 March 2021

Revised:

14 July 2021

Accepted:

1 August 2021

Keywords

item analysis; quality test;
mathematics test

Scan Me:



How to cite:

Manfaat, B., Nurazizah, A., & Misri, M. (2021). Analysis of mathematics test items quality for high school. *Jurnal Penelitian dan Evaluasi Pendidikan*, 25(1), 108-117. doi:<https://doi.org/10.21831/jpep.v25i1.39174>

ABSTRACT

This study aims to determine the quality of the mathematics test items at high school in terms of validity, reliability, differentiation, difficulty level, and distractor effectiveness. This study is an evaluation type of research with a quantitative approach. The subjects in this study were 44 class XII students of SMKN 3 Kuningan and 39 class XII students of SMAN 1 Jalaksana. The results show that the majority (96.67%) of the items are declared valid in content by the experts. The test has very high reliability (0.90). The items have the ideal difficulty level. Most of the questions (70%) have medium difficulty, a few questions (6.67%) are very easy, and a few questions (20%) are difficult, and (3.3%) are very difficult. Most of the items (83.33%) have good discriminating power, and only a few questions (16.67%) have poor discriminating power. Most (90%) of the questions have a well-functioning answer choice, and only a few questions (10%) have the answer choice not functioning properly. Overall, this study can be concluded that the Mathematics Test Questions at SMKN 3 Kuningan are of good quality.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



INTRODUCTION

Assessment is one of the five main tasks of teachers in learning, done by collecting and processing information to measure the achievement of student learning outcomes. Gronlund and Linn interpret assessment as a systematic process that collects, analyzes, and interprets information to decide how far the student or a group of students achieve their learning goals either in knowledge aspect, attitude, or creativity (Kusaeri & Suprananto, 2012). Assessment can also be interpreted as collecting proof using some instruments, which can also be used to decide whether the assessor can be recommended as competent or not (Setiawan, 2018). Thus, it can be concluded that assessment is a process of collecting information that can be done after the learning process to increase the student's achievement in the learning process.

The activity of assessing learning needs to be done to measure the level of success of a student as a student. The goal of assessment is to know student's will of learning in cognitive, affective, and psychomotor (Sukanti, 2011). Another goal of assessment is getting information about the student's learning process and the goals of teaching itself (Ratnawulan & Rusdiana, 2015). Therefore, the purpose of the implementation of the assessment of student learning outcomes is to measure students' success in mastering predetermined competencies and measuring the extent of the level of the student's understanding of the learning material delivered by the teacher. Thus, each teacher must assess the mastery of knowledge possessed by students during the learning phase in school. Besides, teachers must participate in the creation of evaluation methods for student learning outcomes and the analysis of student learning outcomes. One of the tools for assessing teachers as educators is their ability to assess learning.

Given the importance of placing assessment in learning, teachers as evaluators are required to be able to understand and make assessment tools. One of the assessment tools in the learning process is a test. Testing is a technique widely used in education. Goodenough (Baskoro, 2018) defines a test as a task or a series of tasks given to an individual or group of individuals with the intention of comparing their skills with each other. One of the assessment methods used to assess student learning outcomes is the exam. Students should be given tasks as part of the test method to monitor and measure their actions.

It can be considered a good test if the tests are valid and reliable. This means that in terms of preparation, it has met the rules of writing questions, both in terms of construction, substance, or material aspects. Each item should have content validity. This means that the evaluation tool actually contains the material to be measured so that the conformity between the evaluation tool and the content of the material that should be measured is actually displayed in the preparation of the questions. In addition, an assessment tool has high reliability if it shows the accuracy of the results. In other words, the person being tested will get almost the same score if he is tested again with the same test equipment.

In order to measure students' actual abilities, the test kites to measure students' learning outcomes must meet high-quality test qualifications. A higher-quality test instrument can capture students' abilities as individuals or as a group and provide accurate data. This can assist teachers in improving the quality of their students' learning.

To obtain a quality test, a teacher needs to carry out an item analysis. Item analysis is one of the important activities in test development to get quality questions. The goal is to examine each item in order to obtain quality questions before use. In addition, the objective of question analysis is also useful for improving the quality of the items by revising the questions and removing ineffective questions (Kusaeri, 2014). Thus, assessing student learning outcomes or the test can describe a sample of behavior and produce objective and accurate scores.

There are three prerequisites for the assessment or test instrument to be valid (Kusaeri & Suprananto, 2012). A test is said to be valid if it can measure what you want to measure. The consistency of measurement results is referred to as reliability. This means that if a student's answer to the test is tested repeatedly, it will produce consistent results, whereas the usable is related to the practical meaning of the test procedure.

In assessing learning outcomes, the tests are expected to describe behavioral samples, produce objective and accurate scores (Iskandar & Rizal, 2018). According to Suryabrata, a good test is must be standardized, objective, and discriminatory (Suharman, 2018). The test must be standardized, which means that each student must receive the same treatment in terms of test material, scoring, and interpretation of test results so that students who receive a certain score in one location will receive the same score for students on the spot. The test must be objective, which means that one teacher will assess the same test-taker as another teacher. Furthermore, the test must be discriminatory, which means it must be able to detect differences in a symptom found in each student.

The test is said to be a good question if it meets the requirements such as eligibility, level of difficulty, distinguishing power, pattern distribution of answers, and the relationship or correlation of each item with the overall score (Wijayanti, 2020). To find these things, an assessment activity is needed through item analysis which aims to obtain information about questions that have met the requirements of good questions.

Not all tests used are good tests. Research on item analysis has been carried out; for example, Maryanes et al. (2018) showed the results of research on the analysis of the quality of the Mathematics final examination items, namely most (80%) of the items were declared invalid. The reliability coefficient ranges from $0.0824 < 0.70$. This question means that it includes questions that have very low reliability. The difficulty level obtained results 37% in the medium category and 63% for the difficult category. The difference power got 67% bad question category, 10% good question category, 23% enough question category. This is also sup-

ported by Hamimi et al. (2020), who analyzed the Mathematics Exam items for Class VII Odd Semester in the 2017/2018 academic year. The result of the research was that most of the questions used were invalid because there were questions that had low validity, and some were even very low. In addition, this question also has a low degree of reliability or can be said to be unreliable. The results of research conducted by Tilaar and Hasriyanti (2019) show that (1) for the type of multiple-choice questions, few questions (16.67%) had very good quality so that the questions could be saved for reuse. As many as 50.00% of the questions still need to be revised, and 33.33% have very bad quality, so that they cannot be stored in the question bank. (2) For the type of description questions, 40.00% of the items had good quality, 40.00% of the items needed revision, and 20.00% of the items had poor quality. This study indicates that the majority of the questions did not meet the criteria for good questions and must be revised.

Based on interviews with several mathematics teachers at SMKN 3 Kuningan, one of the favorite schools in Kuningan Regency, teachers usually took questions from various mathematics books in compiling tests, both for practice and final semester assessments. However, the results obtained from the test are still many students who get scores below the SKM (*Skor Ketuntasan Minimal*/Minimum Completeness Score). In fact, the teachers at the school already have good educational qualifications, and the school has adequate facilities and infrastructure.

It means that there is a possibility that the student's learning outcomes do not describe the actual situation regarding students' abilities. A learning result is obtained using a test that does not describe a student's learning achievement. It may be the score of each student that higher or lower than the actual ability. The results of this study will provide erroneous information about the achievement of learning outcomes. Therefore, the quality of the test is one of the factors that need to be addressed in learning evaluation activities. In addition, many teachers still do not pay attention to the quality of the items because they have not analyzed the items, so the items used cannot detect whether the questions have met the criteria or not.

From the aforementioned statement, it can be seen that teachers pay less attention to the principles in assessing student learning outcomes (Zuhera et al., 2017), causing the learning outcomes to be out of sync with the students' abilities. This will be risky if the assessment of learning outcomes is not in accordance with students' abilities. Thus, the researchers wanted to know the quality of the mathematics test items used at SMKN 3 Kuningan.

RESEARCH METHOD

This study is an evaluation research. In this study, the researchers evaluate the items contained in the Mathematics Module Class XI book SMKN 3 Kuningan, which aims to determine the quality of the items in terms of validity, reliability, difficulty level, discriminating power, and deceptive effectiveness. The subjects in this study were 44 students of class XII of SMKN 3 Kuningan and 39 students of class XII of SMAN 1 Jalaksana. The average score for the math test at SMKN 3 Kuningan is 32.3, and the average score for the math test at SMAN 1 Jalaksana is 48.3. Thus, the overall math test average score is 39.8.

In this study, the target population was all items of mathematics module for class XI SMKN 3 Kuningan. The available population was all of the items in the mathematics module for class XI published in 2019. The sample consisted of 30 multiple-choice questions drawn from each chapter. In choosing 30 items, the researchers adjusted it to the test time of 120 minutes.

The data obtained from this study are quantitative data and data from instrument validation by an expert or expert judgment. The quantitative data in question are data obtained from students' responses or answers to the questions given to be analyzed in terms of quality. The sources of data in this study are validators or experts (expert judgment) and students. To assess whether the test instrument was valid or not, the researchers asked three validators or experts to assess it.

The technique used by researchers in collecting data is the documentation technique. The researchers collected documents in the form of test sheets taken by the mathematics module and student answer sheets, and the validation results from the expert judgment (expert judgment). This documentation aims to determine the quality of the items contained in mathematics textbooks.

$$CVR = \frac{ne - \frac{N}{2}}{\frac{N}{2}} \dots \dots \dots (1)$$

$$r_{tt} = \frac{2r_{11}}{1 + \left| \frac{r_{11}}{2} \right|} \dots \dots \dots (2)$$

$$\text{Difficulty level} = \frac{\text{numbers of students who answer the item correctly}}{\text{total number of students who answer the item}} \dots \dots \dots (3)$$

$$DP = \frac{RU - RL}{\frac{1}{2}T} \dots \dots \dots (4)$$

$$IP = \frac{P}{(N - B)(n - 1)} \times 100\% \dots \dots \dots (5)$$

Content validation analysis uses the CVR (Content Validate Ratio) formula, namely Formula (1), as suggested by Lawshe (Azwar, 2019). Meanwhile, the reliability analysis uses the Spearman-Brown Formula (2) (Azwar, 2019). The difficulty level of the items will be analyzed using Formula (3) (Jannah et al., 2021). The item difference was analyzed using Formula (4) (Jannah et al., 2021). The effectiveness of the trickster was analyzed using the Formula (5) (Rahayu & Djazari, 2016).

FINDINGS AND DISCUSSION

The results of this study were obtained from students' responses and analyzed using the Anates program, which included reliability, difficulty level, discriminating power, and the effectiveness of questioners. The expert's assessment was analyzed for validity in order to obtain a content-valid test. The results and discussion are elaborated as follows.

Validity

Validity is a concept that measures the extent to which the test can measure what should be measured. The test is valid if it can reveal data from the variables appropriately and does not deviate from the actual situation.

Content validity is defined as the degree to which a test's content is reasonable or relevant, as determined by a qualified panel's objective analysis or expert judgment (Azwar, 2019). This validity refers to the degree to which the test has proof of the content's validity as determined by logical analysis or the test's logic. The query indicator was used to change each item's content validity. After the instrument in the form of a test is assessed by experts or expert judgment and calculated using the CVR formula, the results of the content validity are shown in Table 1.

The data in Table 1 show the validation results by the expert that 3.33% (1 item) are declared as invalid questions, and 96.67% (29 items) are stated as valid questions with a CVR value of 1. When compared with the results of the research by Maryanes et al. (2018), Hamimi et al. (2020), and Tilaar and Hasriyanti (2019), the results of the research on the mathematics test items are better because most of the items (96.67%) are valid.

Table 1. Result Content Validity Analysis

Category	Number of Items	Number of Questions	Percentage
Valid	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29	29	96.67%
Invalid	30	1	3.33%

Thus, it can be concluded that the mathematics test questions at SMKN 3 Kuningan are of good quality because most (96.67%) are declared contentedly valid by experts. This means that most of the items are in accordance with the competency indicators to be measured. In other words, the test is structured based on the content of the material being evaluated. The question points contain relevant content and do not go beyond the boundaries of the learning objectives. As for the follow-up to the results of the analysis of the mathematics test items taken from the Mathematics book of Class XI SMKN 3 Kuningan, namely valid items, they are entered into the question bank and can be reused for learning outcomes tests, while invalid items are declared as failed questions should be thrown away.

Reliability

From the analysis of the Anates program, the reliability coefficient of the test is 0.90. According to Fraenkel et al., if an instrument can be said reliable, that is, when the Spearman-Brown reliability coefficient is more than 0.70 ($r_i > 0.70$) (Yusup, 2018). This means that the test results meet a high level of consistency, so the assessment results of the same test-taker using the test repeatedly will show the same results. This means that this mathematics test is of good quality when viewed from the point of view of the reliability of the questions.

One of the characteristics of the high-reliability questions is that the test used consists of the number of valid questions as the result of the validity analysis. Arifin (2013) states that a reliable test is if it has a high coefficient and low standard error of measurement. Problems with very high reliability mean that the test can show consistent results if tested repeatedly on a test taker as long as the aspects measured in the test taker have not changed. Compared to the results of previous studies that obtained low reliability, this math test has a better degree of consistency so that the test is consistent and reliable to use as a measuring tool. When compared to the results of previous studies, Maryanes et al. (2018) obtained very low reliability, while Hamimi et al. (2020) obtained a low or unreliable degree of reliability, this mathematics test has a better level of constancy so that the test is consistent and reliable to be used as a measuring tool.

Level of Difficulty

The level of difficulty is related to how many or at least students answer the questions correctly. The difficulty of a question in a test is seen from the ability of students to answer the question, not from the assumptions of the teacher who compiled the questions. Because questions that are difficult or easy for teachers are not necessarily difficult or easy for students (Wijayanti, 2020). In order to find out the difficulty level of an issue, it is necessary to calculate the difficulty index. The difficulty index ranges from 0.00 to 1.00 (Zein & Darto, 2012). The smaller the difficulty index obtained, the more difficult the question is for students to answer, and vice versa. In principle, the average score obtained by the testee on the item in question is called the item difficulty level (Sumiati et al., 2018). The description of the difficulty level analysis results using the Anates version 4.0.9 program is presented in Table 2.

Table 2 shows that most of the questions (70%) had medium difficulty, a few questions (6.67%) are very easy, and a few questions (20%) are difficult, and (3.3%) are very difficult. A good question is one that is not too easy or not too difficult (Arikunto, 2013). Problems that

are too easy will not stimulate students to solve a more difficult problem, and questions that are too difficult will result in a loss of enthusiasm when working on the problem because the questions presented are beyond their ability.

Table 2. Results of Difficulty Level Analysis

Difficulty Category	Number of Items	Number of Questions	Percentage
Very easy	10, 22	2	6.67%
Easy	-	-	0%
Medium	1, 3, 4, 5, 6, 7, 9, 11, 13, 14, 17, 18, 19, 20, 21, 25, 26, 27, 28, 29, 30	21	70%
Difficult	2, 8, 12, 15, 16, 23	6	20%
Very difficult	24	1	3.33%

Compared to the research results of [Maryanes et al. \(2018\)](#), which obtained 67% difficult questions, the results of [Tilaar and Hasriyanti \(2019\)](#) show that 46.67% of the questions are categorized as difficult, so this math test shows an enhancement. Compared to the results of previous studies that obtained 67% of difficult questions, this math test shows an increase. Most of the items used in this study had an ideal difficulty level, namely in the medium category. It can be concluded that the questions contained in this math test are of good quality.

Questions that have a moderate level of difficulty can be stored in the question bank for reuse in other tests. As for the easy and difficult items, three possible follow-ups can be done: items are discarded and not reused in the next test; researching or tracing back what causes the items are so easy to answer or too difficult to answer by students. Then improvements were made so that the items could be reused. Easy items can be reused in loose tests with the aim that most test-takers pass the test. In contrast, difficult items can be reused in very strict tests, usually selection tests, so that they can be stored in a bank as a separate question.

Discriminating Power

Discriminating power, namely the level of ability of the items, can distinguish between smart students (students who have mastered the material) and less intelligent students (students who have not mastered the material). Like the degree of difficulty, the discriminating power also has an index called the discriminating power index (DP). To find the discriminating power of a problem, it is necessary to calculate the index of discriminating power. The discriminatory index ranges from 0.00 to 1.00 ([Supriadi et al., 2019](#)). The higher the discriminant index of an item, the better it is at distinguishing between intelligent and less intelligent students ([Kusumawati & Hadi, 2018](#)). Based on the analysis of discriminating power using the Anates program, the research results were obtained, as presented in Table 3.

Table 3. Results of Discriminatory Power Analysis

Category Discriminating Power	Number of Items	Number of Questions	Percentage
Very good	1, 3, 6, 9, 13, 18, 19, 25, 26, 27, 28, 29	12	40%
Good	2, 4, 5, 7, 11, 12, 14, 16, 17, 20, 21, 22, 30	13	43.33%
Enough	8, 15, 23	3	10%
Bad	10, 24	2	6.67%
Very bad	-	0	0%

Table 3 shows that most (83.33%) items have good discriminating power. Only a few questions (16.67%) were poorly differentiated. That is, most of the questions can show the difference between high-ability students and low-ability students. In this case, smart students

or students in the upper group answered more correctly, while less smart students or students in the lower group mostly could not answer the questions correctly.

When compared to the results of the research by [Maryanes et al. \(2018\)](#), which obtain 67% of the distinguishing power in the poor/bad category, 10% in the good category, 23% in the sufficient category, the research results of [Tilaar and Hasriyanti \(2019\)](#) show that 10% results are in a bad category (negative), 26.67% in the poor category, so this math test shows an increase. Most of the items used in this study have the ability to differentiate between students who are smart and less intelligent students. It can be concluded that the questions contained in this math test are of good quality.

Problems with excellent, good, and sufficient discriminating power can be put in the question bank for reuse. However, questions that are categorized as sufficient if they want to be reused must be revised first. Meanwhile, questions that have bad or bad distinctions should be discarded or revisited. There may be several reasons that allow the item not to show differences in student abilities, including having two or more correct answers, not having the correct answer, the measured competency is not clear, the bully does not function, the material asking questions is too difficult so that many students answer the original -casual, or most students are good at misconceptions or misinformation in understanding questions. Therefore, questions with poor discriminatory power should be reviewed and corrected for reuse in other tests.

Deceptive Effectiveness

In the analysis of multiple-choice items, the effectiveness of distractors is one of the requirements for an item to be said to be of good quality. A good item is the one which the deceiver is chosen evenly by all test takers who answer wrongly ([Arifin, 2013](#)). A good distractor is chosen by 5% of the total number of test-takers ([Arikunto, 2013](#)) and good distractors affect the result of the test to differentiate the upper group students from the lower ones ([Hartati & Yogi, 2019](#)). On the other hand, bad items are those in which the test taker unevenly selects the cheater. Based on the results of the decoy effectiveness analysis, the data in Table 4 are obtained.

Table 4. Results of Confusing Analysis

Quality of Swag	Number of Items	Number of Questions	Percentage
Very good	6, 11, 14, 20, 21, 24, 25, 27, 28	9	30%
Good	1, 2, 4, 5, 7, 12, 17, 18, 23, 30	10	33.33%
Enough	3, 8, 9, 10, 13, 16, 19, 26,	8	26.67%
Bad	22	1	3.33%
Very bad	15, 28	2	6.67%

Table 4 shows that the items most (90%) of the cheaters functioned properly. Few (10%) of cheaters are not functioning properly. Seen from the effectiveness of the distractor, this math test question is included in a good quality question compared to the research results of research by [Tilaar and Hasriyanti \(2019\)](#), which has 23.33% distractors in the less category and 20.00% in the poor category. Thus, most of the items in this math test have choices of answers that functioned properly.

According to [Arikunto \(2013\)](#), it was followed up after analyzing the items in terms of the effectiveness of the deceiver, namely: cheaters are accepted because they are categorized as good. This means that all distractors on an item have been chosen by 5% of the total number of test-takers; the bully is rewritten for not being good. This means that the deceiver has not performed its function properly (chosen by less than 5% of test-takers) and was rejected because the cheater is not good. This means that the test taker is not chosen at all (0%).

Quality of Question Items

After analyzing the items based on the validity, reliability, difficulty level, discriminating power, and tricking effectiveness, the next step is to draw conclusions on the quality of each item. In drawing these conclusions, the researchers use four aspects: validity, difficulty level, discriminating power, and deceptive effectiveness. The reliability aspect is to determine the level of consistency or consistency of an item as a whole. Therefore, reliability is not used in determining the quality of each item.

Table 5. Distribution of Overall Question Item Quality Analysis Results

The Quality of the Items	Number of Items	Number of Questions	Percentage
Good	1, 4, 5, 6,7, 9, 11, 14, 17, 18, 20, 21, 25, 27, 29	15	50%
Enough	2, 3, 8, 12, 13, 19, 23, 26, 28, 30	10	33.33%
Bad	10, 15, 16, 22, 24	5	16.67%

Based on the results of the analysis of the quality of the mathematics test questions as a whole, it was found that most (83.33%) items were classified as good quality questions by fulfilling the four criteria for good items, namely valid, had a moderate level of difficulty, had good discriminating power and trickster works fine. This is supported by the results of the reliability analysis with the reliability coefficient of the test of 0.90, which means very high.

As for the items that are not of good quality or fail, the cause should be traced. This is very useful for teachers in compiling questions in order to get good quality. Table 6 presents the causes of the failure of the mathematics test items seen from the aspects of validity, difficulty level, discriminating power, and deceptive effectiveness.

Table 6. Causes of Question Item Failure

The Cause of the Failure of the Item	Number of Items	Number of Questions	Percentage
Validity	30	1	3.33%
Level of difficulty	2, 8, 10, 12, 15, 16, 22, 23, 24	9	30%
Discernment	10, 24	2	6.67%
Deceptive Effectiveness	15, 16, 22, 28	4	13.33%

From Table 6, it can be seen the cause of the drop in good quality items. Dissimilarity is not recommended because the difficulty index is either too low or too high (Rahmat et al., 2020). This means items that are too difficult or too easy cannot discriminate between smart students and less smart students, so they do not have good discriminating power. In addition, discriminatory power can be affected by distractors. The distractor is said to be effective if the test taker chooses it from the lower group. Otherwise, if the test taker chooses it from the upper group, it means that the distractor is not functioning properly. In other words, the test items cannot discriminate between smart and less smart students.

CONCLUSION

In terms of the quality of the question items contained in the Mathematics Module of Class XI SMKN 3 Kuningan book, the results of the analysis of the samples representing each subject are most (96.67%) of the items declared valid in content by the expert. This means that most of the items are in accordance with the competency indicators to be measured. The test has very high reliability (0.90). This means that the test results meet a high level of consistency. The items have the ideal difficulty level. Most of the questions (70%) had medium

difficulty, a few questions (6.67%) were very easy, and a few questions (20%) were difficult, and (3.3%) were very difficult. Most (83.33%) items had good discriminating power. Only a few questions (16.67%) had poor differentiation. This means that most of the questions were able to distinguish high and low-ability students. Most (90%) of the items had functional answer choices. Only a few questions (10%) had the answer choices not working properly. Overall, this study can be concluded that the Mathematics Test Questions at SMKN 3 Kuningan are of good quality.

REFERENCES

- Arifin, Z. (2013). *Evaluasi pembelajaran*. Remaja Rosdakarya.
- Arikunto, S. (2013). *Dasar-dasar evaluasi pembelajaran* (2nd ed.). Bumi Aksara.
- Azwar, S. (2019). *Reliabilitas dan validitas* (4th ed.). Pustaka Pelajar.
- Baskoro, E. P. (2018). *Perencanaan pelaksanaan dan evaluasi pembelajaran*. Eduvision.
- Hamimi, L., Zamharirah, R., & Rusydy, R. (2020). Analisis butir soal ujian Matematika kelas VII semester ganjil tahun pelajaran 2017/2018. *Mathema: Jurnal Pendidikan Matematika*, 2(1), 57–66. <https://doi.org/10.33365/jm.v2i1.459>
- Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better quality test. *English Language in Focus (ELIF)*, 2(1), 59–70. <https://doi.org/10.24853/elif.2.1.59-70>
- Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 22(1), 12–23. <https://doi.org/10.21831/pep.v22i1.15609>
- Jannah, R., Hidayat, D. N., Husna, N., & Khasbani, I. (2021). An item analysis on multiple-choice questions: A case of a junior high school English try-out test in Indonesia. *Leksika: Jurnal Bahasa, Sastra Dan Pengajarannya*, 15(1), 9–17. <https://doi.org/10.30595/lks.v15i1.8768>
- Kusaeri, K. (2014). *Acuan dan teknik penilaian proses dan hasil belajar dalam kurikulum 2013*. Ar Ruzz Media.
- Kusaeri, K., & Suprananto, S. (2012). *Pengukuran dan penilaian pendidikan*. Graha Ilmu.
- Kusumawati, M., & Hadi, S. (2018). An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school. *REiD (Research and Evaluation in Education)*, 4(1), 70–78. <https://doi.org/10.21831/reid.v4i1.20202>
- Maryanes, F., Fitriati, F., & Salmina, M. (2018). Analisis kualitas butir soal ujian akhir semester mata pelajaran Matematika kurikulum 2013 kelas VII SMP Negeri 8 Banda Aceh. *Prosiding Seminar Nasional Pendidikan Dasar 2018*, 576–580. <https://repository.bbg.ac.id/handle/737>
- Rahayu, R., & Djazari, M. (2016). Analisis kualitas soal pra ujian nasional mata pelajaran Ekonomi Akuntansi. *Jurnal Pendidikan Akuntansi Indonesia*, 14(1), 84–94. <https://doi.org/10.21831/jpai.v14i1.11370>
- Rahmat, I., Masi, L., & Anggo, M. (2020). Analisis butir soal Ujian Akhir Semester (UAS) di masa pandemi pada mata pelajaran Matematika untuk tahun ajaran 2019/2020 kelas VII SMPIT Bina Insan Mandiri Al-Masrur Kendari. *Jurnal Penelitian Pendidikan Matematika*, 8(3), 491–504. <https://doi.org/10.36709/jpp.v8i3.16624>
- Ratnawulan, E., & Rusdiana, A. (2015). *Evaluasi pembelajaran*. Pustaka Setia.

- Setiawan, D. F. (2018). *Prosedur evaluasi dalam pembelajaran*. Deepublish.
- Suharman, S. (2018). Tes sebagai alat ukur prestasi akademik. , . Retrieved from. *At-Ta'dib: Jurnal Ilmiah Prodi Pendidikan Agama Islam*, 10(1), 93–115. <https://ejournal.staindirundeng.ac.id/index.php/tadib/article/view/138>
- Sukanti, S. (2011). Penilaian afektif dalam pembelajaran Akuntansi. *Jurnal Pendidikan Akuntansi Indonesia*, 9(1), 74–82. <https://doi.org/10.21831/jpai.v9i1.960>
- Sumiati, A., Widiastuti, U., & Suhud, U. (2018). Workshop teknik menganalisis butir soal dalam meningkatkan kompetensi guru di SMK Cileungsi Bogor. *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, 2(1), 136–153. <https://doi.org/10.21009/JPMM.002.1.10>
- Supriadi, W. O. S., Rahim, U., & Zamsir, Z. (2019). Kualitas tes sumatif mata pelajaran Matematika kelas VIII semester genap SMP Negeri 20 Kendari tahun pembelajaran 2016/2017. *Jurnal Penelitian Pendidikan Matematika*, 6(3), 85. <https://doi.org/10.36709/jppm.v6i3.9142>
- Tilaar, A. L. F., & Hasriyanti, H. (2019). Analisis butir soal semester ganjil mata pelajaran Matematika pada sekolah menengah pertama. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia (JP3I)*, 8(1), 57–68. <https://doi.org/10.15408/jp3i.v8i1.13068>
- Wijayanti, P. S. (2020). Item quality analysis for measuring mathematical problem-solving skills. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, 9(4), 1223–1234. <https://doi.org/10.24127/ajpm.v9i4.3036>
- Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, 7(1), 17–23. <https://doi.org/10.18592/tarbiyah.v7i1.2100>
- Zein, M., & Darto, D. (2012). *Evaluasi pembelajaran matematika*. Daulat Riau.
- Zuhera, Y., Habibah, S., & Mislinawati, M. (2017). Kendala guru dalam memberikan penilaian terhadap sikap siswa dalam proses pembelajaran berdasarkan kurikulum 2013 di SD Negeri 14 Banda Aceh. *Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, 2(1), 73–87. <http://www.jim.unsyiah.ac.id/pgsd/article/view/2534>