REiD (RESEARCH AND EVALUATION IN EDUCATION)
Vol. 6, No. 1, June 2020

Indexed in:
DOAJ DIRECTORY OF OPEN ACCESS JOURNALS ISJD Google Scholar sinta Science and Technology Index

9 772460 699001

REiD

Research and Evaluation in Education

Vol. 6, No. 1, June 2020

Vol. 6, No. 1, 2020 ISSN 2460-6995

Research and Evaluation in Education

REiD

Publisher:
PROGRAM PASCASARJANA UNIVERSITAS NEGERI YOGYAKARTA
in cooperation with
HIMPUNAN EVALUASI PENDIDIKAN INDONESIA

REiD (Research and Evaluation in Education)
ISSN 2460-6995

Publisher
Program Pascasarjana Universitas Negeri Yogyakarta
in Cooperation with Himpunan Evaluasi Pendidikan Indonesia

Published biannually, in June and December
REiD disseminates articles written based on the results of research focusing on assessment, measurement, and evaluation in various educational areas

Editorial
Department of Educational Research and Evaluation, Graduate School of Yogyakarta State University
2nd Floor of General Lecture Building, Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281
E-mail: reid.ppsuny@uny.ac.id, reid.ppsuny@gmail.com

Foreword

We are very pleased that REiD (Research and Evaluation in Education) is releasing its eleventh edition. We are also very excited that the journal has been attracting papers from the neighbouring country, Malaysia. The variety of submissions from different countries will help the journal in reaching its aim in becoming a global initiative.

REiD (Research and Evaluation in Education) contains and spreads out the results of research which is not limited to the area of common education, but also comprises the results of research in education in a broader coverage, childhood education, language learning, learning media testing, mathematics education, natural science, and social sciences, with focuses on assessment and evaluation.

The editorial board expects comments and suggestions for the betterment of the future editions of the journal. Special gratitude goes to the reviewers of the journal for their hard work, contributors for their trust, patience, and timely revisions, and all staffs of the Graduate School of Universitas Negeri Yogyakarta for their assistance in publishing this issue.

Yogyakarta, June 2020

Editor in Chief

TABLE OF CONTENT

REiD (Research and Evaluation in Education), 6(1), 2020, 1-9
Available online at: http://journal.uny.ac.id/index.php/reid

Developing instruments for measuring the level of early childhood development

[*1]I Wayan Gunartha; [2]Tajularipin Sulaiman; [3]Siti Partini Suardiman; [4]Badrun Kartowagiran

[1]Faculty of Language and Arts Education, Institut Keguruan dan Ilmu Pendidikan PGRI Bali
Jl. Seroja, Tonja, Denpasar Timur, Kota Denpasar, Bali 80235, Indonesia
[2]Faculty of Educational Studies, Universiti Putra Malaysia
Persiaran Masjid, 43400 Serdang, Selangor, Malaysia
[3]Faculty of Teacher Training and Educational Sciences, Universitas Ahmad Dahlan
Jl. Ahmad Yani (Ringroad Selatan), Tamanan, Banguntapan, Bantul, Yogyakarta 55166, Indonesia
[4]Faculty of Engineering, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
[*]Corresponding Author. E-mail: w.gunartha@yahoo.com

Submitted: 11 January 2020 | Revised: 17 April 2020 | Accepted: 20 April 2020

Abstract

The aims of the study were to: (1) develop a set of instruments to measure the level of early childhood development (kindergarten group B), and (2) assess the quality of the developed instruments. This study is developmental research. The samples of the study were the students of kindergarten group B. The developed instrument was a set of questionnaires. Instrument testing was carried out in three stages with the number of subjects increased on each stage. The validity analysis of the questionnaire used confirmatory factor analysis (CFA). The reliability estimation of the questionnaire used composite reliability. The results of the study are in the form of instruments for measuring the level of early childhood development, which consists of an instrument to measure religious morality, social-emotional, language, cognitive, and physical-motor development. Based on field study, all instruments have a good fit model, construct validity, and reliability that meet the academic requirements of early childhood education.

Keywords: *instrument development, early childhood, kindergarten, potential*

How to cite: Gunartha, I., Sulaiman, T., Suardiman, S., & Kartowagiran, B. (2020). Developing instruments for measuring the level of early childhood development. *REiD (Research and Evaluation in Education), 6*(1), 1-9. doi:https://doi.org/10.21831/reid.v6i1.21996

Introduction

Early childhood is the most important and fundamental beginning period. Therefore, early childhood is often called the Golden Ages. This period is also often called a sensitive period, the period of play, the critical period because this period will affect the future lives of the child. According to Woolfolk (2007, p. 23), approximately one month after conception, human brain development has begun. Neuron cells appear with incredible speed, i.e. 50,000 to 1,000,000 per second for approximately three months. When born, we have had about 100 to 200 billion neurons, and each neuron has about 2,500 synapses. Synapses that are unused or not getting stimulation from the environment will be trimmed (synaptic pruning). Berk (2007, p. 121) added that the complexity of connections between neurons will determine the child's level of intelligence. The same thing was said by Miller and Cumming (Rushton, 2011, p. 92).

All of the aforementioned statements show that the growth and development of the brain (synapses) are determined by the stimuli or stimulus provided to the child and the activities undertaken by children. Thus, children who are growing and developing must be activated by providing a variety of stimuli and appropriate activities. This is the important role of early childhood education (ECE) as a form of physical and psychosocial stimulation whether at home or in early childhood institutions, besides nutrition and health care. Hence, according to Valentine, Thomson, and Antcliff (2009, p. 196), in Australia, even early education and parenting has got priority policy.

An appropriate stimulation, as well as nutrition, parenting, and also health services at an early age, will develop all of children's potential, including their physical, cognitive, language, art, social-emotional, self-discipline, religious values, self-concept, and also self-reliance, will develop optimally. Therefore, early childhood education is expected to contribute significantly to the improvement of human resources (HR) quality, which will make our nation to be high quality and full of competitiveness in the future.

The importance of stimulation obtained by children in early childhood education institutions is also proved empirically by many experts. Samuelsson (2011, p. 109) in her research on the role of early childhood education said that learning at an early age has an influence in the future, for example, the success of the school, as well as the attitude and attention, will be formed early on. Mann and Reynolds (2006, p. 153) concluded that preschool intervention correlates with a reduction in the incidence, frequency, and severity of the delinquency at age 18 years.

Ashiabi (2007, pp. 205–206) states that a lot of advantages can be attained to let children playing with other children. For example, sociodramatic playing can improve a child's ability to imagine before acting, taking a role, empathy, altruism, and also emotions and rules understanding. Moreover, negotiation and problem-solving skills also increase, such as the ability to work cooperatively with others, share, self-control, and working with the group. In other words, sociodramatic play can enhance children's social and emotional development optimally. A similar study was conducted by Beard and Sugai (2004, p. 408).

By the importance of early childhood education, the government's attention to developing early childhood education becomes greater. Since 2000, Early Childhood Education (ECE) started to become a central issue in education, including in Indonesia, even more, Erman Syamsudin, Director of Early Childhood Development (Ministry of National Education of Republic of Indonesia, 2011, p. vii), stated that early childhood education is one of the priority programs of national education development. Early Childhood Education (ECE) services are expected to nurture, grow, and develop the whole early childhood potential optimally so it can form the basic ability and behavior according to the children's development stage.

In response to the government policy, the public has shown their concern for the problems of education, protection, and care of early-aged children with a variety of services in accordance with their conditions and capabilities. Public awareness of the importance of early childhood education in the optimal development of children's potential has been shown with various active participation in the implementation and improvement of services. Although various policies have been issued by the government, in fact, there are still many problems that exist in the implementation of early childhood services including in Badung Regency, Province of Bali. There are still many children who have not gained early childhood services. This fact acknowledged by the General Director of Early Childhood Education, Non-Formal and Informal that although the policies have been established and socialized related to early childhood development, in fact, of the 28, eight million children aged 0-6 years in late 2009, who gained an early childhood education services is just around 53.7% (Ministry of National Education of Republic of Indonesia, 2011, p. iii).

Research by Hiryanto (2007) about the mapping of the quality achievement level of early childhood programs in Yogyakarta reveals that by the views of the implementation guidelines to the suitability of early childhood education with the real conditions of the program implementation based on the ten benchmarks of national education, in the implementation of early childhood education in the Yogyakarta Province, some problems can be found as follows: (1) The variation in the implementation of education; (2) The existence of the age groupings that do not fit the guidelines because of limited infrastructure and educators; (3) There are still some educators who have not received training; (4) The ratio of the teachers and students number is not ideal.

The research by Hermawati (2007) in a children daycare in Beringharjo, Yogyakarta, found two drawbacks of the input variables, namely, the teacher's educational background and caregiver qualifications that are not relevant to the tasks. In the process variables, the problem is the immeasurability of mentoring by a caregiver. It is associated with the majority of low caregiver's education. In addition, the assistance has not been done regularly by the organizers. The public access to the children's daycare Beringharjo is limited because of limited capacity. Based on the preliminary study conducted at the early childhood institution in Badung Regency, Province of Bali, it was also found many problems related to the implementation of early childhood education (ECE), such as the quality and quantity of early childhood teachers are still relatively low and the number of teachers is on the average of three to four people. In terms of process, the kindergarten student has been taught reading, writing, and arithmetic skills because, according to the teacher, if it is not done, then no parents want to enroll their children to the institution.

In order to give good quality of early childhood education (ECE), in accordance with the existing standards, early childhood education services need to be evaluated regularly. According to Nugraha (2010, p. 3), good quality of early childhood education service is regularly evaluated and the results are acted upon appropriately. The same opinions are also expressed by Mardapi (2012, p. 12), that the improvement of the education quality can be achieved through improving the quality of learning by the improvement of the quality of the assessment system. Therefore, the availability of quality evaluation instruments is very important, in which it can be used by the government to evaluate early childhood education services continuously. By the results of evaluation activities, we will be able to know the things that have been achieved, whether a program can meet the established criteria or not.

Currently, the evaluation of early childhood services internally has not done thoroughly. Likewise, in Bali Province, even in Badung Regency, based on the preliminary study which has been conducted, the Department of Education has never done an evaluation of the existed early childhood services. The quality determination of early childhood institutions is often based on the frequency of competition participation and the number of early childhood learners. This is caused by the absence of evaluation instruments of early childhood services that have been tried and tested, both in terms of validity and reliability. The evaluation results will provide accurate information when obtained through evaluation using reliable instruments. Until now, the government, especially the Badung Department of Education, has not had a standard instrument for evaluation that can be used by the Department of Education or by the head of the kindergarten institution as an internal evaluation.

The education service is a system which consists of interlinked components and they mutually determine each other. These components are the input, process, and product. The component of inputs includes infrastructure, students, teachers, curriculum, and also subject matter. The component of processes includes lesson planning, implementation, and evaluation. The component of products on early childhood education services includes the achievement level of early childhood development, such as moral religious, social-emotional, language, cognitive, and physical-motor development. In evaluating early childhood education services, these components should be evaluated continuously. Therefore, it is necessary to develop an instrument to evaluate the inputs, processes, and products of early childhood services. Based on that description, the instruments developed in this study is only an instrument for measuring the products, that is, an instrument to evaluate the achievement level of the early childhood development, which includes: (a) moral-religious, (b) social-emotional, (c) language, (d) cognitive, and (e) physical-motor. This is caused by the limited costs, energy, and time available.

Based on those aforementioned backgrounds, the problems in this study are as follows. How is the instrument for measuring the level of childhood development? How is the quality of the developed instruments, both in terms of validity and reliability? Based on the problems, the purpose of this research is to develop evaluation instruments that can be used to evaluate the level of early childhood development, particularly for kindergarten group B so as to provide complete and accurate information for program managers and to assess the quality of the evaluation instruments developed.

The products of this study are a set of an evaluation instrument for early childhood education services, particularly for kindergarten group B. The evaluation instrument of early childhood education services limited to instruments tends to measure the achievement level of early childhood development, which includes moral-religious, social-emotional, language, cognitive, and physical-motor development. The development of an evaluation model for early childhood services program is very beneficial, both theoretically and practically. Theoretically, this study is useful as a contribution to developing the existed evaluation methodology to generate new concepts in the field of evaluation science. Practically, the results of this study are useful for teachers, principals of early childhood/ kindergarten, as well as the Department of Education. For teachers in early childhood education (ECE), a kindergarten teacher, in particular, this instrument can be used to measure the effectiveness of the performed services and the results can be useful as a basis to make corrections to educational services.

Method

Development Model

This study is a research and development (R & D), which aims to produce a product in the form of a set of instruments in order to evaluate the level of early-aged children's development (specifically kindergarten group B). The development research adopts the model which was proposed by Borg and Gall (1983, p. 775). The ten steps of development by Borg and Gall were then simplified into four steps, namely: (1) preliminary investigation, (2) design phase, (3) testing, evaluation, and revision, and also (4) implementation.

In the early stages, we conducted several activities, including a preliminary study, review the theory of instrument evaluation models, early childhood education, as well as review the results of research that has been done. In the design phase, the draft instrument was designed in order to measure the level of early childhood development, which consists of instruments for measuring products of services and the test design. In the pilot, evaluation, and revision phase, expert validation and testing of the instruments that have been designed in kindergarten were conducted. The data of test results were then analyzed. If the results of the analysis show that the instrument is not yet good, then it can be revised and tested again until a final prototype eligible fit model (good prototype). Tests were conducted in three phases. In the implementation phase, the instruments that have been well and subsequently tested were implemented.

Development Procedure

Several steps were taken in developing the instrument for measuring the level of early childhood development. Each step is elaborated as follows.

Drafting of the Design

At this stage, evaluation instruments to evaluate the product of early childhood services were structured, which consist of instruments for measuring religious morality, social-emotional, language, cognitive, and physical-motor development. All of those instruments were in the form of a questionnaire on a Likert scale with five points. These instruments are the first draft.

Expert Judgment

In order to check the content validity and refine the instrument draft, it was validated by experts, namely, academicians or lecturers and practitioners (kindergarten teacher), and also the user of the instrument (head/ deputy head of the kindergarten). The expert validation process used FGD (focus group discussions) model. The implementation of the FGD was conducted in two stages. The first FGD was conducted by ten academicians (lecturers) from the post-graduate program of Universitas Negeri Yogyakarta. When the instrument was revised in accordance with academicians' suggestions (lecturers), it was followed by another FGD and readability test by three kindergarten heads and also 17 kindergarten teachers. After the test was carried out, it was continued by the assessment of the instrument.

Tests

The draft of the instrument that has been revised based on the advice obtained in the FGD was piloted in kindergarten to determine the fit model of the measurement, construct validity, and reliability. The instrument test was conducted in three stages, namely, the first, second, and third with the increasing number of test subjects. The numbers of kindergarten were: 10, 13, and 18 and 160, 260, and 360 kindergarten children as the subject.

Data Analysis

The data about the comprehensiveness and also clarity of the instrument which were obtained from the experts were then analyzed descriptively. The data which were taken from the results of the field test were then analyzed using Confirmatory Factor Analysis (CFA) in order to find out the goodness of fit (GoF) as well as determine the validity and reliability, with the 8.8 Lisrel program. In determining the goodness of fit, several indicators were employed, including: (a) the value of chi-square p-value ≥ 0.05, (b) root mean square error of approximation (RMSEA) ≤ 0.08, and goodness of fit index (GFI) ≥ 0.9 (Ghozali & Fuad, 2008, pp. 29–31; Latan, 2012, p. 53). The construct reliability was calculated based on lambda (λ) for each indicator, and the error variance (δ) indicator.

In the descriptive-qualitative analysis, the average score of the quantitative data that were obtained through an assessment instrument was calculated, then were converted into qualitative data with scale 5, and then finally were interpreted qualitatively. The results of the qualitative analysis were used as the basis for determining whether the developed instrument was good or not. In converting the quantitative data into qualitative data with scale 5, a modification of rules which were developed by Sudijono (2011, p. 329) was employed. The criteria of the instrument assessment which were used are presented in Table 1.

Implementation

After the last product of the instrument had been analyzed, a good prototype was implemented in 18 kindergartens. When it is depicted in the chart, the whole developing process of the instrument model of the early childhood development is clearly illustrated in Figure 1.

Table 1. Criteria of Instrument Assessment

Average Score	Qualification	Conclusion
> 4.2	Very good	Can be an example
> 3.4 – 4.2	Good	Can be used without any revision
> 2.6 – 3.4	Quite good	Can be used with a little revision
> 1.8 – 2.6	Less good	Can be used with some revision
≤ 1.8	Bad	Cannot be an example

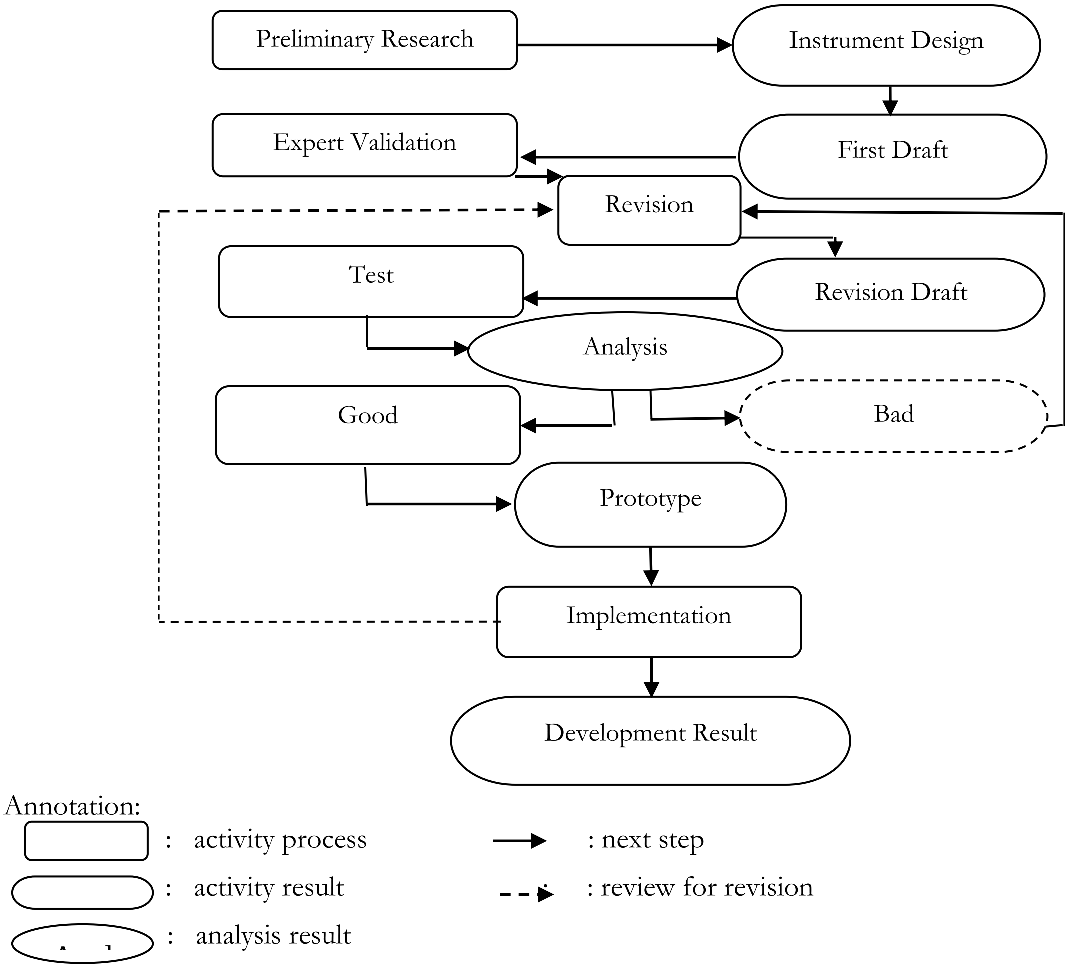


Figure 1. Flowchart of Instrument Model Development Procedure

Table 2. Instrument of Development Result

Instrument for Evaluating Early Childhood Services		
Instrument	**Evaluated Item**	**Instrument Form**
Instrument for measuring the level of early childhood development	Achievement Level of Development:	
	a. Moral-Religious	Questionnaire
	b. Social-Emotional	Questionnaire
	c. Cognitive	Questionnaire
	d. Language	Questionnaire
	e. Physical-Motor	Questionnaire

Findings and Discussion

The instrument for measuring the level of early childhood development consists of five components, namely the instruments for measuring the development of moral-religious, social-emotional, language, cognitive, as well as physical-motor components. The type of the product instrument of the early childhood services developed is clearly presented in Table 2.

Validation Result of Experts and Practitioners

The instrument assessment by experts and practitioners was directed into four main aspects, namely: (a) the clarity of instrument guidance, (b) the completeness of instrument indicators, (c) the suitability of the indicators with the point, and (d) the effectiveness of Indonesian. The assessment used a scale of 5 with the lowest score was 1 and 5 was the highest.

Based on the average score given by the experts, the mean obtained is 4.1 in total. In line with the conversion guidelines, the mean is at intervals of 3.4 to 4.2 and is classified as good. Based on the assessment conducted by teachers and heads of kindergarten, it is obtained a mean score of 4.29 in total. The mean score according to those criteria is also quite good. The total mean score of the two assessors groups is 4.2 It means that the instrument has been well conducted and can be used without any revision, shown in Table 3.

Instruments Measurement Model

Based on the analysis, all items on all instruments of the three pilot phases are significant (t> 1.96), meaning that all items can be used to measure the construct well. In the third test, there are some items of achievement level instruments for language development that have smaller factor loading than 0.5, i.e. 0.49 and 0.48. Since it is approaching 0.5, then it is rounded to 0.5. Thus, all instruments have good construct validity. By looking at the model fit, on the third test, all requirements of model fit are met, both the p-value (≥ 0.05), RMSEA (≤ 0.08), and GFI (≥ 0.9). The construct reliability (CR) of all instruments are above 0.7 in all three stages of the test. Thus, based on the three stages of the test, all of the instruments have good construct validity, reliability, and goodness of fit. Those three phases' analyses are presented in Table 4.

In this study, five instruments for measuring the level of early childhood development were developed, namely: instruments for measuring moral-religious, social, language, cognitive, and physical-motor development. The instrument developed is in the form of a questionnaire. Instrument indicators are based on indicators of the level of achievement of early childhood development contained in the Regulation of the Minister of National Education No. 58 of 2009 on the Standard for Early Childhood Education, specifically the standard level of achievement of

Table 3. Recapitulation of Experts and Practitioners Validation

Validator	Validator Number	Average of Score	Qualification
Experts	10	4.10	Good
Practitioners	20	4.29	Good
Total	30	8.40	-
Mean		4.2	Good

Table 4. Summary of Analysis Result for Instrument Measurement Model of Product and Outcome

Instrument	Number of Point	Test No.	Chi-Square		RMSEA	GFI	λ < 0.5	CR
			Score	p-value				
Moral-Religious Development	25	1	308.30	0.07	0.029	0.87	2	0.89
		2	311.77	0.058	0.023	0.91	-	0.91
		3	307.31	0.075	0.019	0.94	-	0.91
Social-Emotional Development	26	1	330.54	0.081	0.027	0.86	-	0.92
		2	333.69	0.07	0.022	0.91	-	0.91
		3	331.38	0.089	0.018	0.93	-	0.92
Language Development	24	1	282.39	0.060	0.030	0.87	-	0.70
		2	276.32	0.089	0.022	0.90	-	0.75
		3	286.48	0.051	0.02	0.94	2	0.82
Cognitive Deevelopment	26	1	331.27	0.066	0.028	0.86	-	0.87
		2	326.05	0.089	0.021	0.91	2	0.80
		3	330.72	0.075	0.018	0.93	1	0.76
Physical-Motor Development	27	1	356.76	0.077	0.027	0.86	4	0.72
		2	351.72	0.094	0.02	0.91	3	0.85
		3	355.86	0.076	0.018	0.93	-	0.82
Life Skills	30	1	439.09	0.081	0.025	0.84	1	0.72
		2	437.05	0.092	0.019	0.90	-	0.76
		3	447.32	0.055	0.018	0.92	2	0.74

development. The procedure for developing this instrument follows five steps, namely: (1) the design preparation phase, (2) the expert validation phase, (3) the testing phase, (4) the data analysis phase, and (5) the implementation phase. The draft instruments that have been compiled are then validated by experts to see indicator depth, formulation of questions or statements, language effectiveness, and others. The experts who validated the instrument consisted of ten peoples, who came from several fields of science, namely: two measurement experts, three evaluation experts, one education management expert, two primary education experts, and two childhood education experts. The goal is that the instrument can be assessed in various aspects, so as to produce a quality instrument.

After being revised based on the FGD input, the instrument was tested to determine the construct validity and reliability. The trial was conducted in three stages, with the number of trial subjects increased. Two assumptions underlie the thinking of why the test was conducted in three stages, namely: (1) increasing variety and the number of trial subjects three times expected to reach all kinds of characteristics, both kindergarten and existing students, and (2) with the representation of all kindergarten characteristics and students, then a good instrument will be obtained, which is an instrument that can be applied to all existing kindergarten.

Based on the results of the test data analysis that conducted from the first stage to the third stage, the following results were obtained. The results of the first phase of the trial show that the five instruments developed were still lacking. After the items points were revised, the second trial was conducted. The results of the second phase of the trial (main trial) show that the fit model of instrument had become better. Only some items of instruments still have deficiency. The items of instrument was revised again and the third trial was conducted. In this study, all the poor instruments have been revised in two stages, the results of the third stage of the test show that all instruments have good fit model, validity, and reliability. Therefore, all instruments developed have a good measurement model, because: (a) all values of $\chi 2$ are low ($p \geq 0.05$), (b) all RMSEA ≤ 0.08, and (c) all GFI values ≥ 0.9. The coefficients of construct reliability (CR) are all above 0.7. Thus, all instruments developed have good quality.

Conclusion

Based on the research findings, two points of conclusion can be drawn. Each point is elaborated as follows. (1) The instruments for measuring the achievement level of early childhood development developed in this research consist of five components: instrument to evaluate the achievement level of moral-religious, social-emotional, cognitive, language, and physical-motor development. (2) According to the assessment of experts and practitioners, the instruments developed have good quality and can be used without any revision. All developed instruments have good validity, reliability, and goodness of fit.

References

Ashiabi, G. S. (2007). Play in the preschool classroom: Its socioemotional significance and the teacher's role in play. *Early Childhood Education Journal*, *35*(2), 199–207. https://doi.org/10.1007/s10643-007-0165-8

Beard, K. Y., & Sugai, G. (2004). First step to success: An early intervention for elementary children at risk for antisocial behavior. *Behavioral Disorders*, *29*(4), 396–409. https://doi.org/10.1177/019874290402900407

Berk, L. E. (2007). *Development through the lifespan* (4th ed.). Boston, MA: Pearson Education.

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). New York, NY: Longman.

Ghozali, I., & Fuad, F. (2008). *Structural Equation Modeling: Teori, konsep, dan aplikasi dengan program Lisrel 8.80.* Semarang: Badan Penerbit Universitas Diponogoro.

Hermawati, I. (2007). *Evaluasi program Pendidikan Anak Usia Dini (PAUD) bagi* *anak dari keluarga miskin di tempat penitipan anak (TPA) Beringharjo, Yogyakarta*. Yogyakarta: Departemen Sosial RI, Badan Pendidikan dan Penelitian Kesejahteraan Sosial, Balai Besar Penelitian dan Pengembangan, Pelayanan Kesejahteraan Sosial.

Hiryanto, H. (2007). Pemetaan tingkat pencapaian mutu program pendidikan anak usia dini (PAUD) di Provinsi DIY. (Laporan penelitian, tidak diterbitkan). Yogyakarta: Lembaga penelitian UNY. *Diklus: Jurnal Pendidikan Luar Sekolah*, *6*(11), 127–149. Retrieved from https://journal.uny.ac.id/index.php/diklus/article/view/5787

Latan, H. (2012). *Structural Equation Modeling: Konsep dan aplikasi menggunakan program Lisrel 8.80*. Bandung: Alfabeta.

Mann, E. A., & Reynolds, A. J. (2006). Early intervention and juvenile delinquency prevention: Evidence from the Chicago longitudinal study. *Social Work Research*, *30*(3), 153–167. https://doi.org/10.1093/swr/30.3.153

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

Ministry of National Education of Republic of Indonesia. (2011). *Petunjuk teknis penyaluran bantuan alat permainan edukatif*. Jakarta: Directorate of Early Childhood Education Development, Ministry of National Education of Republic of Indonesia.

Nugraha, A. (2010). *Evaluasi pembelajaran untuk anak usia dini*. Bandung: Universitas Pendidikan Indonesia.

Regulation of the Minister of National Education No. 58 of 2009 on the Standard for Early Childhood Education. , (2009).

Rushton, S. (2011). Neuroscience, early childhood education and play: We are doing it right! *Early Childhood Education Journal*, *39*(2), 89–94. https://doi.org/10.1007/s10643-011-0447-z

Samuelsson, I. P. (2011). Why we should begin early with ESD: The role of early childhood education. *International Journal of Early Childhood*, *43*(2), 103–118. https://doi.org/10.1007/s13158-011-0034-x

Sudijono, A. (2011). *Pengantar evaluasi pendidikan*. Jakarta: Raja Grafindo Persada.

Valentine, K., Thomson, C., & Antcliff, G. (2009). Early childhood services and support for vulnerable families: Lessons from the Benevolent Society's Partnerships in Early Childhood program. *Australian Journal of Social Issues*, *44*(2), 195–213. https://doi.org/10.1002/j.1839-4655.2009.tb00140.x

Woolfolk, A. (2007). *Educational psychology* (10th ed.). Boston, MA: Allyn & Bacon.

REiD (Research and Evaluation in Education), 6(1), 2020, 10-19
Available online at: http://journal.uny.ac.id/index.php/reid

Curriculum evaluation of French learning in senior high school

[*1]Irma Nur Af'idah; [2]Amat Jaedun
[1]Graduate School, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
[2]Faculty of Engineering, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
[*]Corresponding Author. E-mail: irmanurafidah15.2017@student.uny.ac.id

Submitted: 5 November 2019 | Revised: 20 December 2019 | Accepted: 24 February 2020

Abstract

The research aims to describe the implementation of French language learning in high schools of Sleman Regency viewed from the components of planning, implementation, and results. This evaluation research uses a quantitative descriptive approach with a countenance model from Stake. Respondents in this research were teachers and students at three senior high schools. Data collection techniques used in this study include research lesson plans, questionnaires, and documentation. The results of this research indicate that: (1) in the planning component, the quality of lesson plan preparation is very good and needs to be maintained because in the lesson plan review, the results obtained are 88.9% and the teacher questionnaire results of 26.6 are included in the excellent category; (2) in the implementation component, it has good results with the acquisition of a total score of 77 and a student questionnaire of 66.19; (3) in the component of the results, good results are obtained with an average value of students that is 86.38 and the results of the teacher questionnaire of 65.7 which is above 61 so that it falls into the good category. Student scores are obtained from the results of the middle semester assessment and teacher questionnaire.

***Keywords**: curriculum evaluation, countenance model, French learning*

How to cite: Af'idah, I., & Jaedun, A. (2020). Curriculum evaluation of French learning in senior high school. *REiD (Research and Evaluation in Education), 6*(1), 10-19. doi:https://doi.org/10.21831/reid.v6i1.28006.

Introduction

Evaluation in education is very broad since it includes various activities such as student assessment, measurement, testing, program evaluation, school personnel evaluation, school accreditation, and curriculum evaluation (Anh, 2018, pp. 140–141). Evaluation has an important role in every research as well as in academic studies. Moreover, important points in the evaluation must meet the values that underlie the curriculum, pedagogy, and results which are the main focus in educational values (Lai & Kushner, 2013, p. 24). Evaluation involves conducting research activities by an evaluator to provide information on the subject and object of the evaluation (Johnson & Christensen, 2000, p. 7; McCormick & James, 2019, p. 13). Evaluation is present when an educational process is carried out by the school and when the teacher takes part in the task of parents in educating (Hasan, 2009, p. 3). Evaluation conducted by the teacher to students is done to find out how the abilities and knowledge of students in understanding the subject matter that has been studied to assess, correct, and improve a program systematically (Tyler, 2013, p. 10). From the definition of evaluation, it is found the definition in curriculum evaluation, which is, scientific research conducted systematically to improve the curriculum applied in education.

The curriculum is an activity and learning experience, as well as everything that affects the personal formation of students, both at school and also outside of school for the school's responsibility to achieve educational goals (Arifin, 2011, p. 5). The curriculum as a learning plan is a facility in an educational program that serves as a guide and tool in teaching students. The curriculum aims to achieve a field in a subject that adheres to the categorization in education (Hamalik, 2008). These objectives make the curriculum as a benchmark and foundation in implementing learning in schools.

The curriculum becomes the operationalization of the concept of a curriculum that is still written in the actual form of learning, where learning in the classroom becomes a place to implement and test the curriculum to ensure the implementation of the curriculum in schools goes well (Majid & Rochman, 2014, p. 23). One of the problems of education in Indonesia in the education system is the frequent change of curriculum. Curriculum development as a curriculum based on character and competence to produce a generation that is competent, innovative, productive, creative, and characterless. In the implementation of the curriculum, as the operationalization of the curriculum concept, it is still written in nature which becomes actual in the form of learning, where learning in the classroom becomes a place to implement and test the curriculum to ensure the implementation of the curriculum in schools runs well.

The aforementioned description shows the need for an evaluation of French language curriculum implementation in high school to get information about the readiness, implementation, and results of the French language curriculum. Readiness includes the readiness of books, teachers, infrastructure, and the condition of lesson plans in each school. The implementation includes the process and evaluation of learning French at school, and the implementation results are the learning outcomes of students. The researchers conduct this research on the implementation of the French language curriculum because French is a cross-field study that is attracting students' interests. Moreover, in the French language curriculum implementation, teachers experience constraints in making French language learning plans that are easy for students to understand in terms of material, readiness, and implementation. Therefore, this study is focused on evaluating the implementation of the French language curriculum in senior high school.

The curriculum is a system usually more contained in written form (Hasan, 2009, p. 32). This dimension gains a lot of attention because its form can be seen and easily read and analyzed (Arifin, 2011, p. 9). Thus, the preparation of the curriculum must be in accordance with the components, rules, and structure in the curriculum. As a basic reference in the implementation of education, the curriculum plays an important and strategic role in the progress of a program, especially in the field of education (Kurniawan, Winarno, & Dwiyogo, 2018). The components in the compiled curriculum must contain planning in the learning process and the development of students in the objectives, content, and teaching materials, which must be in accordance with educational objectives (Arifin, 2011, pp. 6–7; Dündar & Merç, 2017, p. 137), so that, later, in the development and implementation of the curriculum in each subject, it will be following the rules and systems in the educational curriculum. It will be realized that all students will achieve academic success only if the curriculum is brought in line with the leadership skills and the education institution implements the right curriculum (Sorenson, Goldsmith, Méndez, & Maxwell, 2011, p. 5), but there are still many data found in the field that plans exist in the curriculum is still not specific and too general, so the curriculum implementers themselves still cannot understand the curriculum well.

Evaluation and curriculum have characteristics and roles in every education and social research (Hasan, 2009, p. 32), so the two components do have quite dominant relationships. The broad curriculum evaluation is not only about activities in the classroom but also a comprehensive assessment process that involves all educational components such as students, teachers, models and methods of teaching, administration, and facilities (Ismail, 2015, p. 15). The purpose of the curriculum itself is to introduce academic discipline to students so that they can use their knowledge with discipline and wisdom (Schiro, 2017, p. 25). The curriculum in education in Indonesia experiences significant changes and developments. This is intended to make the curriculum itself be able to improve learning implemented in schools. The focus on the curriculum is to ensure that the program achieves the mission and goals it has set. The curriculum in high school which was implemented in this decade is the 2013 curriculum.

Changes in the contents of the Education Unit Level Curriculum or *Kurikulum Tingkat Satuan Pendidikan* (KTSP) and 2013 Curriculum in French should be able to make students able to understand the basic learning of French, especially because it changes into cross-interest subjects. However, the reality that occurred in the three schools that have been observed, they actually experience difficulties because the material to be studied is more complex. The teacher feels it difficult in making learning material that is suitable for the ability of students in learning French. The new challenges in the 2013 curriculum become an important lesson that must be completed by the teacher so that students are able to understand French lessons well. In addition, the main element that must be prepared by a teacher before teaching is to prepare approaches, strategies, techniques, and learning procedures so that they can run the teaching effectively (Dewantara, 2017, p. 20). Based on observations that have been made, problems regarding planning, learning, and student assessment results in teaching French are found. Therefore, an evaluation of the 2013 curriculum in French subjects is needed to fit the objectives in the 2013 curriculum. After evaluating the curriculum, the steps that must be taken are knowing how to implement the improved curriculum, whether it has already been referred to as an improvement in learning and the quality of education, or it has not yet been carried out to the maximum.

Based on the background description of the problem that has been described, the evaluation carried out in this research is an evaluation by the Stake countenance model which includes planning, implementation, and results. This research focuses on preparing the learning implementation plan, learning implementation, and the results obtained by students so that later an accurate evaluation can be made in the implementation of learning French. The formulation of the problems found in this research is as follows: how the implementation of the curriculum of French Subjects in high schools in Sleman Regency is viewed from planning, implementation, and learning outcomes. The purpose of this research is to describe the implementation of French language learning in high schools in terms of the planning, implementation, and results components.

Method

The method of this research was curriculum evaluation. In curriculum evaluation, evaluation becomes a main part of the world of education considering the curriculum is always developing and changing according to the context in its era (Hasan, 2009, p. 41). Curriculum evaluation in this research was carried out on the implementation of the French subject curriculum in high school. The evaluation model used was the Stake Countenance model. This model emphasizes two main things, which are drawing and considering. These two main things are obtained through the evaluation stages, they are: (1) the planning stage (antecedent) which includes planning in learning by looking at the readiness of learning in the preparation of lesson plans; (2) the implementation/process (transaction) stage, which was the implementation of French learning in the preliminary, core, and closing activities; (3) the results and assessment phase, namely the measurement of the results of the French learning assessment which includes aspects of attitude, knowledge, and skills and see the suitability of techniques, instruments, and follow-up conducted by the teacher in the implementation of learning French.

Characteristics in countenance evaluation models are evaluating the interrelation (contingency) at each stage and congruence between planning, implementation, and results to reach the consideration stage. Consideration is given to standards/criteria. The planning, implementation, and learning outcomes in this research are based on the Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 22 of 2016. In addition to the Regulation of the Minister of Education and Culture No. 103 of 2014, the results also refer to the Regulation of the Minister of Education and Culture No. 4 of 2018 and the Minimum Completeness Criteria or *Kriteria Ketuntasan Minimal* (KKM). Sources of data/research respondents were students in class X of senior high school. The sampling technique used was a random sampling technique. Random sampling technique is a method of random sampling from members of the population and is taken using a table/number generator (Sarjono & Julianita, 2011, p. 23). The random sampling technique in this research was conducted by selecting two classes in each school. Data collection techniques in this research used the research of lesson plans, observations, questionnaires, and documentation. The questionnaire in this research is the main instrument used in data collection. Likert Scale is a scale used to measure the attitudes, opinions, and perceptions of a person or group of people towards an event or social situation where the variable to be measured is translated into an indicator variable then the indicator is used as a starting point for compiling question/statement items (Sarjono & Julianita, 2011, p. 6). The questionnaire used was a Likert scale with a rating scale of 1-4. There are two types of respondents in the questionnaire namely teachers and students, three teachers, and 145 students from three schools. The data collection technique used in this research is in the form of Lesson Plan research.

The Lesson Plan research was used to find out the planning components that exist in implementing French learning in the three high schools in Sleman Regency where the research was conducted. The documentation used in this research is the value of students used in the results component. This research used content validity and construct validity. The content validity used Aiken validity and the construct validity used Exploratory Factor Analysis with the help of SPSS. In this research, the content validity was carried out by five experts (expert judgment), namely three lecturers who were experts in the field of language learning. The results obtained from 117 items from 71 indicators are that there is one statement that is failed because it does not have relevant relevance so that there are 116 items tested. In conducting trials and research conducted on three teachers, 145 students, and three Lesson Plan, 116 validated items were used.

The construct validity in this research was proven using factor analysis. Factor analysis is a statistical method that is commonly used in the development of measuring tools to analyze the relationship between variables (Azwar, 2018, p. 121). Thus, factor analysis answered the relationship and validity of the items in the instrument. The exploratory factor analysis (EFA) procedure helps develop tests in recognizing and identifying various factors that help construct by finding the largest score variance with the least number of factors expressed in the form of eigenvalue >1. Construct validity according to Nunnally and Fernandes (Retnawati, 2014, pp. 2–3) is validity which shows the extent to which the instrument reveals a certain theoretical ability or construct that is intended to measure. Construct validity is related to the provenience of the measurement result score. The construct validity can be proven by testing that the instrument construct does exist and empirically proven to confirm the existence of the construct of an instrument. The validity test model used was using KMO which is said to be valid if the KMO number is greater than 0.5 and the significance is the senior high school of more than 5%. On the diagonal axis anti-image correlation, all must be greater than 0.5 if there are less than 0.5 then the item is removed (Priyatno, 2009). Factor analysis is used to test the correlation between variables. To test the correlation between variables, the Barlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) test were used. If the results are significant with a KMO value above 0.5, then there is a significant correlation with several variables. The construct validity in this research was used on the student questionnaire with the result that five statements fell out of the 26 items that existed. The final results of the acquisition of KMO and Bartlett's test and the Rotated Component Matrix are as follows. KMO is used to determine whether all data that have been taken are sufficient to be factored measuring the adequacy of the sampling (sampling adequacy). This value compares the magnitude of the observed correlation coefficient with a partial correlation coefficient, a small KMO value indicates that the correlation between pairs of variables cannot be explained by other variables. If the sum of the squares of partial correlation coefficients among all pairs of variables is of small value compared to the sum of the squares of the correlation coefficient, it will produce a KMO value close to 1. The KMO value is considered to be sufficient if more than 0.5. From those results, it can be said that the sampling that has been met can be used for further analysis.

Based on Table 1, the KMO from the SPSS calculation is 0.815, so it is greater than 0.5, and Bartlett's Test is 0.000 so it is said to be good. The conclusion obtained is that the data can be used for further testing. From the results of the calculation of the Rotated Component Matrix, it is known that there are six factors that affect the 21 items with details, namely component/factor 1, that is apperception and preparing a learning plan affecting items 1, 2, 3, 4; component 2, namely core activities affecting items 5, 10, 15, 16, 17, 18, 19; component 3, namely mastering the material taught that influences point 14; component 4, containing the use of media in learning influencing items 5, 6, 7; component 5, regarding asking how the understanding and involvement of students influence points 20, 21; and on factor/component 6 about ending learning influencing points 23, 24, 25, 26.

Instrument reliability in this study was estimated by looking at the Alpha coefficient. Reliability estimation is done by reliability analysis using SPSS program computer ver.22.0 for Windows. To find out the alpha coefficient, the Alpha-Cronbach value for the reliability of all items in one variable was observed. The reliability test is said to be good if it is more than 0.7 (Mardapi, 2017, p. 25). The reliability test results in this study were 0.77 and 1 and more than 0.7. It shows that the student questionnaire reliability is good so that it can be used to test the implementation of the curriculum in the implementation of French language learning in high school.

The analysis technique used in this study is a descriptive statistical analysis technique using the SPSS program through a quantitative approach. It also uses a normal distribution with the following details (Azwar, 2018, p. 148): if the results are said to be not good, if the results obtained are said to be not good, if the results are said to be good, if the results obtained are said to be very good, if it is the average overall score, if it is the standard deviation of the overall score, and if it is the score achieved by students. In the planning category for the teacher questionnaire, if a score of x <12.25 is obtained, the results are said to be not good; if the score is between 12.25-17.74, the results are said to be not good; if the score is between 17.75-22.75, the score is said to be good; and if the score is more than 22.75, the results are stated to be very good. Furthermore, in the implementation category in the teacher questionnaire, if a score of x <41 is obtained, the results are said to be not good; scores between 41-52.00 are said to be not good; scores between 52.01-63.00 are said to be good; and if the score is more than 63.01, the results are stated to be very good. For the implementation category for the students' schedule, if a score of x < 37 is obtained, the results are said to be not good; if the score is between 37-52.75, it is said to be not good; the scores between 52.76 - 68.25 are said to be good; and if the score is

Table 1. KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.815
Bartlett's Test of Sphericity	Approx. Chi-Square	1091.993
	df	210
	Sig.	.000

more than 68.25, it is said to be very good. In the results category for the teacher questionnaire, if a score of x <34 is obtained, the results are said to be not good; a score between 34-42.5 is said to be poor; a score between 42.6-51 is said to be good; and if the score is more than 51, the results are stated to be very good.

Findings and Discussion

Planning in learning is done by using the Lesson Plan. The Lesson Plan is an important component that must be present and made by the teacher before carrying out learning, because it is a plan in describing a teacher in carrying out learning to start learning, giving material, using media, and using assessment instruments that are appropriate to the method and given to students.

In this research, there were three lesson plans analyzed and 37 items in the lesson plan review instrument using a score of 0.0. The calculations in the review of the Lesson Plan are used as the main instrument in planning (antecedent) with Formula (1). The results were analyzed with the planning table criteria (Arikunto, 2018, p. 35) presented in Table 2.

$$\text{Score} = \frac{96}{108} \times 100 = 88{,}9 \text{ .... (1)}$$

Table 2. Lesson Plan Results

Percentage	Result
80– 100 %	
66 – 79 %	
56 – 65 %	88.9%
40 – 55%	
< 40 %	

Table 2 is obtained from the evaluation standard criteria by Arikunto (2018). Descriptive percentages are used to facilitate the analysis of the evaluation of the French language curriculum in high schools based on established standards. The results are then interpreted and presented with numbers at the description stage, not until the generalization stage. Quantitative data analysis using descriptive techniques is used to process data from the questionnaire results obtained that are used to be able to evaluate concerning the techniques used.

From the results of the lesson plan analysis, it is found that it received a presentation score of 88.9%. Then the score is compared with the planning criteria by knowing that the preparation of the Lesson Plan 100% has good results when viewed from the criteria. It can be said that the Lesson Plan of the five schools has a very good suitability of 88.9%. In this planning component, besides using the lesson plan, there is also a teacher questionnaire instrument consisting of seven statement items with the following categorization. The results obtained from the teacher questionnaire in the planning components of preparing lesson plans, designing learning, and evaluating French learning are equal to 25.7, so that it falls into the very good category.

In the component of implementation (transaction) in this research, 45 items of teacher questionnaire and 26 items of student questionnaire were used. The student questionnaire was filled in by 145 respondents, namely students consisting of five schools, namely Depok 1 High School, Kalasan 1 High School, and Angkasa Adisucipto High School, located in Sleman Yogyakarta. Based on the results of the research as a whole, the results of the implementation of French Language Learning in the three schools are included in either category. From the 26 statements of the student questionnaire in the implementation of learning, five statements fall after the Exploratory Factor Analysis (EFA) test using SPSS.

The results of student questionnaire calculations in the implementation of French learning in senior high school obtained an average value of 66.19 so that it is included in the good category. The next aspect is the presentation of student questionnaire results in the implementation component of French learning.

In addition to using student questionnaires, the implementation component also uses a teacher questionnaire instrument which amounts to 21 statements with scores ranging from 1-4, like the student questionnaire. From the calculation of the teacher's questionnaire, a value of 77 is obtained. The results are above 63 so it is included in the very good category.

The results in this research were carried out using the Mid-Semester Assessment, teacher questionnaires, and interviews. The results of this research were obtained by looking at the behavior and assessment results obtained by the teacher with a total of 17 statement items. Based on calculations in the component results from the teacher questionnaire, a value of 65.7 is obtained, so it is included in the very good category.

Planning in evaluating the implementation of the curriculum has an important role so it is known how the preparation of lesson plans and teacher responses in implementing learning that will be done to students, in this case the cross-interest subjects in French. In the planning component, it is measured using the Lesson Plan research instrument and teacher's questionnaire. From the review of the Lesson Plan, it is found that the preparation of the Lesson Plan is known to be very good and the preparation reached 88.9%. Whereas, in the teacher questionnaire, the planning component achieved 93.52% success, so the planning is included in the excellent category.

In the French Lesson Plan, all components meet good requirements in the preparation of the Lesson Plan in line with the syllabus and the Ministry of Education and Culture. Based on research that has been done, the planning component using the main instrument, namely the Lesson Plan review, is supported by a teacher questionnaire that gets very good result. Research that supporting the results of this planning component is found in research by Abrory and Kartowagiran (2014) that planning in preparing lesson plans has been included in the good category, even though the 2013 curriculum has just been applied. Other research that supports the planning component in this study is the study by Lukum (2015) which makes learning plans in the good category so that teachers are known to be able to compile lesson plans well. Another relevant research is conducted by Dewantara (2017) which shows that in planning Indonesian learning, it has been done well and shows the suitability of planning with the standard policy process that is being applied.

The main instrument used in this research is a questionnaire, namely the teacher's questionnaire and student questionnaire. The interview and observation were used as supporting instruments. In evaluating the curriculum, the implementation is a provider of information as an input in decision making (Hasan, 2009, p. 42), then the implementation must meet the criteria to achieve the results and objectives set. The implementation of the French language learning in the three high schools in Sleman Regency obtained good results.

Research supporting the results in this research is a study by Prasojo, Kande, and Mukminin (2018) which state that the implementation of learning is still not in accordance with the standard process because it is hampered by the process of motivating learning, learning media, and identification of students' abilities, even though the results in the questionnaire were already well. Thus, there needs to be a deeper review. Another research relevant to this study is a research by Kurniawan et al. (2018) that the implementation component is good but there are still some components that do not meet the qualifications of the process standard.

In the implementation of learning, one of the main keys to success is the qualification of an educator. Hence, educators who already have a lot of teaching experience still need self-development as lifelong learners and need to open themselves to various educational innovations that can support learning (Sumual & Ali, 2017, p. 348). These studies indicate that many factors affect achievement in the implementation of learning so that all indicators must be reviewed and considered. The outcome component of this research was seen using the teacher questionnaire instrument and supporting instruments using interviews. From the results of teacher questionnaires, it is known that the preparation, reporting, remedial, and follow-up have been done well by the teacher by looking at the results of the grades obtained by students. In this case, the teacher is greatly helped by the assessment criteria that have been deter-mined from the specificity of the specified curriculum, namely the assessment of knowledge, attitude assessment, and skills assessment. Guidelines regarding assessments in learning the 2013 curriculum for high schools are contained in the Regulation of the Minister of Education and Culture No. 4 of 2018. Teacher activities to find out the results obtained by students are conducting assessments, planning follow-up activities in the form of remedial learning, enrichment programs, counseling services, and or assigning assignments groups and individuals in line with student learning outcomes.

Assessment of learning outcomes by educators is inseparable from the learning process. Therefore, the assessment of learning outcomes by educators shows the ability of teachers as professional teachers. The purpose of conducting an assessment according to the Regulation of the Minister of Education and Culture No. 4 of 2018 is to determine the level of mastery of competencies in attitudes, knowledge, and skills that have been and have not been mastered by a/group of students to be improved in remedial learning and enrichment programs and, establish mastery requirements learners' learning competencies in a certain period of time, i.e. daily, midterm, one semester, one year, and the period of research of the education unit, establish improvement or enrichment programs based on competency mastery levels for those identified as learners who are slow or fast in learning and achieving learning outcomes, improving the learning process at the next semester meeting.

In terms of the output component, the implementation of the assessment in learning French as a cross-interest lesson obtains good results by looking at the results of the midterm examination that has been conducted. The value gained by students varies because of the different characters they have. The average score obtained is 86.38 so that the learning carried out has been said to be good because all students have reached the Minimum Completion Criteria or *Kriteria Ketuntasan Minimal* (KKM), with a KKM in this subject that is 75. However, there are still students who have not yet met the KKM in the middle semester assessment because of the different characteristics and abilities of diverse students, even though the teacher has given special treatment.

Research that supports the study in this component is a study by Lukum (2015) which shows that in the components of the students' assessment results reached 65% and is included in the category of sufficient, but still not met in achieving the KKM because there is no match between the planning and implementation of the standard process. The implementation of learning needs to be improved and adjusted again to the standard process. Moreover, this study still has shortcomings in the assessment because the results show that there are still students who have not reached the KKM even though the teacher has done variations in learning to ensure students can understand the subject matter well. Other research in line with this study is by Abrory and Kartowagiran (2014) that the quality of student outcomes has not yet reached maximum results because the value of attitudes, knowledge, and skills has not shown any conformity and achievement in accordance with the planned targets so it can be concluded that in the learning process that has not yet reached perfect results, it is necessary to develop each assessment carried out in learning as in this study. There are previous studies that are relevant to this research, namely research by Sumual and Ali (2017) that a learning outcome is very much determined by the experience and way of the teacher in teaching and giving direction to students. The results in this study are included in the good category because the teacher also has competence in teaching French well. This research has the uniqueness compared to other studies in terms of the planning with a French Language Learning Plan that is adapted, and the learning outcomes of students viewed from the results of the midterm examination that has been carried out to see and evaluate clearly the implementation of the French Language curriculum in high school. The results of the 2013 curriculum implementation are expected to be able to create interesting and meaningful learning for students, especially in French, as a cross-interest lesson that is encouraged by students. For this reason, in implementing the 2013 curriculum in French Language, schools need to continue to encourage the realization of national standards in schools.

Conclusion

Based on the evaluation results of the implementation of the French subject curriculum that have been conducted at senior high school, the following conclusions are drawn: the planning (antecedent) of learning French contained in the Lesson Plan and the teacher questionnaire obtained a very good result; the implementation (transaction) of French learning in the teacher questionnaire and student questionnaire is in a good category. However, in reality, the learning undertaken is still not procedurally and structurally following the Lesson Plan so that the conclusions in the implementation of learning French subjects are included in the good category and need to be improved. The results (outcomes) in French Learning of the teacher's questionnaire are included in the very good category. Hence, overall, French learning is included in the good category and still needs improvement.

References

Abrory, M., & Kartowagiran, B. (2014). Evaluasi implementasi Kurikulum 2013 pada pembelajaran matematika SMP negeri kelas VII di Kabupaten Sleman. *Jurnal Evaluasi Pendidikan*, *2*(1), 50–59. Retrieved from http://journal.student.uny.ac.id/ojs/index.php/jep/article/view/73

Anh, V. T. K. (2018). Evaluation models in educational program: Strengths and weaknesses. *VNU Journal of Foreign Studies*, *34*(2), 140–150. https://doi.org/10.25073/2525-2445/vnufs.4252

Arifin, Z. (2011). *Konsep dan model pengembangan kurikulum*. Bandung: Remaja Rosdakarya.

Arikunto, S. (2018). *Evaluasi program pendidikan*. Jakarta: Rineka Cipta.

Azwar, S. (2018). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.

Dewantara, I. P. M. (2017). Stake evaluation model (Countenance model) in learning process Bahasa Indonesia at Ganesha University of Educational. *International Journal of Language and Literature*, *1*(1), 19–29. https://doi.org/10.23887/ijll.v1i1.9615

Dündar, E., & Merç, A. (2017). A critical review of research on curriculum development and evaluation in ELT. *European Journal of Foreign Language Teaching*, *2*(1), 136–168. https://doi.org/10.5281/zenodo.437574

Hamalik, O. (2008). *Manajemen pengembangan kurikulum*. Bandung: Remaja Rosdakarya.

Hasan, S. H. (2009). *Evaluasi kurikulum*. Bandung: Remaja Rosdakarya.

Ismail, F. (2015). The evaluation of curriculum implementation at Tarbiyah Faculty IAIN Raden Fatah Palembang. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *1*(1), 12–27. https://doi.org/10.21009/JISAE.011.02

Johnson, R. B., & Christensen, L. (2000). *Educational research: Quantitative, qualitative, and mixed approaches*. Thousand Oaks, CA: SAGE Publications.

Kurniawan, R., Winarno, M. E., & Dwiyogo, W. D. (2018). Evaluasi pembelajaran Pendidikan Jasmani, Olahraga, dan Kesehatan pada siswa SMA menggunakan model Countenance. *Jurnal Pendidikan: Teori, Penelitian, Dan Pengembangan*, *3*(10), 1253—1264. https://doi.org/10.17977/jptpp.v3i10.11599

Lai, M., & Kushner, S. (Eds.). (2013). *A developmental and negotiated approach to school self-evaluation*. https://doi.org/10.1108/S1474-7863(2013)14

Lukum, A. (2015). Evaluasi program pembelajaran IPA SMP menggunakan model Countenance Stake. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 25–37. https://doi.org/10.21831/pep.v19i1.4552

Majid, A., & Rochman, C. (2014). *Pendekatan ilmiah dalam implementasi kurikulum 2013* (E. Kuswandi, ed.). Bandung: Remaja Rosdakarya.

Mardapi, D. (2017). *Pengukuran, penilaian, dan evaluasi pendidikan* (2nd ed.). Yogyakarta: Parama Publishing.

McCormick, R., & James, M. (2019). *Curriculum evaluation in schools*. London: Taylor & Francis.

Prasojo, L. D., Kande, F. A., & Mukminin, A. (2018). Evaluasi pelaksanaan standar proses pendidikan pada SMP Negeri di Kabupaten Sleman. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 61–69. https: //doi.org/10.21831/pep.v22i1.19018

Priyatno, D. (2009). *5 Jam belajar olah data dengan SPSS 17* (J. Widiyatmoko, ed.). Yogyakarta: Mediakom.

Regulation of the Minister of Education and Culture No. 103 of 2014 on Lerning in Primary and Secondary Education. , (2014).

Regulation of the Minister of Education and Culture No. 22 of 2016 on the Process Standard of Primary and Secondary Education. , (2016).

Regulation of the Minister of Education and Culture No. 4 of 2018 on the Assessment of Learning Outcomes by the Educational Unit and the Government. , (2018).

Retnawati, H. (2014). *Analisis kuantitatif instrumen penelitian*. Yogyakarta: Parama Publishing.

Sarjono, H., & Julianita, W. (2011). *SPSS vs LISREL: Sebuah pengantar, aplikasi untuk riset*. Jakarta: Salemba Empat.

Schiro, M. S. (2017). *Teori kurikulum: Visi-visi yang saling bertentangan dan kekhawatiran tanpa henti* (B. Sarwiji, ed.; E. Sulistyowati, trans.). Jakarta: Indeks.

Sorenson, R. D., Goldsmith, L. M., Méndez, Z. Y., & Maxwell, K. T. (2011). *The principal's guide to curriculum leadership*. Thousand Oaks, CA: Corwin Press.

Sumual, M. Z. I., & Ali, M. (2017). Evaluation of primary school teachers' pedagogical competence in implementing curriculum. *Journal of Education and Learning*, *11*(3), 343–350.

Tyler, R. W. (2013). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

REiD (Research and Evaluation in Education), 6(1), 2020, 20-31
Available online at: http://journal.uny.ac.id/index.php/reid

Testing the influence of SDES instructional media on the results of cryptography learning

[*1]Soeprijanto; [1]Aodah Diamah; [1]Prasetyo Wibowo Yunanto
[1]Faculty of Engineering, Universitas Negeri Jakarta
Jl. Rawamangun Muka, Rawamangun, Pulo Gadung, Kota Jakarta Timur, DKI Jakarta 13220, Indonesia
[*]Corresponding Author. E-mail: soeprijanto@unj.ac.id

Submitted: 6 February 2020 | Revised: 2 April 2020 | Accepted: 6 April 2020

Abstract

This study aims to examine the influence of Simplified Data Encryption System (SDES) simulation on student learning outcomes in Cryptography lessons. The research employed a quasi-experiment. Data analysis to test the SDES simulation model was performed using ANOVA 2x2. The U-Mann Whitney Test was chosen to examine differences in student learning outcomes of treatment groups and control groups, while the effectiveness of the media is determined by differences in student learning outcomes between the pre-test and post-test results in the two groups. The test results show that: (1) There is a difference between the treatment group and control group, indicated by the U-Mann Whitney Test result ($U_{count} = 15 < U_{table} = 23$; $\alpha = 0.05$), which means there is a difference of student learning outcome between students given learning by DES simulation media and those by PowerPoints Media. (2) There is a difference in the cryptography learning outcomes for the students with the high initial ability between the treatment group and the control group. The test result is $U_{count} = 0.5 < U_{table} = 2$; at $\alpha = 0.05$. (3) There is no difference in student learning outcomes for low initial ability student groups using the DES simulation media, with high ability students group using PowerPoints Media; the statistical test results show $U_{count} = 11 > U_{table} = 2$; at $\alpha = 0.05$. This study concludes that using U-Mann Whitney, it can prove that the SDES simulation model developed is effective for improving student learning outcomes in Cryptography lessons.

Keywords: *SDES, student learning outcomes, cryptography lessons*

How to cite: Soeprijanto, S., Diamah, A., & Yunanto, P. (2020). Testing the influence of SDES instructional media on the results of cryptography learning. *REiD (Research and Evaluation in Education), 6*(1), 20-31. doi:https://doi.org/10.21831/reid.v6i1.30024.

Introduction

Learning media plays an important role in achieving learning objectives by providing an opportunity for teachers to develop students' knowledge, motivation, and classroom engagement. One of the learning media is a computer simulation that can be used by teachers for developing students' conceptual understanding. Computer simulation can facilitate students to develop knowledge and construct their understanding of the topics. In the computer simulation, students can repeat and explore the process to understand the concepts.

Researchers have successfully developed a Data Encryption System (DES) simulations using effective simulation design principles and avoiding cognitive overload on students. Excess Simulation DES development results include: (1) attention cueing to make it easier for students to focus and understand the simulations presented; (2) navigation and control feature to enable students to control the simulation; and (3) only use dynamic visualization if necessary. After all of the activities during the assessment, analysis, design, and development, are completed, then we are ready for summative evaluation, to judge the effectiveness of the solution.

Data Encryption Standard (DES) and Simplified Data Encryption Standard (SDES) are designed to assist students in learning of modern cryptoanalytic techniques. Properties and structure in SDES are similar to those in DES, but SDES is simpler and makes students easier for encryption and decryption by hand with a pencil and paper. The simplified DES is designed only for educational purposes. Learning SDES provides insights on DES and other block ciphers and insights on various cryptanalytic approaches. Four levels of evaluation are identified, including reaction, learning, behavior, and results, to see the effectiveness of media influence of DES simulation of the development result on students' learning outcomes.

The research employed experimental research to prove the improvement of student learning outcomes. The problem arises when the number of students who joined the course of Cryptography as a respondent is limited, so the data obtained during potential research is not normally distributed. As an alternative data analysis solution is no longer conducted with parametric statistics, instead, to prove differences in student learning outcomes between treatment groups and control groups, Non-Parametric Statistics was used.

The Mann−Whitney U-test and the Kolmogorov−Smirnov two-sample test are non-parametric statistical procedures for comparing two independent samples. The parametric equivalent to these tests is the t-test for independent samples.

This research problem includes: (1) whether through the characteristic of the U-Mann Whitney Test, the differences in student learning outcomes between treatment groups and control groups can be demonstrated; (2) whether there is a difference in student learning outcomes of the group of students who have low initial ability given cryptographic learning with DES simulation media and a group of high initial ability students who were given cryptographic learning with PowerPoint media; (3) whether there is any influence of using DES simulation to the student learning outcomes of Cryptography. Meanwhile, the novelty of this study is the existence of a solution to the testing of educational media toward a relatively small number of student samples where the data obtained are not normally distributed and to support the learning process effectively in the cryptography course. Thus, this study aims to examine the influence of Simplified Data Encryption System (SDES) simulation on student learning outcomes in Cryptography lessons.

Cryptographic Learning Outcomes

Learning outcomes are abilities obtained by individuals to get learning experiences. According to Briggs (1979, p. 149), learning outcomes are all competencies that are obtained through the learning process.

Further, Bloom, Englehart, Hill, Furst, and Krathwohl (1978, p. 7) state that learning outcomes can be classified into three domains: cognitive, affective, and psychomotor domains. Gagné (1983, pp. 27–28) believes that learning outcomes are competencies that include verbal information, intellectual skills, motor skills, attitudes, and cognitive strategies and values. Verbal information and cognitive skills are students' knowledge or understanding of theory, while motor skills are students' skills, and attitudes are the values of student work, all of that as learning outcomes. Based on the aforementioned opinions, in this study, learning outcomes are defined as individual competencies including knowledge, skills, and attitudes obtained by students through the learning process.

Cryptography comes from the word *Crypto* which means secret, and *Graphy* which means writing (Sasongko, 2005, p. 160). Ariyus in Pratama and Latifah (2014, p. 19) asserts that in general, cryptography consists of three important main parts, namely, the encryption section, the description, and the key sections. The encryption algorithm is a function used to perform encryption and decryption work. According to Kromodimoeljo (2010, p. 5), the encryption technique is a way in which the original text is changed using an encryption key into a random script that is difficult to read by someone who does not have a decryption key to decrypt the key using the so-called "decryption key" in order to get the original data back.

In modern cryptography, there are various kinds of algorithms that are intended to secure information sent over a computer network. According to Insights for Professionals (IFP) (2018, p. 1), modern cryptographic algorithms consist of three parts: (1) Symmetric Algorithm, (2) Asymmetric Algorithm, and (3) Hybrid Algorithm. Symmetric algorithm is an algorithm that uses the same key for encryption and description. The application of the symmetric algorithm is used by several prayer algorithms, one of which is the Data Encryption Standard (DES).

Based on the aforementioned studies, it can be concluded that Cryptographic learning outcomes are students' knowledge and skills towards data encryption and description techniques, as well as attitudes obtained by students through the learning process of cryptography. Operationally, the learning outcomes measured in this study are only the knowledge and skills of students about cryptography, while the aspects of attitude are not measured.

SDES Instructional Media

Levie and Lentz in Arsyad (2016) suggest four functions of instructional media, especially visual media, namely: (1) attention function, which sees that visual media is interesting and directs the students' attention to concentrate on the content of the lesson; (2) affective function, i.e. the visual media seen in student's enjoyment when studying; (3) cognitive function, i.e. the visual or image symbols that facilitate the learning outcome of goals for understanding and remembering information; (4) the compensatory function, that is, to provide a context for understanding the text and help the weak student in reading to organize the information in the text and recall it. Thus, the use of media in the learning process can generate new desires and interests, ease in remembering information, and assist students in organizing and recalling text lesson material that will ultimately affect student learning outcomes. Cheung (2009, p. 9) states that media production goes beyond mere comprehension and analysis in Bloom's taxonomy. Those involved in media production have to include the production of meaning and design using a range of symbol systems in evaluating the availability of a wide range of media resources.

Media production is not just mere understanding and analysis as in Bloom's taxonomy. Media production must include meaning and design using various symbol systems in evaluating the availability of various media resources. The effectiveness of media influences can be determined at least with two criteria, namely, (1) the difference in the mean of a result of student learning when compared with other media usage, and (2) an increase in average student learning outcomes when the learning media is used. Moreover, Lee and Owens (2004, p. 162) insist that successful multimedia development methodologies tend to include these elements: (1) *Design-time prototyping:* creating an early application-system prototype so as to review, test, and approve the interface design, media elements, script, or map. This is an efficient method for rapid development. (2) *Evolutionary development:* using each stage of prototyping and development as the basis from which to evolve the next prototype. For this to be successful, design decisions that do not involve the content must be locked in. (3) *The use of rapid development tools (RDT):* templates are useful for parallel development projects. They are particularly useful in projects where content is added in an iterative process, as it is made available. Templates are created and used as a framework for content as it is identified. Computer simulation design to support the learning process effectively should consider several factors, one of which, according to Plass, Homer, and Hayward (2009), is a control and navigation feature that allows students to simulate Plass et al's opinion is in line with the arguments of Hennessy et al. (2007) and Windschitl (1998). Controls that allow students to stop, repeat, or manage speed simulations, facilitate them to consolidate what they are learning. Another factor to consider is the cognitive load that students will experience when running the simulation. The information that is dynamically visualized in the simulation according to

Plass et al. (2009) requires a more severe mental process than information which is presented in the static form, however, properly designed dynamic visuals can help students learn more effectively. Cueing is one of the effective ways examined by de Koning, Tabbers, Rikers, and Paas (2007). Cueing in dynamic visualization can help students focus on specific processes they need to understand. According to de Koning et al., cueing in visualization can be a color or arrow that guides the students to an important aspect of the simulation. Associated with cognitive loads, according to Höffler and Leutner (2007), dynamic visualization will only be more effective than static visualization if its nature does represent the process to be studied and not merely decorative. Data Encryption Standard (DES) is one of the topics in Cryptography Courses. DES is originally designed to be implemented only in hardware systems and is, therefore, extremely slow in software applications (Rabah, 2005, p. 312). DES is a symmetric-key algorithm for the encryption of digital data. Compared to classical cryptographic algorithms, DES includes complex and elusive algorithms. DES was originally designed by IBM before it became the standard set by the National Institute of Standards and Technology in 1977. Technically, the DES algorithm was resolved when published a scientific article containing an analysis for brute-force attack DES (Biham & Shamir, 1991, p. 4). However, at that time to carry out the attacks proposed by Biham and Shamir, it takes a lot of plaintexts so that the attack is not practical to do. When the computer becomes faster, the attack becomes possible and triple-DES and AES finally appear in place of DES. Nevertheless, DES remains widely used (Burr, 2006). In addition, DES is an important algorithm studied due to the basis of the triple-DES algorithm and its AES continuation algorithm. Due to the long process of DES, a simplified version of the DES called Simplified Data Encryption System (SDES). Cohen (2007, p. 14) believes that SDES was developed by Professor Edward Schaefer of Santa Clara University. The SDES algorithm is instructive and is not a secure encryption algorithm. As seen in Figure 1 and Figure 2, SDES has a process and structure similar to DES, but all the parameters have been made as simple as possible. For example, 16 rounds on DES are simplified into two rounds. According to Schaefer, with simpler structures and parameters, SDES will be more easily understood by undergraduate students.

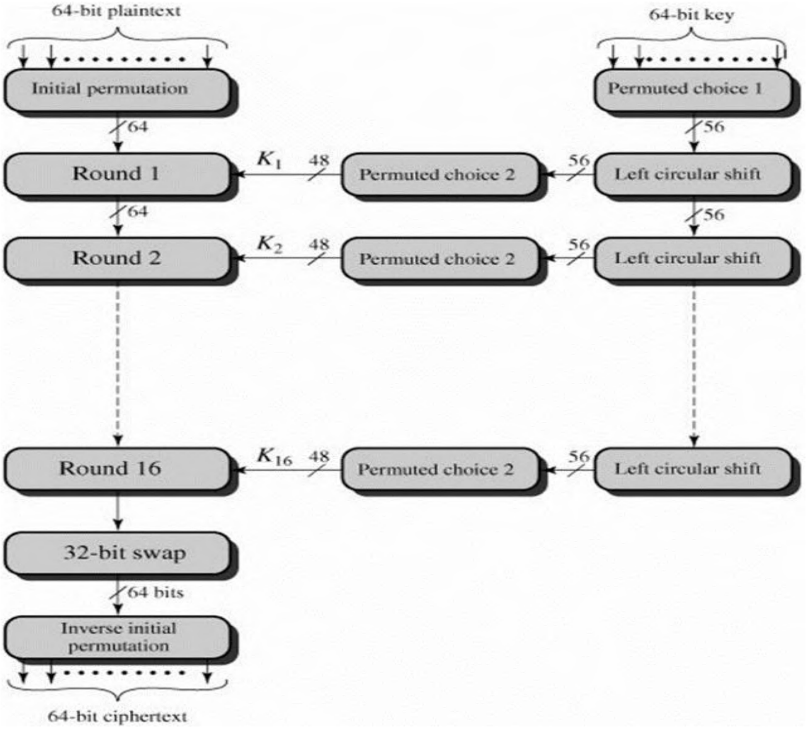


Figure 1. Data Encryption Standard (DES) (Stallings, 2002)

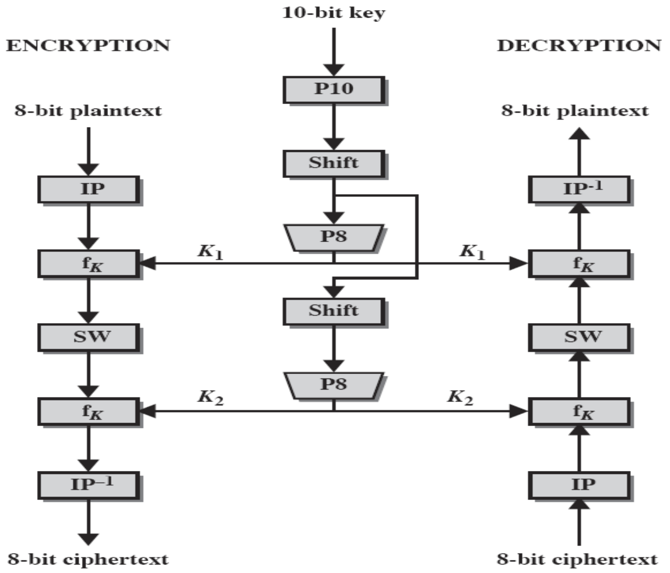


Figure 2. Structure of Simplified Data Encryption Standard (SDES) (Stallings, 2002)

The comparison between DES and SDES can be seen in Table 1. The purpose of SDES for education is that students can more easily learn about modern cryptanalytic techniques. Table 1 shows the differences between DES and SDES. SDES is similar to DES but has simpler properties and structures that are easier to understand.

Table 1. Comparison of DES and SDES

	DES	SDES
Key	64 Bit	10 Bit
Sub Key	56 Bit	8 Bit
Plain Text Processed	64 Bit	10 Bit
Number of Rounds	16	2

The Mann-Whitney (U-test)

Corder and Foreman (2014, pp. 69–70) explain that the Mann−Whitney U-test is non-parametric statistical procedures for comparing two samples that are independent, or not related. The parametric equivalent to these tests is the t-test for independent samples. Mann−Whitney U-test is used to compare two unrelated, or independent, samples. The two samples are combined and rank-ordered together. The strategy is to determine if the values from the two samples are randomly mixed in the rank-ordering or if they are clustered at opposite ends when combined. A random rank-ordered would mean that the two samples are not different, while a cluster of one sample's values would indicate a difference between them. According to Ho (2014, p. 518), the Mann-Whitney test is a non-parametric statistic used to find out whether there are differences in responses from two independent data populations when data are weaker than the interval scale. This test can be likened to a t-test test for two independent groups when a violation of the assumption of normality or data scale is not appropriate for the t-test. From Corder and Foreman (2014), Ho (2014), and Berry, Mielke Jr., and Johnston (2012, pp. 9–11), we can conclude that the U-Mann Whitney has a characteristic as an alternative test to the independent group t-test when the assumption of normality is not met. Formula (1) is used to determine a Mann-Whitney U-test statistic for each of the two samples (Corder & Foreman, 2014, p. 70). The smaller two U statistics is the obtained value:

$$Ui = n_1.n_2 + \frac{ni(ni+1)}{2} - \sum R_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Annotation:
U_i is the test statistic for the sample of interest,
n_i is the number of values from the sample of interest,
n_1 and n_2 are the numbers of values from the first and second sample,
Σ_{Ri} is the sum of the ranks from the sample of interest.

According Susetyo (2017, p. 236), the level of significance uses α =0.05, and rejection criteria H_0 for one side if $U_{count} \leq U_{table}$ formulated at an opportunity value (p) compared to the specified real level. U_{price} is selected as the smallest value from the calculation results in each group.

Method

This study utilizes the characteristic of U-Mann Whitney's Test to prove the effectiveness of learning media influence SDES simulation development result in learning cryptography. The research was conducted in Informatics and Computer Technology Education Study Program, Faculty of Engineering, in Jakarta. The study was conducted in the even semester of the academic year 2016/ 2017. The research method used to test the DES Simulation model is a quasi-experimental method with treatment by level. The quasi-experiment method is meant to see the causal relationship between two factors deliberately caused by the researcher by eliminating other disturbing factors. This research uses treatment design by level 2x2 because there are two types of treatment on independent variables. This study also controls two attribute variables consisting of two levels. The variable has the potential to affect the dependent variable. Experimental design treatment by level 2 x 2 is described in Table 2.

The variables studied consist of the independent variable and bound variable. The dependent variable is the student learning outcome of the independent variable consisting of one active variable and one attribute variable. The pre-test is considered an attribute, while the active variables in the form of learning by using PowerPoints (PPT) media. The hypothesis was tested with two levels. This variable has the potential to affect the dependent variable. Experimental ANOVA model requires sample data requirement that is used come from a population that has normal and homogenous distribution, so, before data analysis is done or hypothesis testing is done, Normality and Homogeneity need to be tested first. When it is not normal, the use of Parametric Statistics cannot proceed. In this research, the experiment was conducted by the steps illustrated in Figure 3.

Table 2. Design of ANOVA Experiments (2x2)

Level	SDES Simulation Media (Treatment Group)	PowerPoint Media (Control Group)
High initial ability (Pre-Test)	A_1B_1	A_2B_1
Low initial ability (Pre-Test)	A_1B_2	A_2B_2

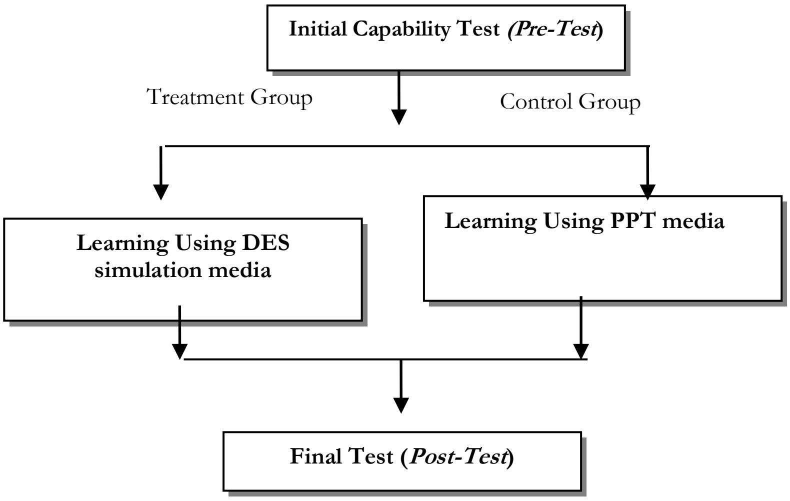


Figure 3. Experiment Steps

Table 3. Distribution of Research Respondents

Level of Ability	SDES (Treatment Group)	PPT (Control Group)	Total
High Ability Pre-Test Results	5	5	10
Low Ability Pre-Test Results	5	5	10
Total	10	10	20

Table 4. Data of Pre-Test and Post-Test Results

	Mean	Median	Modus	Deviation Standard
Pre-Test	55.70	53.00	44.00	14.40
Post-Test (Treatment Group)	87.60	80.00	80.00	0.00
Post-Test (Control Group)	65.00	65.00	65.00	9.78

Table 5. The Value of *Lilliefors*

Group	L_{count}	L_{table}	Conclusion
Simulation DES(Treatment)	0.6895	0.2580	Abnormal
Power Points (Control)	0.0763	0.2580	Normal
Treatment High Level Pre-Test	0.6134	0.3370	Abnormal
Treatment Low Level Pre-Test	0.5707	0.3370	Abnormal
Control High Level	0.6188	0.3370	Abnormal
Control Low Level	0.5870.	0.3370	Abnormal

Each step in Figure 3 is elaborated as follows. (1) A pre-test is the initial test performed using an objective test in the form of multiple-choice questions. Preliminary test results were used to divide the respondents into two groups: high initial ability group and low initial ability group. Based on the results of the grouping of respondents at each level, 50% taken to be treated as the Treatment group is taught using DES simulation media, while others are given lessons by using PowerPoint media. (2) In the experimental step, the participants were divided into two groups, namely, the treatment group and the control group. Each group consists of students who have low initial ability and who have high initial ability. Treatment Group was taught by using DES Simulation Media. The control group was taught using PowerPoint media. (3) In the provision of Post-Test, the final test is done with the same problem as the initial stage. The final test result is used to test the research hypothesis.

The sample size is 20 students. The sample distribution in each group is presented in Table 3.

Findings and Discussion

The research data are presented under the form of a summary of information, including the minimum, maximum, mean, or median, standard, deviation, variance, and theoretical ranges of each variable. This research data are obtained from 20 respondents. Data of the research results consist of initial ability and result data of the Post-Test. The description of the research results for each complete variable can be seen in Table 4.

Test Requirements Analysis

Before the data analysis was carried out to test the hypothesis, the analysis requirements need to be tested first. One of the tests is the normality test. The normality test was performed using the Lilliefors test of the null hypothesis which states that the sample originated from a normally distributed population versus an alternative hypothesis states that the sample is from a population that is not normally distributed. The calculation value of Lilliefors is presented in Table 5.

From the calculation of the value of Lilliefors, it turns out that from almost all of the groups tested, the data are not normal. Only one variable has normal data, namely, on the control group student learning outcomes. On that basis, the researchers decided that ANOVA analysis cannot proceed. Instead, non-parametric statistics are used to prove the difference in student learning outcomes between the treatment group and the control group. The statistic used to test the hypothesis is the U-Mann Whitney Test. U-Mann Whitney Test can be equated with a t-test for two independent groups drawn from one population-scale lower than interval and assumption of the distribution of sample normality (Ho, 2014).

Hypothesis Testing

Hypothesis was tested using a formula previously presented in Formula (1) to determine the Mann−Whitney U-test statistic for each of the two samples. Meanwhile, the value of U_2 is calculated by the formula $U_2 = n_1.n_2\text{-}U_1$. The level of significance uses $\alpha = 0.05$, while the criteria rejection H_0 if the U_{value} of the calculated result is less than the value of U_{table} at probability 0.95 or at $\alpha = 0.05$. According Susetyo (2017, p. 236), the level of significance uses $\alpha = 0.05$ and rejection citeria H_0 for one side if $U_{count} \leq U_{table}$ formulated at an opportunity value (p) compared to the specified real level.

The First Hypothesis

The first hypothesis is elaborated as follows. H_0: there is no difference in student learning outcomes between the treatment group and the control group. H_1: there are differences in student learning outcomes between the treatment group and the control group.

Table 6. Data of Students' Learning Outcomes

Treatment Group	Rank	Control Group	Rank
95	20	75	13.5
85	19	70	10.5
75	13.5	70	10.5
80	16.5	65	7.5
80	16.5	65	7.5
70	12	55	3
80	16.5	60	4.5
50	2	40	1
65	7.5	60	4.5
80	16.5	65	7.5
Total Rank	**140**		**70**

The data on the students' learning outcomes are presented in Table 6. From Table 6, the value of $R_2 = 70$. When this value is entered to Formula (1), the value of U_1 obtained is elaborated as illustrated in Formula (2):

$$U_1 = 10.10 + \frac{10(10+1)}{2} - 70 = 85 \ldots\ldots\ldots\ldots (2)$$

Meanwhile, the value of $U_2 = n_1.n_2 - U_1 = 10.10 - 85 = 100 - 85 = 15$. The calculation results obtained value U arithmetic of 15 When this value is confirmed in table U for $n_1 = 10$ and $n_2 = 10$, $\alpha = 0.05$ got U table value of 23. Thus, statistical test results prove that $U_{count} < U_{table}$ (15<23). It is concluded that H_0 is rejected and H_1 is accepted. It means that there is a difference between the treatment group and the control group (see Table 7). The result of the calculation of the mean obtained that the mean to the treatment group is 87.5. This value is higher than the mean of the control group amounted to 65. It shows that the students' learning outcomes of the group of students who were given learning from the DES simulation media are higher than the students who were given the learning by using PowerPoints (PPT) media.

The Second Hypothesis

The second hypothesis is elaborated as follows. H_0: there is no difference in student learning outcomes of cryptography between those using SDES simulation media on high-ability cryptography (A1B1) and high-ability student group learning cryptography using PowerPoint media (A2B1). H_1: there is a difference in student learning outcomes of cryptography between those using DES simulation media on high ability students (A1B1), and those using PowerPoint media on the high-ability students (A2B1).

This hypothesis was tested in two stages. The first stage is to test the significance of differences in student learning outcomes between the treatment group and control group with the U-Mann Whitney Test. The second stage is to compare the mean values of both.

The result of the difference test of student learning outcomes at the students with high initial ability obtained U_{table} price with probability of 0.95 (U-α) or α (0.05) with the sample number 1 (n_1) and number 2 respectively = 5 and 2. The value U calculation result is 0.5, so the statistical test results $U_{count} < U_{table}$ (0.5<2). Thus, it is concluded that H_0 is rejected and H_1 is accepted (see Table 8).

Table 7. Hypothesis Test Results 1

Hypothesis 1	**U_{count} = 15**	**U_{table} = 23**	**H_0 Rejected**	**H_1 Accepted**
Mean Value	$\overline{A1}$ =87.5	$\overline{A2}$ =65		
Conclusion	There is a difference in post-test results between student learning outcomes with SDES (A_1) and with the result of learning with PowerPoint (A_2)			

Table 8. Hypothesis Test Result 2

Hypothesis 2	**U_{count} = 0.5**	**U_{table} = 2**	**H_0 Rejected**	**H_1 Accepted**
Mean Value	$\overline{A1B1}$ = 83	$\overline{A2B1}$ = 70		
Conclusion:	There is a difference between student learning outcomes with DES simulation media and student learning outcomes with PowerPoint media, in the both groups of high ability students			

Table 9. Hypothesis Test Result 3

Hypothesis 3	**U_{count} = 11**	**U_{table} = 2**	**H_0 Accepted**	**H_1 Rejected**
Mean Value	$\overline{A1B2}$ =69.00	$\overline{A2B1}$ =70.00		
Conclusion:	There is no difference in student learning outcomes between a group of low-level students who learn cryptography to using DES simulation media and high-ability student group learning cryptography using PowerPoint (PPT)			

The Third Hypothesis

The third hypothesis is elaborated as follows. H_0: there is no difference in student learning outcomes between low-grade students who learned cryptography using SDES (A1B2) simulation media and high-ability student group learning cryptography using PowerPoint media (A2B1). H_1: there is a difference in student learning outcomes between a group of low-performing treatment (A1B2) and a high initial-ability control group (A2B1).

Through the U Mann Whitney Test, U_{table} price is obtained with probability 0.95 (U-α) or at α (0.05) with sample number 1 (n_1) and sample 2 (n_2) respectively = 5 and 2. The value of U calculation result is 11, so the statistical test results is $U_{count} < U_{table}$ (11>2). Thus, it can be concluded that H_0 is accepted and H_1 is rejected, meaning that there is no difference in student learning outcome in the low-skilled student group treated with learning using DES simulation media (A1B2) and student learning outcome in the group of high-ability students who are not treated (control group) (A2B1), as presented in Table 9.

Viewed from the average indigo obtained, it shows that the mean of the Treatment Group is 69 and the control group's average rating is 70. Therefore, it can be concluded that the student learning outcomes of low-skilled students who were given lessons with DES simulation media are more comparable than the high initial ability students who were given learning using PowerPoint (PPT) media.

When examined thoroughly from testing Hypotheses 1 to 3, the influence of learning media DES simulation on the development of students' cryptography learning results can be proven. More detail information is presented in Table 10.

Other findings through the first hypothesis calculation through U-Test also prove that the learning outcomes of the treatment group students differ from the learning outcomes of the control group students. In other words, the results of the students who were given the lesson of cryptography using DES simulation media and those who were given the lesson of cryptography using PowerPoint media is different.

This finding is also supported by the fact that the average post-test result from the treatment group reached 87.5 is much higher than the average over the control group's post-test outcome of 65. From this first hypothesis, it also shows that the U test results also correspond with the result of the student's average grade.

In line with the findings of the second hypothesis that tested the use of DES simula-

Table 10. U-test Value, Mean Comparison, and Conclusion

Hypothesis	U-Test Value	Conclusion	Findings	Mean Comparison
1	$U_{count}=15<U_{table}=23$	H_0 rejected H_1 accepted	There is a difference in student learning outcomes of cryptography between those using SDES simulation media and those using PowerPoint on all samples	87.5 : 65
2	$U_{count}=0.5<U_{table}=2$	H_0 rejected H_1 accepted	There is a difference in the student learning outcomes of studying cryptography on student with high ability between those who were taught using SDES simulation media and those taught using PowerPoint media	83 : 70
3	$U_{count}=11>U_{table}=2$	H_0 accepted H_1 rejected	There is no difference in the students' cryptography learning outcomes between the group of low ability students when learning by SDES simulation media, and a group of high-ability students learning by PowerPoint.	69 : 70

tion media in the group of high-ability students, a match of the U-Mann Whitney statistical test results with the mean of each test group is also found, where the second hypothesis proves that there is a difference between the students' learning outcomes using DES simulation media and the students' learning outcomes using PowerPoint media in a group of high-ability students, by comparison of the mean of 83.00 compared to 70.00.

Thus, the first and second problems raised in this study were answered that this study proves that the U-Mann Whitney Test can prove differences in student learning outcomes between treatment group, i.e. groups of students given learning by DES simulation media and control group, i.e. students given learning using PowerPoint. The result of statistical analysis to prove the fourth hypothesis also explains at the same time answer the third problem in this research. The results of the analysis prove that DES simulation media is an effective development result for use as a medium in teaching cryptography. It is shown that the learning outcomes of students with low initial ability can be increased so that they are not different from the high-ability students given learning cryptography using conventional media (PowerPoint). Hence, the developed DES simulation works well and can be recommended as an alternative media for cryptographic learning, especially in achieving the competence of DES mastery goals.

Conclusion

Through the U-Mann Whitney Test, it is proven that there are differences in the result of cryptography learning between students taught using SDES simulation media and those taught using PowerPoint. Then, the learning media obtained from the developed DES simulation works well and improves the students' learning in cryptography.

According to the research findings and discussion, several conclusions are drawn as follows. (1) There is a difference in terms of the post-test results between the learning outcomes of students taught using SDES (A1) and the learning outcome of students taught using PowerPoint. (2) There is a difference between the learning outcomes of students taught using DES simulation media and the learning outcomes of those taught using PowerPoint media in both groups of high ability students. (3) There is no difference in terms of the learning outcomes between the group of low-ability students who learn cryptography using DES simulation media and the group of high-ability students learning cryptography using PowerPoint (PPT).

References

Arsyad, A. (2016). *Media pembelajaran*. Jakarta: Raja Grafindo Persada.

Berry, K. J., Mielke Jr., P. W., & Johnston, J. E. (2012). The two-sample rank-sum test: Early development. *Electronic Journ@l for History of Probability and Statistics*, *8*(December), 1–26. Retrieved from http://www.jehps.net/decembre2012/BerryMielkeJohnston.pdf

Biham, E., & Shamir, A. (1991). Differential cryptanalysis of DES-like crypto-systems. *Journal of Cryptology*, *4*(1), 3–72. https://doi.org/10.1007/BF00630563

Bloom, B. S., Englehart, M. D., Hill, W. H., Furst, E. J., & Krathwohl, D. R. (1978). *Taxonomy of educational objectives - The classification of educational goals; Handbook I: Cognitive domain*. New York, NY: David McKay.

Briggs, L. J. (1979). *Instructional design: Principles and applications*. Englewood Cliffs, NJ: Prentice Hall.

Burr, W. (2006). Cryptographic hash standards: Where do we go from here? *IEEE Security & Privacy*, *4*(2), 88–91.

Cheung, C.-K. (Ed.). (2009). *Media education in Asia*. Dordrecht: Springer Netherlands.

Cohen, A. E. (2007). *Architectures for cryptography accelerators* (Doctoral thesis, University of Minnesota, Minneapolis, MN.). Retrieved from https://pqdtopen.proquest.com/doc/304824471.html?FMT=ABS

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2007). Attention cueing as a means to enhance learning from an animation. *Applied Cognitive Psychology*, *21*(6), 731–746. https://doi.org/10.1002/acp.1346

Gagné, R. M. (1983). *The conditions of learning*. New York, NY: Holt, Rinehart and Winston.

Hennessy, S., Wishart, J., Whitelock, D., Deaney, R., Brawn, R., Velle, L. la, … Winterbottom, M. (2007). Pedagogical approaches for technology-integrated science teaching. *Computers & Education*, *48*(1), 137–152. https://doi.org/10.1016/j.compedu.2006.02.004

Ho, R. (2014). *Handbook of univariate and multivariate data analysis with IBM SPSS* (2nd ed.). Boca Raton, FL: CRC Press.

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, *17*(6), 722–738. https://doi.org/10.1016/j.learninstruc.2007.09.013

Insights for Professionals (IFP). (2018). 3 Types of encryption to protect your data. Retrieved from Tech Insights for Professionals website: https://www.insightsforprofessionals.com/it/security/types-of-encryption-protect-your-data

Kromodimoeljo, S. (2010). *Teori dan aplikasi kriptografi*. Sunnyvale, CA: SPK IT Consulting.

Lee, W. W., & Owens, D. L. (2004). *Multimedia-based instructional design: Computer-based training, web-based training, distance broadcast training, performance-based solutions* (2nd ed.). San Francisco, CA: John Wiley & Sons.

Plass, J. L., Homer, B. D., & Hayward, E. O. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing in Higher Education*, *21*(1), 31–61. https://doi.org/10.1007/s12528-009-9011-x

Pratama, R. K. P., & Latifah, F. (2014). Implementasi enkripsi dekripsi pesan teks menggunakan model Julis Caesar berbasis Object Oriented Programme. *Jurnal Techno Nusa Mandiri*, *XI*(1), 17–26. https://doi.org/10.33480/techno.v11i1.167

Rabah, K. (2005). Theory and implementation of Data Encryption Standard: A review. *Information Technology Journal*, *4*(4), 307–325. https://doi.org/10.3923/itj.2005.307.325

Sasongko, J. (2005). Pengamanan data informasi menggunakan kriptografi klasik. *Dinamik*, *10*(3), 160–167. Retrieved from https://www.unisbank.

ac.id/ojs/index.php/fti1/article/view/25

Stallings, W. (2002). The advanced encryption standard. *Cryptologia*, *26*(3), 165–188. https://doi.org/10.1080/0161-110291890876

Susetyo, B. (2017). *Statistik untuk analisis data penelitian*. Bandung: Refika Aditama.

Windschitl, M. A. (1998). A practical guide for incorporating computer-based simulations into science instruction. *The American Biology Teacher*, *60*(2), 92–97. https://doi.org/10.2307/4450426

REiD (Research and Evaluation in Education), 6(1), 2020, 32-40
Available online at: http://journal.uny.ac.id/index.php/reid

Alternative item selection strategies for improving test security in computerized adaptive testing of the algorithm

***Iwan Suhardi**
Faculty of Engineering, Universitas Negeri Makassar
Jl. Daeng Tata Raya, Parang Tambung, Mannuruki, Tamalate, Kota Makassar, Sulawesi Selatan 90224, Indonesia
*Corresponding Author. E-mail: iwan.suhardi@unm.ac.id

Submitted: 5 March 2020 | Revised: 21 April 2020 | Accepted: 29 April 2020

Abstract

One of the ability estimation methods that is widely applied to the Computerized Adaptive Testing (CAT) algorithm is the maximum likelihood estimation (MLE). However, the maximum likelihood method has the disadvantage of being unable to find a solution to the ability estimation of test-takers when the test takers' scores do not have a pattern. If there are test takers who get either score of 0 or perfect score, then the abilities of test-takers are usually estimated using the step-size model. However, the step-size model often results in item exposure where certain items will appear more often than other items. This surely threatens the security of the test because items that often appear will be easier to recognize. This study tries to provide an alternative strategy by modifying the step-size model and randomizing the calculation results of the information function obtained. Based on the results of the study, it is found that alternative strategies for item selection can make more varied items appear to improve the security of tests on the CAT.

Keywords: *item selection strategy, item exposure, step-size, adaptive testing*

How to cite: Suhardi, I. (2020). Alternative item selection strategies for improving test security in computerized adaptive testing of the algorithm. *REiD (Research and Evaluation in Education), 6*(1), 32-40. doi:https://doi.org/10.21831/reid.v6i1.30508.

Introduction

The development of item response theory (IRT) and computer technology that is faster and in a large capacity allows the development of computerized adaptive testing (CAT) (Haryanto, 2013, pp. 49–50). It is called "computerized" testing because the testing process no longer uses paper and pencil, but rather uses a computer device. It is called "adaptive" testing because the items that appear are chosen in such a way and adjusted to the ability of the test takers independently. CAT is a test conducted for test-takers where the items are determined based on the answers of the test takers (Winarno, 2013, p. 577). The efficiency of CAT compared to conventional testing models has been supported by several studies. The results of research by Eignor concluded that at the same level of measurement precision, adaptive tests only required a test length that was less than half of the computer-based test (CBT) device (Eignor, Stocking, Way, & Steffen, 1993; Grist, 1989, p. 2; Rudner, 1998, p. 2). McBride and Martin concluded that to achieve the same level of reliability, conventional testing required 2.57 times more items than adaptive testing (McBride & Martin, 1983).

The method widely used to estimate the ability of test-takers is the maximum likelihood estimation (MLE). The application of the maximum likelihood method has the disadvantage of being unable to find a solution when there are test takers who get extreme scores where all answers are always incorrect or always correct. To overcome this problem, the step-size method is generally employed. However, the application of the MLE and step-size model often leads to item exposure, which is the frequent appearance of certain items given to test takers. Although CAT is more efficient and reliable, the security of this testing is not guaranteed because certain items appear repeatedly. The items are easily recognized because they appear frequently, especially at the beginning of the item sequence. Therefore, modifications are needed to the conventional CAT algorithm to minimize the appearance of these easily noticeable items. The procedures that are commonly used in developing conventional CAT algorithms are elaborated as follows (Thissen, 1990).

Starting CAT

CAT generally starts with the selection of items with the difficulty level of moderate (Mills, 1999, p. 123; Santoso, 2010, p. 70; Vispoel, 1999). A test taker who answers incorrectly will then be given items with the difficulty level of easy. Conversely, if test taker answers correctly, they will be given items with the difficulty level of hard.

Estimating the Ability of the Test-Takers

The method commonly used to estimate the ability of test-takers is MLE (Baker, 1992; Birnbaum, 1968). The estimation of the ability of test-takers using the maximum likelihood method is calculated using the Newton-Raphson iterative procedure (Hambleton & Swaminathan, 1985, p. 83). The Newton-Raphson iterative procedure is performed first by subtracting the ratio of the first derivative to the second derivative from the initial $\hat{\theta}$ value so that it results in new $\hat{\theta}$. This procedure is repeated by using the new $\hat{\theta}$ and calculating the value of the new derivative ratio. The estimated value of θ at (m + 1) iteration can be expressed using the iterative relation as presented in Formula (1). Meanwhile, the error value is a correction factor that is formulated as seen in Formula (2), where *u* equals 1 if student's answer is correct and *u* equals 0 if student's answer is incorrect. Besides, *P* is probability of participants answering the items correctly, which is obtained by Formula (3).

$$\theta_{m+1} = \theta_m + error \text{ ....} (1)$$

$$error = \frac{\sum 1.7\, a\, (u-P)(P-c)/(P(1-c))}{\sum\, [-1.7^2\, a^2\, ((1-P)/P)]\, [(P-c)/(1-c)]^2} \dots (2)$$

$$P = c + \frac{(1-c)}{\left(1+exp\left(-1.7\, a\left(\theta_{duga\,0}-b\right)\right)\right)} \text{....} (3)$$

The iteration process is stopped when the error value $< \varepsilon$, with ε as limiting number whose value is very small. In this study, the ε value of 0.0001 was used.

One problem with the application of the MLE method in adaptive testing is the inability of the MLE method to find solutions when there are test takers who get an extreme score, which is either a score of 0 or a perfect score. To overcome the problem of the inability of the MLE method to estimate the ability of test-takers when their responses did not have a pattern, the step size method can be used (Dodd, 1990). Based on the step size method, the test taker's ability level is upgraded or degraded by a certain constant as long as the test taker's responses do not have a pattern, for example, by using a step size of 0.5.

Selection of the Next Item

After the test taker's ability is successfully estimated, the CAT algorithm will then select the next item. Lord recommended the use of the maximum item information procedure to select the next item (Lord, 1977). This method guarantees a highly accurate estimation of the ability of test-takers (Eignor et al., 1993). Items that have the greatest information function value on the ability of certain test takers are selected to be presented to them. The item information function is calculated at each ability level with the equation in Formula (4) (Hambleton, Swaminathan, & Rogers, 1991, p. 107).

$$I(\theta) = \frac{2.89\, a_i^2\, (1-c_i)}{\left[\left(c_i + exp\left(1{,}7\, a_i(\theta-b_i)\right)\right)\right]\left[1 + exp\left(-1{,}7\, a_i(\theta-b_i)\right)\right]^2} \text{....} (4)$$

Formula (4) shows that the information value only depends on the characteristic value of item parameters (for example the values of b, a, and c for the 3PL model) and the level of ability (θ). Thus, for each ability level (θ), the information function contribution for each item in the question bank can be calculated.

The test information function is the sum of the information functions of the test item and is written as in Formula (5). Meanwhile, the test information function illustrates the accuracy of the test set in estimating different levels of ability. The greater the information at the given ability level, the more accurate the ability is estimated from the test kit. The standard error of measurement (SEM) is expressed by the equation in Formula (6) (Hambleton & Swaminathan, 1985, p. 95).

$$TIF = \sum_{i=1}^{n} I_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$SEM = {}^{1}\!\big/\!\sqrt{TIF} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Termination of CAT

CAT termination uses criteria of equal measurement precision and a fixed number of items. Equal measurement precision criteria aim to produce test scores with the same measurement error level for each test taker. The standard error of measurement is limited to 0.30, which is equivalent to reliability of 91% on conventional tests (Thissen, 1990). By using the criteria, the number of items the test takers must work on can vary (where the number of items is not the same). However, to avoid the test process that may not be converging, the criterion of a fixed number of items is also used in the CAT termination rules by limiting the maximum items that appear, for example, as many as 20 items.

Giving Score to the Ability of the Test-Takers

The score of the ability estimation of the test-taker derives from the conversion of the value θ that is obtained by Formula (7).

$$Score = 50 + \left(\frac{50}{3}\theta\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

In this study, the CATs assessment results, which were the conventional CAT model (by taking the information value of items or the largest I (θ)) and the alternative CAT model (by taking some of the largest I (θ) values, then taken randomly to determine the value of I (θ) that would be used), were compared. After that, the alternative CAT model was treated using the step-size method with an additional variable of response time when the test takers' responses did not have a pattern.

The assumption underlying the response time variable is those test-takers who have a high level of ability will be able to answer the items correctly in a shorter time than those who have a lower level of ability. Lidia Martinez compared groups of test-takers who took a test using CBT and found that the groups that spent the shortest average time responding to the initial test item obtained a higher average score (Martinez, 2009). Phil Higgins' research results showed that in CBT, if the item difficulty index was higher, then test-takers would need more time to answer and review the items (Higgins, 2009). This showed that the test taker's response time in working on the items correctly correlated with the estimation of the test taker's ability level.

Method

This study used a Research and Development (R&D) approach. The study began with the development of a question bank to obtain 265 items based on the 1-parameter logistic item response theory (1PL IRT) model. Characteristics of items in the form of parameters of the difficulty level of 265 items were obtained from the validation of processed results using the BILOG-MG software, obtained from the response test using CBT. The total number of items before validation was originally 290 items. A summary of the question bank validation statistics developed and used in this study is presented in Table 1.

Table 1. Summary of Item Statistics on Question Bank

General Information	Based on 1PL IRT Number of items = 265 items
Criteria of Item Difficulty Index (*b*)	Hard category = 40 items Moderate category = 128 items Easy category = 97 items

In the 1PL IRT model, the probability of a person with a certain ability (*θ*) answering the items correctly depends only on the difficulty level of the items (*b*). In this study, the estimation methods of the ability of test-takers are the MLE and step-size methods.

Next, two adaptive test designs developed were the conventional and the alternative CAT model. In this study, the development of CAT software referred to the incremental model (Pressman, 2001, pp. 35–36).

In the conventional CAT model, the first item selection method employs a difficulty level of moderate, starting with a range of *b* values from -0.5 to 0.5 chosen randomly. The ability level estimation is calculated using the MLE method. However, when the test-takers' responses have not had a pattern, their ability is estimated using the step size method with a value of 0.5. The next item that is selected is the item that has the greatest information function value on a particular ability.

The alternative CAT model has the same principles as the conventional CAT model. The difference is in the selection of the second and subsequent items, which uses the principle of randomizing the value of the information function in the 5-4-3-2-1 pattern. The pattern rule of 5-4-3-2-1 used was that the second item was selected from one item randomly from the five items that had the largest information function, the third item was selected from one item randomly from the four items that had the largest information function, the fourth item was selected from one item randomly from three items that had the largest information function, the fourth item was selected from one item randomly from three items that had the largest information function, and the fifth item was selected from one item randomly from two items that had the largest information function. Meanwhile, for the sixth and subsequent items, the item selection criteria revert to the maximum information function criteria or revert to the conventional CAT model.

To estimate the ability of test-takers on the alternative CAT model when their responses have not had a pattern, a step-size method is used with the addition of the response time variable. The test-takers' estimated initial ability level is selected at the ability level of *θ*0. Moreover, the step-size interval changes constantly by **k** (where in this study, the value of **k**=0.5). If the test taker responds by answering incorrectly, the test-taker's estimated ability level becomes ***θ*0 – k** or equal to 0-0.5 = -0.5. Meanwhile, if the test taker answers correctly, the estimated ability level becomes ***θ*0+*x* k** or **0.5 . *x***, where ***x*** is a positive constant multiplier and the value depends on the category of students' response time when their answer is correct.

Table 2 shows a simulation procedure to estimate the test taker's ability level with a step-size interval added to the response time factor. Test takers were given 300 seconds to respond to each item. If for more than 300 seconds there is no response from test taker, the response is declared incorrect and easier-level items will be displayed. In this study, the criterion for test termination is that the test is terminated if the SEM value has reached 0.30. An SEM value of 0.30 is equivalent to the reliability of 0.91 in conventional tests such as paper and pencil tests (Thissen, 1990).

Table 2. Estimation of Ability of Test-Taker in the Response-Time-Based Step-Size Method

Annotation: θ_0 = Initial ability = 0 **k** = step size = 0.5 **x** = constant multiplier $\theta_{ke\text{-}i} = \theta_{i-1} + xk$ (for correct response) $\theta_{ke\text{-}i} = \theta_{i-1} - k$ (for incorrect response)	**Responding with Correct Answer in Consecutive Times**			**Responding with Incorrect Answer in Consecutive Times**		
	Item 1	**Item 2**	**Item 3**	**Item 1**	**Item 2**	**Item 3**
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
Very fast: x = 1.4 (≤ 30 seconds)	0.7	1.4	2.1	-0.5	-1.0	-1.5
Fast : x = 1.3 (31 to 60 seconds)	0.65	1.3	1.95	-0.5	-1.0	-1.5
Moderate: x = 1.2 (61 to 90 seconds)	0.6	1.2	1.8	-0.5	-1.0	-1.5
Slow : x = 1.1 (91 to 120 seconds)	0.55	1.1	1.65	-0.5	-1.0	-1.5
Very slow : x = 1 (≥ 121 seconds)	0.5	1.0	1.5	-0.5	-1.0	-1.5

Table 3. Testing Results of Conventional CAT Model when Responses of Answers Have Not Had Pattern Yet

Item 1	Item 1 was taken randomly with the difficulty level of moderate (-0.5 ≤ b ≤ 0.5)			
Item	**Test Takers' Responses are Always Correct**		**Test Takers' Responses are Always Incorrect**	
	Value of θ	**List Number of Item**	**Value of θ**	**List Number of Item**
Item 2	- 0.5	209	0.5	275
Item 3	- 1	164	1	081
Item 4	- 1.5	113	1.5	002
Item 5	- 2	044	2	091
Item 6	- 2.5	237	2.5	115

Findings and Discussion

Before the answers have a pattern, the conventional CAT model will use the step-size method with an interval of 0.5. This means that if the test taker always responds with the correct answer, then the second and subsequent items that will appear are items that have the largest information function value at the ability level (*θ*) of 0.5, 1, 1.5, 2, 2.5, and 3 respectively. Meanwhile, for test-takers who always respond with the incorrect answer, the second and subsequent items that will appear are items that have the largest information function value at θ of -0.5, -1, -1.5, -2, -2.5, and -3 respectively. The results that were obtained in the conventional CAT model are summarized in Table 3.

From the results of the study, it was found that items with list numbers of 209, 164, 113, 044, 237, 275, 081, 002, 091, and 115 were items that appeared more often than other items. The items that often appear will make the security of the test in the conventional CAT model degrade because they may be items that have been recognized by the test takers.

From the results of conventional CAT model testing, it was found that the number of items with difficulty index of moderate, which was indicated by the difficulty index value (b) ranging from -0.5 to +0.5, was 128 items. This meant that the probability of the first item having a chance to appear was 128 items chosen randomly. This was indeed in accordance with the criteria applied to the conventional CAT model design algorithm, that the initially selected items were items with difficulty index of moderate (-0.5 to +0.5).

After the first item displayed and was responded by the test taker, the second item was presented by using the step-size method. This meant that if students responded to the item with the correct answer, then the second item displayed was the item with maximum information for θ = 0.5. However, if students responded to items with incorrect answers, then the second item that was displayed was an item with maximum information for θ = -0.5. Thus, it was certain that in the conventional CAT, the second item only consisted of the possibility of 1 of 2 items only. In this study, the second item presented was question item number 275 (if the answer was correct) and question item number 209 (if the answer was incorrect). The frequent appearance of item number 275 and item number 209 made the security of CAT threatened due to the familiarity with the question.

Another case that also often arises is that there has not been a pattern in students' answers so that the step-size method is used. For example, if students answered questions correctly, the items that would appear were questions that had a maximum information value for θ = 0.5, 1.0, 1.5, 2.0, and 2.5, which were the second item whose item number was 275, the third item whose item number was 081, the fourth item whose item number was 002, the fifth item whose item number was 091, and the sixth item whose item number was 115.

However, if students always answered the question incorrectly, then the item that appeared was questions that had a maximum information value for θ = -0.5, -1.0, -1.5, -2.0, and -2.5, i.e., the second item with item number 209, third item with item number 164, fourth item with item number 113, fifth item with item number 044, and sixth item with item number 237. In the conventional CAT model, if the responses of the test takers have the same pattern, then the items that appear will also be the same. This is what makes the security level of the conventional CAT model suboptimal.

If students' responses already had patterns (where the responses already consisted of correct and incorrect answers), then the items that appeared next had been quite varied because the first item that appeared already had a relatively large variety of items (128 items). However, by using the maximum information function value model to search for items that corresponded to the estimated level of test-takers' abilities, it was very possible that many items could not be presented because they never obtained the maximum function value for each level of ability.

The alternative solution proposed was to use the step-size method based on the student's response time in answering correctly. Student responses were grouped into groups based on the time spent by students in answering the questions correctly. In the step-size method based on response time, the step-size value formula was given an additional constant multiplier based on the response time. The faster the students answered correctly, the greater the constant multiplier became.

An additional solution proposed was to randomize the maximum information function value. If the conventional CAT model determined the items that appeared based on the value of the (single) maximum information function, then the alternative CAT model determined the items that appeared by randomizing the maximum information function values based on groups of 5–4–3–1–1. For example, one of the results of testing the alternative CAT model is presented in Table 4.

From Table 4, the calculation procedure for the alternative CAT model can be observed. From the table, it can be seen that the items that appear in the alternative CAT model are more varied compared to those in the conventional CAT model. The algorithmic procedure in the alternative CAT model can be explained as follows.

The First Item that Appeared was Item Number 239 with *b* = -0.416

The first item appeared in accordance with the criteria that items were taken randomly with a difficulty index of moderate whose b value ranged from -0.5 to 0.5. Item number 239 fulfilled the criteria. Because students' answers did not have a pattern, the method of estimating the ability level was the step-size of 0.5. Students' answers were declared correct (value 1). The time that was spent to work on the first item was 34 seconds, so it was included in the fast category (between 31 and 60 seconds) with a multiplier factor = 1.3. Thus, the value of θ was 0.5 x 1.3 = 0.64.

Table 4. Results of Alternative CAT Model Testing

No.	Item	*b*	Response	Time (second)	θ	IIF	TIF	SEM
1	239	-0.416	1	34	0.65	0.7224	0.7224	1.18
2	182	0.662	1	40	1.3	0.7223	1.4447	0.83
3	192	1.32	0	8	1.1809	0.7225	2.1672	0.68
4	042	1.181	0	49	0.8579	0.7225	2.8897	0.59
5	132	0.861	1	20	1.3204	0.7225	3.6122	0.53
6	192	1.32	0	26	1.1161	0.7225	4.3347	0.48
7	152	1.119	1	10	1.5224	0.7225	5.0572	0.44
8	002	1.524	0	14	1.3846	0.7224	5.7796	0.42
9	161	1.396	0	7	1.2399	0.7224	6.502	0.39
10	013	1.251	1	9	1.5831	0.7225	7.2245	0.37
11	127	1.579	1	15	1.9486	0.7217	7.9462	0.35
12	060	1.987	0	12	1.8848	0.7223	8.6685	0.34
13	062	1.867	0	17	1.8118	0.7222	9.3907	0.33
14	163	1.787	0	19	1.7339	0.7214	10.1121	0.31
15	124	1.687	1	14	2.0656	0.7214	10.8335	0.3

The Second Item that Appeared was Item Number 182 with *b* = 0.662

The second item appeared because it had the five largest information function values at the value of θ = 0.65, according to the use of randomization with the principle of 5–4–3–2–1. From the five alternative values of the largest information function (see Table 5), the item with number 182 was selected randomly. The second item was answered correctly (then the response value was 1). Because students' answers did not have a pattern, the method of determining the estimated ability level was the step-size of 0.5. The item was done in 40 seconds and included in the fast category (between 31 to 60 seconds) with a multiplier factor of 1.3. Thus, the value of θ = 0.65 + (0.5 x 1.3) = 1.3.

Table 5. The Five Alternative Values of the Largest Information Function

Rank	Information Function	Item	*b*
1	0.722495	153	0.647
2	0.722492	274	0.654
3	0.722474	202	0.643
4	0.722425	182	0.662
5	0.721861	003	0.685

The Third Item that Appeared was Item Number 192 with *b* = 1.32

The third item appeared because it had the four largest information function values at the value θ = 1.3 according to the use of randomization with the principle of 5–4–3 –2–1. Of the four alternative values for the largest information function (see Table 6), the item with number 192 was randomly selected. The third item was responded with an incorrect answer (so the response value was 0). Because students' answers did not have a pattern, the method for estimating the level of ability was MLE. The value of θ obtained was = 1.1809.

Table 6. The Four Alternative Values for the Largest Information Function

Rank	Information Function	Item	*b*
1	0.722349	053	1.317
2	0.722291	192	1.32
3	0.72227	179	1.321
4	0.722091	145	1.272

The Fourth Item that Appeared was Item Number 042 with *b* = 1.181

The fourth item appeared because it had the three largest information function values at the value θ = 1.1809 according to the use of randomization with the principle of 5–4–3 –2–1. Of the three alternative values for the largest information function (see Table 7), item with number 042 was randomly selected. The fourth item was responded with an incorrect answer (so the response value was 0). Because students' answers did not have a pattern, the method for estimating the level of ability was MLE. The value of θ obtained was = 0.8579.

Table 7. The Three Alternative Values for the Largest Information Function

Rank	Information Function	Item	*b*
1	0.7225	042	1.181
2	0.7225	057	1.181
3	0.722449	021	1.171

The Fifth Item that Appeared was Item Number 132 with *b* = 0.861

The fifth item appeared because it had the two largest information function values at the value θ = 0.8579 according to the use of randomization with the principle of 5–4–3 –2–1. Of the two alternative values for the largest information function (see Table 8), the item with number 132 was randomly selected. The fifth item was responded with the correct answer (so the response value was 1). Because students' answers did not have a pattern, the method for estimating the level of ability was MLE. The value of θ obtained was = 1.3204.

Table 8. The Two Alternative Values for the Largest Information Function

Rank	Information Function	Item	*b*
1	0.722495	132	0.861
2	0.722474	242	0.865

The Sixth Item that Appeared was Item Number 192 with *b* = 1.32

This sixth item appeared because it had one largest information function value at the value θ = 1.32 according to the use of randomization with the principle of 5–4–3 –2–1. Of the one alternative value for the largest information function (see Table 9), the item with number 192 was randomly selected. The sixth item was responded with an incorrect answer (so the response value was 0). Because students' answers were patterned, the method for estimating the level of ability was MLE. The value of θ obtained was = 1.1161.

Table 9. The One Largest Information Function Value

Rank	Information Function	Item	*b*
1	0.7225	192	1.32

The subsequent items (i.e. the seventh to fifteenth items) used the same method to determine the item that had the largest information function at its value of θ. The fifteenth item became the last item because the criterion for termination rule had been met (SEM = 0.3). It was converted to a numerical value of 85.

This alternative CAT model has been proven to be able to overcome a fundamental shortcoming in the conventional CAT model, which was the frequent appearance of certain items. From Table 3, it can be seen that in the conventional CAT model, several similar items would appear, especially in the initial patterns of CAT execution. Meanwhile, in Table 4, there were many variations on the possible items that appeared on the alternative CAT model, even though the patterns of students' answers were the same. The many variations of items that appear in the alternative CAT model can reduce the level of item exposure on CAT so that it will make the CAT more secure. The item variations that appeared in the alternative CAT model actually had item difficulty index that was not much different from those that appeared in the conventional CAT model, so it did not increase the test length or reduce the efficiency of the estimation of the ability of the test takers.

Conclusion

From the results of this study, it can be concluded that the alternative CAT model was able to decrease the level of item exposure on the CAT, thereby increasing the security of the CAT without increasing the test length or reducing the efficiency of the CAT. The strategy adopted by the alternative CAT model was to select items using the step-size method based on response time and randomization of the maximum information function value with the criteria of 5–4–3–1–1 by applying the maximum likelihood estimation (MLE) to estimate the ability level of the test takers. The strategy has been proven to be able to present items with more variations, but still with item difficulty index which was not much different in the response patterns of the same test takers.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental rest scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, *14*(4), 355–366. https://doi.org/10.1177/014662169001400403

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation*. https://doi.org/10.1002/j.2333-8504.1993.tb01567.x

Grist, S. (1989). Computerized adaptive tests. In *ERIC Digest No. 107*. Retrieved from https://files.eric.ed.gov/fulltext/ED315425.pdf

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item* *response theory*. Newbury Park, CA: Sage Publications.

Haryanto, H. (2013). Pengembangan computerized adaptive testing (CAT) dengan algoritma logika Fuzzy. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1), 47–70. https://doi.org/10.21831/pep.v15i1.1087

Higgins, P. (2009). *Candidate measured ability and use of time*. Retrieved from https://www.rasch.org/mra/mra-10-09.htm

Lord, Frederic M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*(1), 95–100. https://doi.org/10.1177/014662167700100115

Martinez, L. (2009). *Time usage and candidate performance*. Retrieved from http://www.rasch.org/mra/mra-06-09.htm

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224–236). New York, NY: Academic Press.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Mahwah, NJ: Lawrence Erlbaum Associates.

Pressman, R. S. (2001). *Software engineering: A practitioner's approach* (5th ed.). New York, NY: McGraw-Hill Higher Education.

Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial*. Retrieved from http://edres.org/scripts/cat

Santoso, A. (2010). Pengembangan computerized adaptive testing untuk mengukur hasil belajar mahasiswa Universitas Terbuka. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(1), 62–83. https://doi.org/10.21831/pep.v14i1.1976

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 161–186). Hillsdale, NJ: Erlbaum.

Vispoel, W. P. (1999). Creating computerized adaptive tests of music aptitude: Problems, solutions, and future directions. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 151–176). Mahwah, NJ: Lawrence Erlbaum Associates.

Winarno, W. (2013). Pengembangan computerized adaptive testing (CAT) menggunakan metode pohon segitiga keputusan. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*(2), 574–592. https://doi.org/10.21831/pep.v16i2.1132

REiD (Research and Evaluation in Education), 6(1), 2020, 41-50
Available online at: http://journal.uny.ac.id/index.php/reid

NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model

***1Roudloh Muna Lia; 1Ani Rusilowati; 1Wiwi Isnaeni**
[1]Graduate School, Universitas Negeri Semarang
Jl. Kelud Utara III, Petompon, Gajahmungkur, Kota Semarang, Jawa Tengah 50237, Indonesia
*Corresponding Author. E-mail: aliamoetz@yahoo.co.id

Submitted: 10 February 2020 | Revised: 27 April 2020 | Accepted: 1 May 2020

Abstract

The instrument of measuring test attributes must be valid and reliable. This study was carried out since the validity and reliability testing of the chemistry items used by the testee is necessary. This study aims to estimate the validity and determine the reliability of chemical test instruments oriented Next Generation Science Standards (NGSS). The research was conducted through a quantitative descriptive approach in two vocational schools of engineering program which had 130 testees. The instrument used was an NGSS-oriented chemistry test instrument containing 35 items and an expert validation questionnaire. The obtained test participant's response from the test instrument was collected through the documentation method. Item in NGSS test were presented to three subject matters experts. The validities used were the content validity and the construct validity. The reliability was tested through internal consistency and interrater consistency approaches. The results show that content validity (Aiken's V) is at a range of 0.50 to 1.00. The value of the unexplained variance is less than 10%, which means that it is well-categorized. This analysis is strengthened by CFA which has a goodness of fit and a good measurement model fit. The parameters used to test model fit are CFI, NFI, RMSEA and the value of loading factor. Some results values are over 0.90 and RMSEA is 0.00 and more than 0.3 of loading factor value on each item. All scales had alpha reliability more than the criteria of 0.70. Thus, the developed chemical test item were proven as valid and reliable instruments.

Keywords: *validity, reliability, NGSS*

How to cite: Lia, R., Rusilowati, A., & Isnaeni, W. (2020). NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model. *REiD (Research and Evaluation in Education), 6*(1), 41-50. doi:https://doi.org/10.21831/reid.v6i1.30112.

Introduction

In the Government Regulation No. 32 of 2013, it is written that learning process in the education unit is carried out interactively, inspiratively, pleasently, defiantly which motivates students to participate actively, as well as providing sufficient space for initiative, creativity, and independence by following their talents, interests, physical and psychological development of students. Educators or teachers are required to carry out the mandate of government regulation. The implementation of learning will be achieved based on the goals set if it is suitable for the students' talents and interests. Students from the Engineering Program of vocational school will be less suitable if Business Economics subject is taught because it does not match with interests and expertise areas of students, likewise the Chemistry lessons that are applied at Vocational High School (VHS). The existence of chemistry subjects in the Engineering Skills Program can support the development of learners' competencies if the material is adjusted to the expertise area of students (Wena, 2009 in Banne, 2018, p. 45). If the chemistry is taught separately and it is not associated with productive subjects in the expertise area which is occupied, the chemistry subject will be irrelevant (Astuti, Sunarno, & Sudarisman, 2016).

Facts in the field from the results of questionnaire distribution in vocational students showed as many as 76 % of students stated that chemistry was a difficult subject. The reason is that students are less interested in chemistry lessons because they consider that chemistry subject is not important for them (Lia & Isnaeni, 2018, p. 403). Chemistry as an adaptive subject in VHS is expected to be in accordance with productive material needs. One way to present chemistry subjects to be in accordance with productive material in learners' expertise area is through Next Generation Science Standards (NGSS) (Lia, 2019, p. 113).

NGSS provides the opportunity to include engineering in science (National Research Council, 2013, p. xviii). One of the assessment challenges in NGSS is creating assignments that include the practical side of science and engineering (Damelin, 2017). NGSS offers a new standard combining content and practice in science and engineering (National Research Council, 2013). NGSS creates a new vision for science education based on the idea that science is a unity of knowledge and a set of practices related to developing knowledge (Penuel, Harris, & DeBarger, 2015, p. 45). This teaching and learning approach is built on decades of research that identifies problems through learning in science classes and promising strategies to make learning to be more meaningful and effective for students (Reiser, 2013).

NGSS-oriented chemistry learning had been successfully developed by Lia (2019). After the learning process has been implemented, it is followed by an assessment activity. Assessment is an activity conducted to measure and assess the curriculum achievement level (Sudrajat, 2016, p. 1). Through assessment, any lacks in learning can be identified and can be evaluated.

The assessment instrument in measuring the question attributes as students' evaluation material must be valid and reliable. Therefore, further research on the development of the NGSS learning model, namely the preparation of chemical items needs to be conducted. The NGSS-oriented chemistry items developed provide breakthroughs to give students a more meaningful assessment. Assessment becomes more meaningful because it is associated with technical material by following the field occupied by students. Before carrying out the test, some practicums were oriented towards NGSS which made the chemical side more desirable (Lia, 2019, p. 113).

The NGSS-oriented chemistry question items must have two important requirements. Those are having a good validity and reliability level. Validity and reliability will be fulfilled if the questions have been arranged. Item analysis is analyzed in order to obtain the adequate quality of the question, and data processing and interpretation of the assessment result (Kadir, 2015, p. 71). Reynolds, Livingston, and Willson (2010, p. 144) state that validity means the extent to which theoretical and empirical evidence supports the meaning and interpretation of test scores. In addition, Dewi and Sukadiyanto (2015, p. 230) explain that a valid test is a test that can measure accurately and thoroughly the symptoms which are to be measured). Reliability is test consistency (Bhakti, 2015; Khumaedi, 2012). It means that a reliable test must have consistent results even if tested repeatedly at different times. It is in accordance with the theory explained by Reynolds et al. (2010, p. 91) that reliability is the accuracy or stability of the assessment results. The measuring tools used by evaluators when carrying out evaluation activities must have accuracy, consistency, and stability so that the measurement results obtained can measure accurately (Amalia & Susilaningsih, 2014). A set of tests must have accuracy when it is used. It also should be consistent and stable in the sense that there is no change from one measurement time to another (Utami, 2018, p. 5).

This study aims to estimate the validity and determine the reliability of chemical test instruments oriented NGSS to measure the level of understanding of chemical material in engineering. Research on the validity and reliability of the test instruments has been conducted by Mohamad, Sulaiman, Sern, and Salleh (2015), Kusaeri, Sutini, Suparto, and Wardah (2019), and Iskandar (2017). The differences between previous and current research are the analysis of the validity of the construct using the confirmatory factor analysis (CFA) modification and the Rasch model. It is expected that research on validity and reliability will increase knowledge in the field of teaching, especially in the evaluation of learning.

Rasch model used in this study has several advantages which can identify the error response, predict missing data scores, distinguish the ability of respondents with the same raw score, and also identify any indications of guesses and cheaters (Sumintono & Widhiarso, 2015, pp. 44–45). These advantages make the Rasch model more accurate (Lord in Nurcahyo, 2016). Rasch modeling can produce standard error measurement values which can improve the accuracy of calculations (Ardiyanti, 2016, p. 261). Sabekti and Khoirunnisa (2018, p. 69) confirm that the Rasch model is more recommended to be used in the development of test instruments.

An assessment of the appropriateness of the item's display and/or content validity becomes the earlier steps. Assessments carried out by a panel of experts and chemistry teachers are also included in the expert panel (Ismail, Permanasari, & Setiawan, 2016, p. 239). Instruments that have been compiled and validated by experts are then validated empirically through trial instruments in small classes (Prabowo & Ristiani, 2011, p. 80).

The high of agreement among experts who assess the feasibility of an item can be estimated and quantified. Then, the statistical calculation is used as an indicator of the item content validity and the test content validity. This study used an assessment procedure in measuring validity thorough a content validity coefficient (the content validity of the test with a V index) proposed by Aiken's V. The construct validity was tested using CFA with the help of Lisrel 8.8 software. Proof of construct validity used first order confirmatory factor analysis which calculated the estimated value of the item against its latent variable. According to Sitninjak and Sugiarto in Rusilowati (2014, p. 131), the validity of an observed variable can be seen from the factor loading of the variable against latent variable. Variables are labelled as good construct validity when the goodness of fit and the measurement model fit are met.

Method

The study was conducted in two vocational high schools in Engineering Program with a total of 130 testees. The instrument used was an NGSS-oriented chemical test instrument, amounting to 35 items and validation sheet. Based on the test instrument, the result of the test participants' answers was obtained and collected through the documentation method.

Three experts were assessing to obtain three sheets of questionnaire result. The validity was estimated by content validity, validity in large class trials, and construct validity. Then, the reliability was estimated through internal consistency and interrater consistency approaches. To analysis the content validity, the Aiken's V Formula was used. The construct validity with CFA was used with the help of Lisrel 8.8 software. The internal consistency reliability used in this study is the Spearman-Brown's formula in small class trials, whereas in large class trials, the Rasch alpha Cronbach model and interrater reliability using three raters tested using two-way ANOVA with Ebel formula were used.

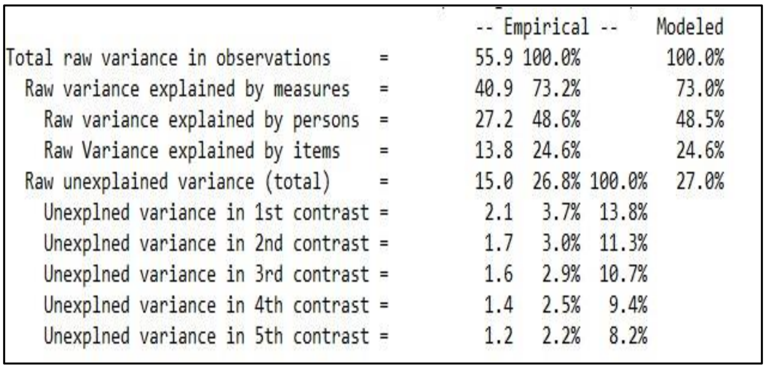
Findings and Discussion

Validity Test

Content validity was estimated with Aiken's V index. Items in NGSS test were presented to three experts to assess the compatibility of the material, construction, language and compatibility with NGSS. The experts also filled out a questionnaire containing the conclusions of the experts' assessment of chemistry-oriented items in NGSS. Quantitative data that present a summary of quantitative expert agreement coefficient data are shown in Table 1.

Table 1. Coefficient Data of Expert Agreement

Item Number	Aiken's V Index	Criterion	Conclusion
1, 5, 6, 7, 8, 9, 11, 12, 14, 16, 17, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35	1.0	valid	eligible
2, 3, 4, 13, 15, 18 19, 20	0.8	not valid enough	little revision
21	0.7	not valid enough	little revision
10	0.5	not valid enough	little revision



```
                                          -- Empirical --    Modeled
Total raw variance in observations     =     55.9 100.0%         100.0%
  Raw variance explained by measures   =     40.9  73.2%          73.0%
    Raw variance explained by persons  =     27.2  48.6%          48.5%
    Raw Variance explained by items    =     13.8  24.6%          24.6%
  Raw unexplained variance (total)     =     15.0  26.8% 100.0%   27.0%
    Unexplned variance in 1st contrast =      2.1   3.7%  13.8%
    Unexplned variance in 2nd contrast =      1.7   3.0%  11.3%
    Unexplned variance in 3rd contrast =      1.6   2.9%  10.7%
    Unexplned variance in 4th contrast =      1.4   2.5%   9.4%
    Unexplned variance in 5th contrast =      1.2   2.2%   8.2%
```

Figure 1. Unidimensionality Test

Based on the results of the data analysis in Table 1, 25 of 35 items are valid and 10 items are not valid enough, which means that there are some revisions. Comparing to previous studies, the quality classification research items analyzed are better than the result of research by Hasnah (2017) in which only nine of 40 items are categorized well.

Construct validity was proven by combining the factor analysis of the Rasch model and CFA (using Lisrel 8.8 software). The first step to see the construct validity with the Rasch model is through Output Diagnosis Item Polarity (Hayati & Lailatussaadah, 2016, p. 173). All items have a positive Point Measure Correction (Pt. Mea- Corr). A total of 14 items have strong or high correction numbers. One of the items (question number 5) has a moderate correlation number (0.57). It is in accordance with the opinion of Othman, Salleh, Hussein, and Wahid (2014, p. 117) that the high *Pt. Mea Corr* (0.68- 1.00) shows that a question item can distinguish respondents' ability.

The result of the correlation figures on *Pt. Mea Corr* is strengthened to the results of the unidimensionality test through the output table unidimensionality. The output table unidimensionality is presented in Figure 1.

The raw variance in Figure 1 shows a high number (73.2%). According to the opinion of Hakiki, Fitri, and Agung (2018, p. 42), the results of the analysis which have a unidimensionality requirement of more than 60 % show special meaning. The instrument which is developed can measure what should be measured. Variance values that cannot be explained (unexplained variance) successively are 3.7; 3.0; 2.9; 2.5; and 2.2. It shows that the variances which cannot be explained by the instruments are all less than 10%. It indicates that the unidimensionality in the instruments falls into a good category (Wibisono, 2014, p. 744).

The construct validity test on Rasch is only for the response of the tested item, whereas to find out the covariance between the test items, the CFA model with the Lisrel or Amos or SPSS programs is needed. About specifying a model for a data set, the procedures for CFA appear to be more advanced, simpler, and more user-friendly than those developed for Rasch (IRT). The CFA model can calculate an accurate estimate of the chi-square size of the fit model and related degrees (Reise, Widaman, & Pugh, 1993, pp. 554–563). Therefore, the researchers strengthened the construct validity test through the Lisrel program.

Conceptually, to make a test across NGSS, three components should be recked, namely DCIs, SEs, and also CCs. DCIs are very dependent on the material that will be made from the instrument. Then, SEPs and CCs are the characteristics of NGSS-oriented statistics. SEPs consist of six aspects with 15 indicators. CCs consist of three aspects with 14 indicators. The results of the NGSS instrument construct validity with CFA prove that the dimensions of CCs which consist of three aspects with 14 indicators are evidenced by the factor loading value and item compatibility parameters. The analysis of CCs components consisting of three aspects and 14 indicators is generated in a diagram presented in Figure 2.

Analysis through CFA proved that CCs dimensions which consisted of three aspects with 14 indicators are evidenced by the value of loading factor and items that are compatible with the parameters. All factor loading's value shows that there are more than 0.3. Factor loadings which are less than 0.5 are removed (Arifin, Yusoff, & Naing, 2012). The parameters that are used to test model fit are CFI, NFI, and RMSEA. CFI and NFI are over 0.90 (CFI=0.92; NFI=0.90) and RMSEA is 0.00. It is compatible with the theory that the expected CFI and NFI values are above 0.90 (Zehir, Akyuz, Eren, & Turhan, 2013, p. 9). RMSEA is recommended to be under 0.05 though acceptable up to 0.08 (Sohail & Jang, 2017). In Rusilowati (2014, p. 134), it is stated that the compatibility of the model that is developed by empirical data at a minimum can be seen from three match sizes that represent the three categories of match test different models. When two of the three categories are significant, the model developed is compatible with the data. All model fits were acceptable and according to the literature, the validity of the measurements in the current study met the criteria.

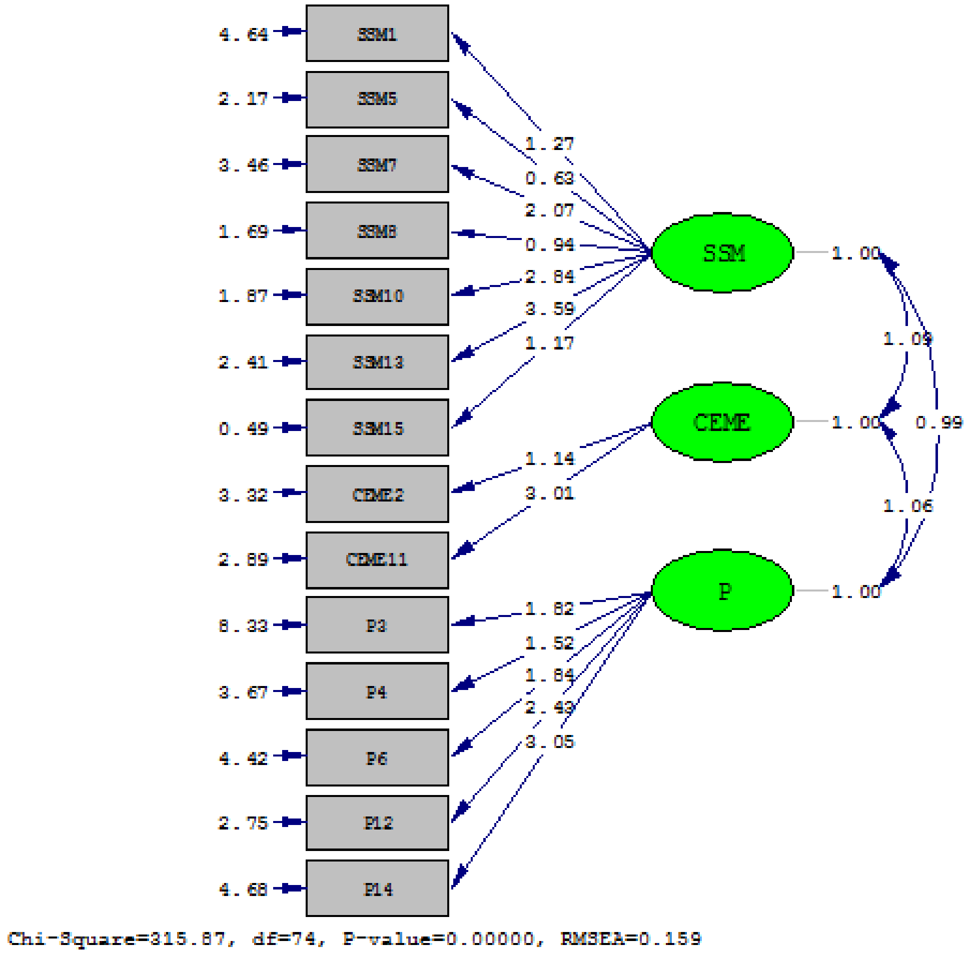


Figure 2. CCs Path Diagram

The validity of the large class trial phase was analyzed using the Rasch through the Output model, item fit order. The output is presented in Table 2.

Table 2. Item Fit

Item's	Outfit		
Number	MNSQ	ZSTD	PT. Mea Corr
1	2.11	3.9	0.68
3	1.99	3.2	0.75
6	1.47	1.7	0.75
2	1.33	1.4	0.73
7	1.33	1.7	0.75
9	1.16	0.8	0.74
12	1.08	0.5	0.80
5	1.03	0.2	0.57
8	0.94	-0.2	0.69
10	0.80	-1.0	0.87
4	0.81	-0.8	0.82
14	0.77	-1.0	0.83
13	0.56	-2.0	0.90
15	0.59	-1.9	0.78
11	0.48	-1.6	0.86

The item fit information is useful for identifying the indications of misconception (Sumintono & Widhiarso, 2015, p. 77). In Table 2, based on MNSQ, ZSTD, and Pt. Mea Corr, it can be concluded that 15 items were classified as valid, but there is one item namely question number 1 which is indicated as a misconception. The MNSQ value is 2.11 and the ZSTD is 3.9 which represents unexpected data. The cause of outlier MNSQ and ZSTD values is from some testee's answers. Those are reversed between "the oxidation-reduction reaction and the reason", but Pt. Mea Corr is still within the limit of more than 0.4 and less than 0.85. Therefore, 15 items have been used to measure the quality of education because these questions have been analyzed. It is in accordance with the opinion of Pancoro (2011, p. 94) that test questions need to be first analyzed to have the same characteristics so that they can be used to measure the quality of education.

Reliability Test

The reliability test consists of (a) inter-rater reliability, (b) small-scale trial reliability, and (c) large-class trial reliability. Based on Table 3, the values of the reliability of the tests are 0.17, 0.82, and 0.94. Inter-rater reliability (among experts) is very low, the reliability of small class trials is very high, and the reliability of large classes is special. A discussion of the three reliability tests is elaborated as follows.

Inter-rater Reliability

Inter-rater reliability is a preliminary part of a study (Dockrell et al., 2012, p. 633). Interrater reliability was calculated after calculating the content validity among three validators. Level agreement between three validators can be explained through the reliability coefficient between rater (assessors) using two-way ANOVA-analysis with the Ebel formula. Two-way ANOVA analysis through SPSS 16.0 is presented in Table 4.

In Table 4, it can be explained that Rater is the assessor and Item is a matter of Items. The mean square value of Rater is 0.495, the value of the item is 0159 and the interaction between Rater and Item (Rater * Item) is 0.132. These values are entered in the Ebel formula and produce a reliability coefficient of 0.17. The reliability coefficient of r value is less than 0.2. The reliability

Table 3. Reliability Data Analysis

Trial Phase	Reliability	N of Items
Expert (Expert Judgment)	0.17	35
Small Class	0.82	25
Big Class	0.94	15

Table 4. Output Reliability of Two-Way ANOVA

Source	Mean Square
Rater	0.495
Item	0.159
Rater*Item	0.132

among the assessors in assessing the contents of the instrument is still not consistent (Rusilowati, 2014, p. 29). When the reliability coefficient obtained is not high enough, there are inconsistencies among raters (Pinilih, Budiharti, & Ekawati, 2013, p. 25). The reason for this inconsistency in this research is the difference in viewpoints in evaluating chemical test instruments. For example, expert 1 puts more emphasis on its chemical content while expert 3 is more inclined in evaluating the appearance and suitability of the answers.

Small Class Trial Reliability

Reliability using the Spearman-Brown formula was applied to small classes and searched using the Anastes Description application. The reliability coefficient of small class tests based on Table 3 shows that the coefficient number is 0.82. Figures for reliability coefficient is 0.8 r < 1.0, which indicates very high reliability.

Big Class Trial Reliability

In the big class stage, the reliability is seen with the help of Winstep 3.73 program. Reliability in the Rasch model is illustrated by the presence of a separation index. The separation indexes reported are the item reliability and the person reliability which are supplemented by Cronbach Alpha KR-20 of reliability coefficient figures. Those are three successive coefficient numbers (0.91, 0.98 and 0.94). All three of these figures indicate very high reliability. Separation reliability (item or person reliability) is categorized as high value because the study sample and grain difficulty level have a wide range and produce a small measurement error. Broad grain means that the item has a difficulty level from the easiest to the most difficult. Similarly, in the study sample, a broad sample means that the sample can spread from the smartest to the least clever (Linacre, 2016, p. 256). The output reliability can be seen in Table 5. In Table 5, in addition to the reliability coefficient, there is also important information related to the statistical summary of the test participant's overall response patterns, namely (a) INFIT MNSQ ZSTD, and OUTFIT MNSQ ZSTD, and (b) Separation.

INFIT MNSQ ZSTD and OUTFIT MNSQ ZSTD

The MNSQ INFIT and MNSQ OUTFIT values are 0.99 and 1.21, respectively for persons as well as 0.98 and 1.10 for MNSQ INFIT values and MNSQ OUTFIT items. It is categorized as having a good value because the ideal value is 1 (the closer to 1 the better). The value of INFIT ZSTD and OUTFIT values are 0.99 and 1.21, respectively for persons as well as 0.98 and 1.10 for MNSQ INFIT values and MNSQ OUTFIT items. It is also categorized as having a good value because the ideal value is 1 (the closer to 1 the better). The value of INFIT ZSTD and OUTFIT ZSTD in sequence person and item are 0.0, 0.2, -0.1, 0.3. The ZSTD value is ideally 0.0, so that the ZSTD value including ideal except for the value of INFIT ZSTD in the item shows a negative value (not good).

Table 5. Output Reliability of Rasch Model

	Measured Person			
	Infit		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
Mean	0.99	0.0	1.21	0.2
Separation			3.11	
Person Reliability			0.88	
	Measured Item			
	Infit		**Outfit**	
	MNSQ	**ZSTD**	**MNSQ**	**ZSTD**
Mean	0.98	-0.1	1.18	0.3
Separation			6.37	
Item Reliability			0.97	
KR-20 Test Reliability			0.94	

Separation

The greater the value of separation, the quality of the instrument in terms of overall respondents and grain is getting better. The separation value on the items developed is 8.45 by entering the formula H that has been explained. Score 8.45 rounded up to 8, which means that eight groups of items can be interpreted as groups of varied items.

Conclusion

This test instrument has been proven for content validity, construct validity, inter-rater reliability, and reliability with the Rasch model. The test instrument has fulfilled the content validity with expert judgment as evidenced by the acquisition of agreement index (Aiken index) ranging from 0.50 to 1.00. The lowest score (0.5) is caused by each value's interconsistence. The raw variance value in the analysis of the Rasch model's construct validity is 73.2% with a special category. Variance values that cannot be explained are less than 10%, consecutively 3.7; 3.0; 2.9; 2.5; 2.2 indicating that unidimensionality in the instrument is in a good category. The parameters used to test model fit are CFI, NFI, RMSEA, and the loading factor value. Some results values are over 0.90 (CFI=0.92; NFI=0.90) and RMSEA is 0.00, and more than 0.3 of loading factor value on each item which indicates that the variable has good validity to the construct. The test instrument increases the number of reliability coefficients at each step of the trial, i.e. 0.17, 0.82, and 0.94. The characteristics of the Rasch model items analyzed can reveal interpretations in terms of items, personnel, and instruments. Thus, the chemistry test items developed are tested to be valid, reliable and have adequate characteristics.

References

Amalia, N. F., & Susilaningsih, E. (2014). Pengembangan instrumen penilaian keterampilan berpikir kritis siswa SMA pada materi asam basa. *Jurnal Inovasi Pendidikan Kimia*, *8*(2), 1280–1389. Retrieved from https://journal.unnes.ac.id/nju/index.php/JIPK/article/view/4443

Ardiyanti, D. (2016). Aplikasi model Rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karir siswa. *Jurnal Psikologi*, *43*(3), 248–263. https://doi.org/10.22146/jpsi.17801

Arifin, W. N., Yusoff, M. S. B., & Naing, N. N. (2012). Confirmatory factor analysis (CFA) of USM Emotional Quotient Inventory (USMEQ-i) among medical degree program applicants in Universiti Sains Malaysia (USM). *Education in Medicine Journal*, *4*(2), 1–22. https://doi.org/10.5959/eimj.v4i2.33

Astuti, R., Sunarno, W., & Sudarisman, S. (2016). Pembelajaran IPA dengan pendekatan ketrampilan proses sains menggunakan metode Eksperimen Bebas Termodifikasi dan Eksperimen Terbimbing ditinjau dari sikap ilmiah dan motivasi belajar siswa. *Proceeding Biology Education Conference*, *13*(1), 338–345. Retrieved from https://jurnal.uns.ac.id/prosbi/article/view/5742

Banne, K. (2018). Meningkatkan aktivitas belajar kimia (Redoks) siswa kelas XII TKR SMK Negeri 1 Sumarorong melalui penerapan model pembelajaran kooperatif tipe NHT dengan materi berbasis kontekstual. *Jurnal MEKOM (Media Komunikasi Pendidikan Kejuruan)*, *5*(1), 45–50. https://doi.org/10.26858/mekom.v5i1.8223

Bhakti, Y. B. (2015). Pengaruh jumlah alternatif jawaban dan teknik penskoran terhadap reliabilitas tes. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, *5*(1), 1–13. https://doi.org/10.30998/formatif.v5i1.168

Damelin, D. (2017). Using technology to enhance NGSS-aligned assessment tasks for classroom formative use. Retrieved from The Concord Consortium website: https://concord.org/newsletter/2017-spring/using-technology-enhance-ngss-aligned-assessment-tasks/

Dewi, P. C. P., & Sukadiyanto, S. (2015). Pengembangan tes keterampilan olahraga woodball untuk pemula. *Jurnal*

Keolahragaan, *3*(2), 228–240. https://doi.org/10.21831/jk.v3i2.6254

Dockrell, S., O'Grady, E., Bennett, K., Mullarkey, C., Mc Connell, R., Ruddy, R., … Flannery, C. (2012). An investigation of the reliability of Rapid Upper Limb Assessment (RULA) as a method of assessment of children's computing posture. *Applied Ergonomics*, *43*(3), 632–636. https://doi.org/10.1016/j.apergo.2011.09.009

Government Regulation No. 32 of 2013, on National Education Standard. , (2013).

Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analisis properti psikometri subtes Merkaufgaben (ME) dengan Rasch model. *Jurnal Psikologi*, *14*(1), 40–49. https://doi.org/10.24014/jp.v14i1.4900

Hasnah, H. (2017). Analisis kualitas soal matematika Ujian Sekolah kelas XII IPA SMA Negeri di Watansoppeng berdasarkan Teori Respon Butir. *PEP Educational Assessment*, *1*(1), 27–33. Retrieved from https://ojs.unm.ac.id/UEA/article/view/3776

Hayati, S., & Lailatussaadah, L. (2016). Validitas dan reliabilitas instrumen pengetahuan pembelajaran aktif, kreatif dan menyenangkan (PAKEM) menggunakan model Rasch. *Jurnal Ilmiah Didaktika*, *16*(2), 169–179. https://doi.org/10.22373/jid.v16i2.593

Iskandar, A. (2017). *Teknik analisis validitas konstruk dan reliabilitas instrument test dan non test dengan software LISREL*. https://doi.org/10.31227/osf.io/nbhxq

Ismail, I., Permanasari, A., & Setiawan, W. (2016). STEM virtual lab: An alternative practical media to enhance student's scientific literacy. *Jurnal Pendidikan IPA Indonesia*, *5*(2), 239–246. https://doi.org/10.15294/jpii.v5i2.5492

Kadir, A. (2015). Menyusun dan menganalisisi tes hasil belajar. *AL-TA'DIB : Jurnal Kajian Ilmu Kependidikan*, *8*(2), 70–81. https://doi.org/10.31332/atdb.v8i2.41 1

Khumaedi, M. (2012). Reliabilitas instrumen penelitian pendidikan. *Jurnal Pendidikan Teknik Mesin*, *12*(1), 25–30. Retrieved from https://journal.unnes.ac.id/nju/index.php/JPTM/article/view/5273

Kusaeri, K., Sutini, S., Suparto, S., & Wardah, F. (2019). The validity and inter-rater reliability of project assessment in mathematics learning. *Beta: Jurnal Tadris Matematika*, *12*(1), 1–13. https://doi.org/10.20414/betajtm.v12i1.266

Lia, R. M. (2019). *Pengembangan butir soal Kimia berorientasi NGSS dan analisisnya menggunakan model Rasch*. Master thesis, Universitas negeri Semarang, Semarang.

Lia, R. M., & Isnaeni, I. (2018). Evaluation of Chemistry learning programs at vocational high school Semarang on Vehicle Engineering field. *Proceedings of the International Conference on Science and Education and Technology 2018 (ISET 2018)*, 403–407. https://doi.org/10.2991/iset-18.2018.82

Linacre, J. M. (2016). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia - Social and Behavioral Sciences*, *204*, 164–171. https://doi.org/10.1016/j.sbspro.2015.08.129

National Research Council. (2013). *Next Generation Science Standards: For states, by states*. https://doi.org/10.17226/18290

Nurcahyo, F. A. (2016). Aplikasi IRT dalam analisis aitem tes kognitif. *Buletin Psikologi*, *24*(2), 64–75. https://doi.org/10.22146/buletinpsikologi.25218

Othman, N. B., Salleh, S. M., Hussein, H., & Wahid, H. B. A. (2014). Assessing construct validity and reliability of competitiveness scale using Rasch model approach. *The 2014 WEI International Academic Conference Proceedings*, 113–120. Retrieved from https://www.westeastinstitute.com/wp-content/uploads/2014/06/Suria-Mohd-Salleh.pdf

Pancoro, N. H. (2011). Karakteristik butir soal ulangan kenaikan kelas sebagai persiapan bank soal Bahasa Inggris. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1), 92–114. https://doi.org/10.21831/pep.v15i1.1089

Penuel, W. R., Harris, C. J., & DeBarger, A. H. (2015). Implementing the Next Generation Science Standards. *Phi Delta Kappan*, *96*(6), 45–49. https://doi.org/10.1177/0031721715575299

Pinilih, F. W., Budiharti, R., & Ekawati, E. Y. (2013). Pengembangan instrumen penilaian produk pada pembelajaran IPA untuk siswa SMP. *Jurnal Pendidikan Fisika*, *1*(2), 23–27. Retrieved from https://jurnal.fkip.uns.ac.id/index.php/pfisika/article/view/2798

Prabowo, A., & Ristiani, E. (2011). Rancang bangun instrumen tes kemampuan keruangan pengembangan tes kemampuan keruangan Hubert Maier dan identifikasi penskoran berdasar teori Van Hielle. *Kreano, Jurnal Matematika Kreatif-Inovatif*, *2*(2), 72–87. https://doi.org/10.15294/kreano.v2i2.2618

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Reiser, B. J. (2013). What professional development strategies are needed for successful implementation of the Next Generation Science Standards. *The Invitational Research Symposium on Science Assessment*, 1–23. Retrieved from http://www.ets.org/Media/Research/pdf/reiser.pdf

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2010). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Rusilowati, A. (2014). *Pengembangan instrumen penilaian*. Semarang: Unnes Press.

Sabekti, A. W., & Khoirunnisa, F. (2018). Penggunaan Rasch model untuk mengembangkan instrumen pengukuran kemampuan berpikir kritis siswa pada topik ikatan kimia. *Jurnal Zarah*, *6*(2), 68–75. https://doi.org/10.31629/zarah.v6i2.724

Sohail, M. S., & Jang, J. (2017). Understanding the relationships among internal marketing practices, job satisfaction, service quality and customer satisfaction: An empirical investigation of Saudi Arabia's service employees. *International Journal of Tourism Sciences*, *17*(2), 67–85. https://doi.org/10.1080/15980634.2017.1294343

Sudrajat, D. (2016). Portofolio: Sebuah model penilaian dalam Kurikulum Berbasis Kompetensi. *Intelegensia*, *1*(2), 1–9. Retrieved from http://ejurnal.unikarta.ac.id/index.php/intelegensia/article/view/257

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Komunikata.

Utami, B. N. (2018). *Praktik evaluasi penyuluhan pertanian*. Malang.

Wibisono, S. (2014). Aplikasi model Rasch untuk validasi instrumen pengukuran fundamentalisme agama bagi responden muslim. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, *3*(3), 729–750. https://doi.org/10.15408/jp3i.v3i3.10731

Zehir, C., Akyuz, B., Eren, M. S., & Turhan, G. (2013). The indirect effects of servant leadership behavior on organizational citizenship behavior and job performance: Organizational justice as a mediator. *International Journal of Research in Business and Social Science (2147-4478)*, *2*(3), 1–13. https://doi.org/10.20525/ijrbs.v2i3.68

REiD (Research and Evaluation in Education), 6(1), 2020, 51-65
Available online at: http://journal.uny.ac.id/index.php/reid

Item parameters of Yureka Education Center (YEC) English Proficiency Online Test (EPOT) instrument

[1]Endrati Jati Siwi; *[1]Rosyita Anindyarini; [1]Sabiqun Nahar
[1]Yureka Education Center Yogyakarta
Jl. Palem Hijau No. 120, Sidoarum, Godean, Sleman, Yogyakarta 55264, Indonesia
*Corresponding Author. E-mail: anin@eurekatour.com

Submitted: 1 April 2020 | *Revised: 29 April 2020* | *Accepted: 15 May 2020*

Abstract

Yureka Education Center (YEC) is one of the institutions which has developed an online-based English proficiency test. The test is called the English Proficiency Online Test (EPOT) which follows the TOEFL ITP (Institutional Testing Program) framework. Thus, this study aimed to analyze the characteristics of EPOT instruments consisting of Listening, Structure, and Reading subtests, which later the quality of each EPOT test item is identified. This study used a descriptive quantitative approach by describing the characteristics of EPOT test items in terms of item difficulty index, item discrimination index, test information's function, and test measurement's errors. The data were collected through EPOT trials conducted by 2,652 online test-takers as participants from 20 provinces in Indonesia. The collected data were then analyzed using the Item Response Theory (IRT) approach using the BILOG program on all logistic parameter models which began with the item compatibility test against the model. Based on the results of the analysis, all subtests match the 3-PL model. Most of EPOT's test items had a good range of difficulty index and discrimination index. The EPOT information's function shows that accurate items are used on the 3-PL model for a certain capability range. This study is expected to point out that the EPOT test could be used as an alternative English proficiency test that is easy to use and useful.

Keywords: *analysis, parameter, EPOT, listening, structure, reading*

How to cite: Siwi, E., Anindyarini, R., & Nahar, S. (2020). Item parameters of Yureka Education Center (YEC) English Proficiency Online Test (EPOT) instrument. *REiD (Research and Evaluation in Education), 6*(1), 51-65. doi:https://doi.org/10.21831/reid.v6i1.31013.

Introduction

In this era of globalization or better known as free trade, each individual is required to prepare reliable skills, especially in the communication field. In the current situation, English has a big role related to global communication between countries. Therefore, each individual is expected to be able to master English actively both oral and written. As in Indonesia, English is one of the foreign languages learned at school. Nowadays, foreign languages, especially English, have an important role, especially in careers. The working world will give high appreciation to the people who have good English ability (Handayani, 2016, p. 106). English ability is needed for various job positions, such as teachers, employees, receptionists, security guards, programmers, and job seekers. Many companies, government agencies, including the selection process for civil servant candidates (*Calon Pegawai Negeri Sipil* or CPNS) require English proficiency, one of which is proved by a Test of English as a Foreign Language (TOEFL) certificate (Arnani, 2019).

In addition to functioning as a requirement for studying abroad and applying for work, the usage of TOEFL in Indonesia has an additional function as a test instrument. This addition gives a chance for several institutions to develop and organize a test measuring an individual's English proficiency level. Sharpe states that there are 180 countries that take the TOEFL test every year in language institutions spread throughout the world (Sharpe, 2002, p. 3).

Yureka Education Center (YEC) is one of the institutions which develop English proficiency tests as a test instrument following one of ETS products, TOEFL ITP (Institutional Testing Program). English Proficiency Online Test (EPOT) is a TOEFL Prediction Test which has been developed by YEC since 2018. As the name implies, EPOT measures an individual's English proficiency level in three aspects which are Listening, Structure and Written Expression, and Reading skills which can be done online.

EPOT gives several benefits for the test takers. One of the benefits is that the test can be done almost anywhere and anytime, as long as the test takers are connected to the internet. Moreover, the result of EPOT can be delivered instantly after the test ends. Test takers will receive a digital certificate sent to their registered email. EPOT is a web-based proficiency test, therefore, the test takers are not required to download any software or applications. They can take the test using a web browser on their laptops or personal computers.

EPOT has a test structure which refers to TOEFL ITP, consisting of three sections, namely: Listening Comprehension, Structure and Written Expression, and also Reading Comprehension. EPOT is held for 115 minutes. The exercises are in multiple-choice with four answer choices. Table 1 is a comparison table of the number of questions and estimation time between TOEFL ITP and EPOT YEC.

To find out the quality of EPOT YEC test items, it is necessary to prove that each EPOT's test item is also capable of measuring someone's English proficiency as TOEFL ITP. The researchers verified each EPOT's test item using Item Response Theory (IRT) since the developed EPOT's test items do not depend on the ability of the test takers and vice versa. This means that the items' level of difficulty and discrimination do not depend on the test-takers (Anderson & Morgan, 2008, p. 76; Olufemi, 2013, p. 378; Yang & Kao, 2014, p. 171). In addition, Fan also said that the analysis using IRT emphasizes more on the level of test items' information, whereas, in classical test theory, the analysis emphasizes more on the level of the test's set information (Fan, 1998, p. 359). Thus, an analysis using IRT will give more detailed and accurate results (Pollard, Dixon, Dieppe, & Johnston, 2009, p. 3).

EPOT's items produce data with dichotomous scores in the form of correct (1) and incorrect (0). For dichotomous data, it can be analyzed using a latent linear model, perfect scale model, latent distance model, normal ogive parameter model, as well as the logistic parameter (de Ayala, 2009, p. 120; van der Linden & Hambleton, 1996, p. 18). This analysis of EPOT's test items chooses to use the parameter logistic model because the mathematical calculation is simpler using a logistic distribution model than using a normal distribution (Chung, 2005, p. 41).

Table 1. The Comparison between TOEFL ITP and EPOT YEC

Section	**TOEFL ITP**	**EPOT YEC**
Section 1: *Listening Comprehension*	Number of questions: 50 (35 minutes)	Number of questions: 50 (35 minutes)
Section 2: *Structure & Written Expression*	Number of questions: 40 (25 minutes)	Number of questions: 40 (25 minutes)
Section 3: *Reading Comprehension*	Number of questions: 50 (55 minutes)	Number of questions: 50 (55 minutes)

Several previous studies about item analysis to measure the cognitive skills of the students used classical test theory. Still, the analysis using classical test theory did not yield enough information to find out the effectiveness of test items. The reason was the existing assumptions that could not be met. Item statistics depended on the test takers' characteristics and standard error of estimator score which applied to all of the test takers. Therefore, there was no estimator score for each of the test-takers and test items. Nowadays, there are several studies which are using IRT because this theory is considered to be more detailed and valid to reveal the test items' quality.

The main advantages of IRT are that (1) the item parameters are invariant function or the response curve unchanged; and (2) the item selection can be done based on the amount of item information and test information (Hambleton, Swaminathan, & Rogers, 1991, p. 7). According to Naga, there are two types of parameters that are related to one another. In this case, participant characteristic parameters can be known if the parameter characteristics of the items are known or also known as a logistic model estimation. This model estimation is then developed into a logistic model one-to-three parameter. Likewise, the parameter features of the items can be measured if the parameter characteristics of the participants are known as the maximum likelihood estimation or the estimation of the maximum probability of occurrence (Naga, 1992).

According to the logistic distribution, IRT model is classified based on the number of test item's parameter into three types namely one-parameter logistic model (1-PL), two parameters logistic model (2-PL), and also three-parameter logistic model (3-PL) (Hambleton, 1989, p. 148; Hambleton et al., 1991, p. 7; Magis, 2013, p. 305). The 1-PL model only has one parameter which is the level of difficulty; the 2-PL model has two parameters, namely, the level of item difficulty and discrimination index; while the 3-PL model displays the parameter of difficulty index, discrimination index, and also pseudo-guessing.

Item difficulty index (b) shows the difficulty level of an item. Item discrimination index (a) shows how each test item differentiates test takers' ability in answering that test item. Meanwhile, pseudo-guessing (c) shows the probability of test-takers with low ability to correctly answer a test item. In order to apply the theory, the researchers need to determine a suitable model with the analyzed data. For statistical model selection, from the three models, then the compatibility of the items was made based on the Chi-square values. If an item has a probability of the Chi-square value ≥0.05, then that item is considered fit or compatible with the model. For this reason, the logistic model in data that has the most compatible items will be chosen as the model for data analysis (Retnawati, 2014, p. 25).

A research of the Test of English Proficiency (TOEP) developed by *Direktorat Pendidikan SMA* or the Directorate of Senior Secondary Education has been done by several researchers using Three-Parameter Logistics (3PL). It was in contrast with test items developed by private English courses. Currently, there are many institutions which offer online TOEFL Prediction test which can be easily accessed. However, the quality of test items they developed cannot be validated since it was not tested and evaluated properly. There were many test takers like college students or fresh graduates who have taken these tests to find out their English proficiency. As one of the institutions which develop TOEFL Prediction like test called English Proficiency Online Test (EPOT) and an online course, YEC makes serious efforts to analyze its test items using the IRT approach. This study was conducted to analyze and describe the parameter of EPOT's test items based on the parameter logistics which suited to the responses of EPOT's test-takers.

Method

The study is aimed at finding out the parameters or the characteristics of EPOT's test items through the trial results. The parameter of EPOT's test items can be observed from the difficulty, discrimination, and also pseudo-guessing level of each test item. There were 2,652 participants from 20 provinces throughout Indonesia which become the research subjects. Most of them are fresh graduates who wanted to apply for a job and students who wanted to continue their study. A simple random sampling technique was used in order to gather samples from the population. The samples were picked randomly neglecting any difference in the population. This method is used if the members of a population are considered homogeneous (Sugiyono, 2014). The samples were fresh graduate students from bachelor level with the minimum age of 23 years old. Most of the samples were taking EPOT since they needed a TOEFL certificate to apply for job vacancies or to continue their studies. Others were taking EPOT to test their proficiency level since EPOT's framework is equivalent to the TOEFL ITP.

All of the research subjects took EPOT online test through the official Yureka Education Center's website yec.co.id. A set of EPOT test consists of 50 listening comprehension questions, 40 questions of structure and written expression, and 50 questions of reading comprehension. The test should be done in 115 minutes. Previously, the testing of EPOT's validity and reliability has been conducted. The content validity testing was done by three English experts, examining the content and structure of the test. The results of the validity testing showed that there were four test items that were not valid since their Aiken's V index was less than 0.67 (Azwar, 2017, p. 113). These four items were then being revised and tested again to achieve a good Aiken's V index. The distribution of Aiken's V value is shown in Figure 1.

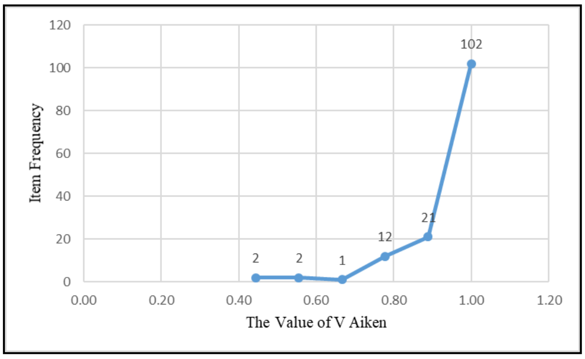


Figure 1. V Aiken Value Distribution

The face validity test was conducted by two experts on learning media. The experts examined the test appearance and the item context compatibility with the aim of the test. As the results, for the test appearance, YEC should add a button to change audio volume; recheck the audio playback; change the test instructions' placement; fix the test items' placement; fix the consistency of font size; and fix the writing whether it should be capital, italic, or bold. After the revision was done and the appearance of the test was improved, it can be considered that the face validity has been met (Azwar, 2017, p. 43). The reliability test of EPOT showed that it has Cronbach's Alpha score of 0.908. It meant that 90.8% of the observed score variant resembled the true score. According to the literature, the reliability score of 0.908 showed that EPOT's test instrument has good reliability (Gliem & Gliem, 2003; Guilford, 1956). Therefore, the developed EPOT's test instrument is assumed to highly reliable. The results of the reliability test are shown in Table 2.

Table 2. Reliability Index

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.908	.910	140

The item analysis on EPOT used the logistic parameter model. In IRT theory, the item's difficulty level can be labeled as good if the value is in the range -2 up to 2 (de Ayala, 2009, p. 15; Fan, 1998; Hambleton et al., 1991, p. 13). Theoretically, the item discrimination index is in the scale $-\infty \leq a \leq \infty$, but practically, the a value is in the range 0 up to 2 (Hambleton et al., 1991, p. 15). Meanwhile, c value was considered as a good item if it is in the range of 0 up to 1 or 1/k that k is the total answer choices (Hulin, Drasgow, & Parsons, 1983). After going through the comparison process from the three logistic parameters, the 3-PL model was considered to be the most suitable model for EPOT trial result data.

The item analysis used Bilog-MG software. The computer program for maximum likelihood estimation was the Bilog-MG fit program that was used for one, two, or three-parameter model. The Bilog-MG program was able to estimate multiple-choice items and also for estimating latent skills in huge amounts (Crocker & Algina, 1986, p. 354; Hambleton et al., 1991, pp. 43–50; Yen & Fitzpatrick, 2006, pp. 131–132). Based on the output of the Bilog-MG program, it could be obtained item difficulty index (b) or threshold, item discrimination index (a) or slope, and pseudo guessing (c) or asymptote. The difficulty index, discrimination index, and the ability of items to be guessed by a participant will be shown in a graph. Besides, the Item Characteristics Curve (ICC) graph would show the quality of several items, and the Test Information Curve (TIC) graph will show the quality of EPOT.

Findings and Discussion

EPOT consists of three sections, namely Listening Comprehension, Structure and Written Expression, and Reading Comprehension. The summary of difficulty index, discrimination index, and matched item can be seen in Table 3.

If the data are accumulated in 1-PL, there will be only 71 items from Listening, Structure, and Reading which has Chi-square ≥ 0.05. In the 2-PL model, there are 117 items which have Chi-square ≥ 0.05. Meanwhile, in the 3-PL model, there are 123 items which have Chi-square ≥ 0.05 or can also be considered as fit items. In conclusion, the logistic model that fits the EPOT test-takers answers results is the 3-PL model. The selection of the 3-PL model is also caused by some test-takers who already fulfilled the requirements for the use of the 3-PL model. Other than that, it also reinforces the assumption that proficiency tests using multiple-choice formats are examples of situations where the 3-PL model is suitable. Test takers tend to choose the best answer which they found most interesting if they could not find the correct answer, so the guessing factor is considered in this study (Huriaty, 2019, pp. 35–36).

Table 3. Summary of Item Parameters' Characteristics and Matched Item Analysis

Section	Model	Item's Description	Number of Good Item/ Item Fit	Percentage
Listening	1PL	*b*	49	98%
		Fit Item	27	54%
	2PL	*a*	45	90%
		b	48	96%
		Fit Item	45	90%
	3PL	*b*	46	92%
		a	50	100%
		c	10	20%
		Fit Item	48	96%
Structure	1PL	*b*	34	85%
		Fit Item	25	62.5%
	2PL	*a*	35	87.5%
		b	40	100%
		Fit Item	26	65%
	3PL	*b*	40	100%
		a	39	97.5%
		c	12	30%
		Fit Item	27	67.5%
Reading	1PL	*b*	44	88%
		Fit Item	19	38%
	2PL	*a*	46	92%
		b	45	90%
		Fit Item	46	92%
	3PL	*b*	44	88%
		a	49	98%
		c	3	6%
		Fit Item	48	96%

The first section, Listening, consists of 50 questions with a duration of 35 minutes. Based on the test-takers' response data, it is found out that EPOT Listening has various difficulty index, discrimination index, and pseudo-guessing which can be seen in Figure 2, Figure 3, and Figure 4.

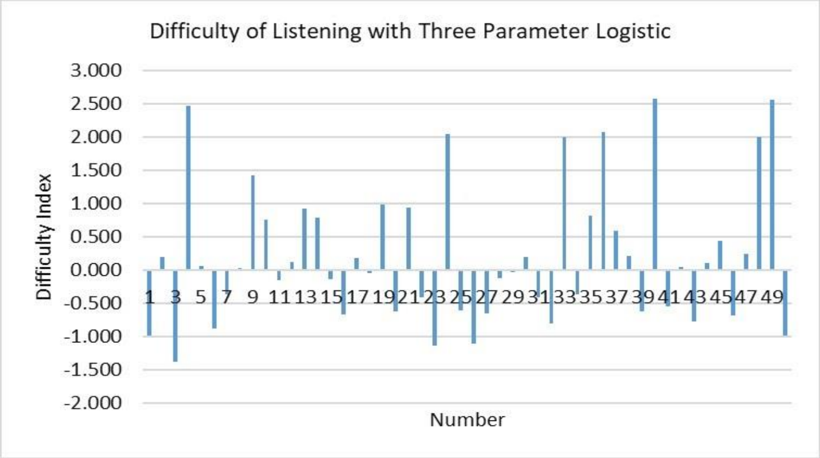


Figure 2. Difficulty Index of EPOT Listening

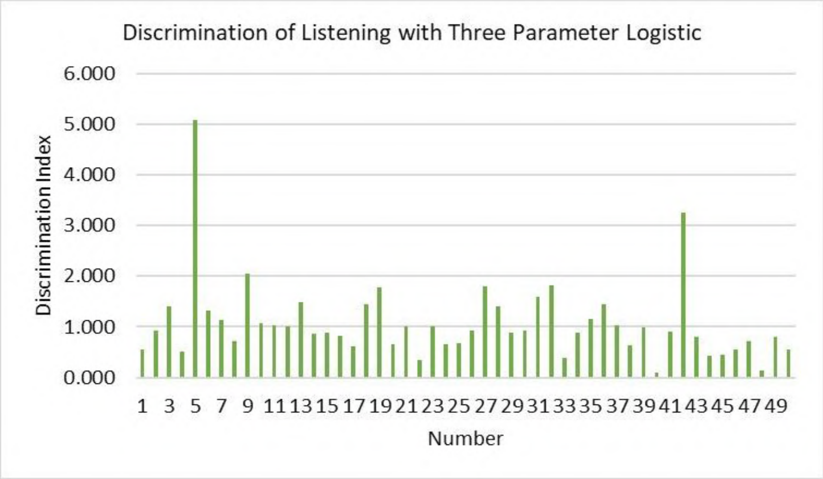


Figure 3. Discrimination Index of EPOT Listening

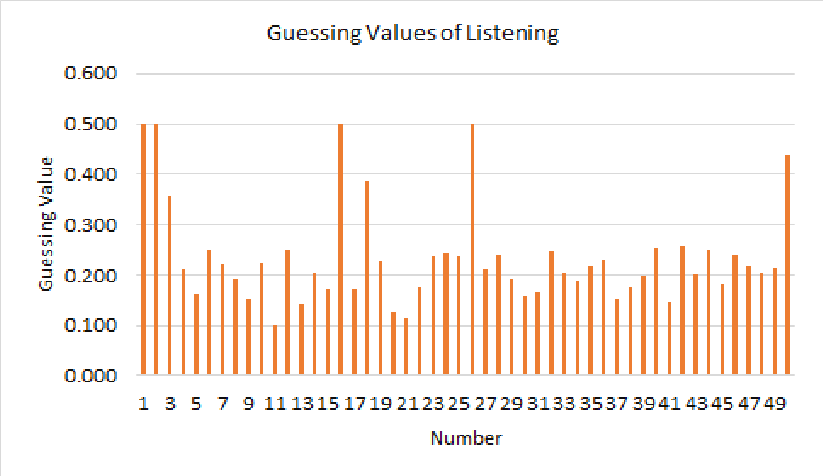


Figure 4. Pseudo Guessing Values of EPOT Listening

According to Figure 2, it can be concluded that there are 46 items out of 50 which have good difficulty index while four items are considered as poor. Those four items are number 4 (b = -2.473), number 36 (b = 2.068), number 40 (b = 2.572) and number 49 (b = 2.552). Number 36, 40 and 49 are considered too difficult because the b > 2, while number 3 is considered too easy because b < 2. It causes the answer responses' patterns tend to be poor and not able to show the difficulty index parameter. In Figure 3, it can be seen that the items in the Listening section have shown the various difficulty index and are distributed well. All 50 test items show a good discrimination index with the range between 0 up to 2. Accordingly, the high and low ability of the test takers can be shown by the EPOT Listening test items.

On the other hand, Figure 4 shows that the Listening section has 43 items with good pseudo guessing. It means there are only 14% out of all items that can be answered correctly because there is an element of guessing. The next analysis is about the item fit analysis on Listening which gives an illustration in the form of Item Characteristic Curve (ICC) as presented in Figure 5 and Figure 6.

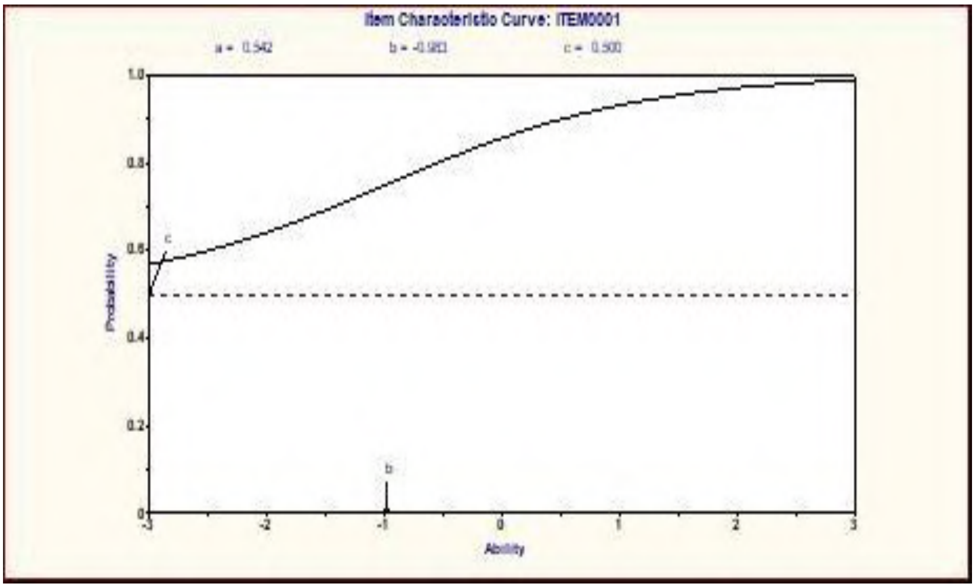


Figure 5. An ICC Example of Listening Item Number 1

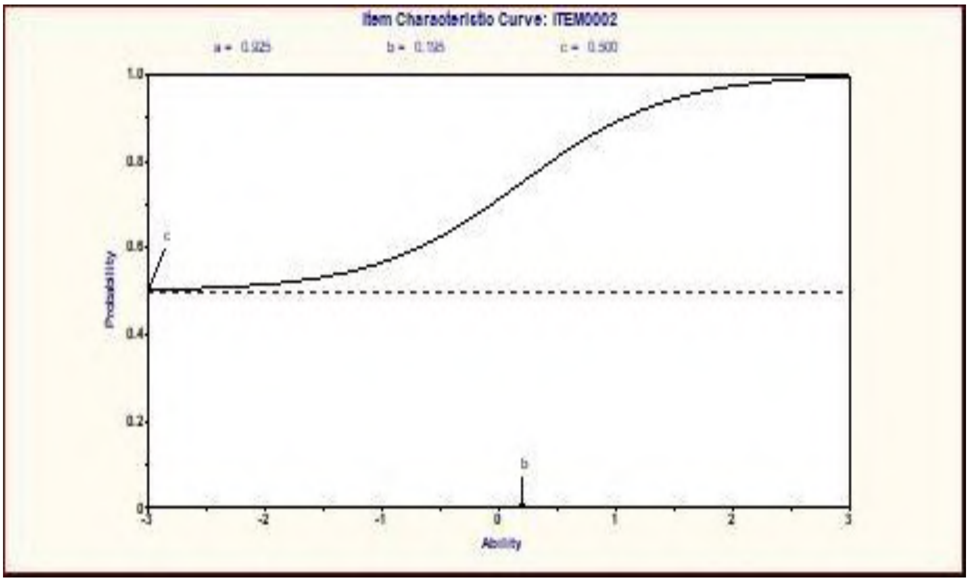


Figure 6. An ICC Example of Listening Item Number 2

Figure 5 and Figure 6 are examples of test-takers' responses pattern toward EPOT Listening test items number 1 and 2. Figure 5 shows a graph of the relationship between test takers' ability and parameter estimation item number 1 with b = -0.983; a = -0.542; and c = 0.500. Figure 6 illustrates the relationship between test takers' ability and parameter estimation item 2 with b = 0.195; a = -0.925; and c = 0.500.

EPOT Structure section consists of 40 items done in 25 minutes. According to the data of test-takers' responses, 40 items of EPOT Structure also have various difficulty and discrimination index. These findings can be seen in Figure 7, Figure 8, and Figure 9.

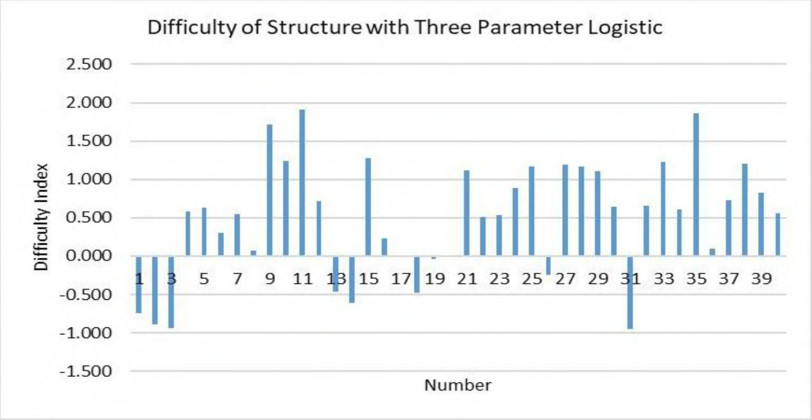


Figure 7. Difficulty Index of EPOT Structure

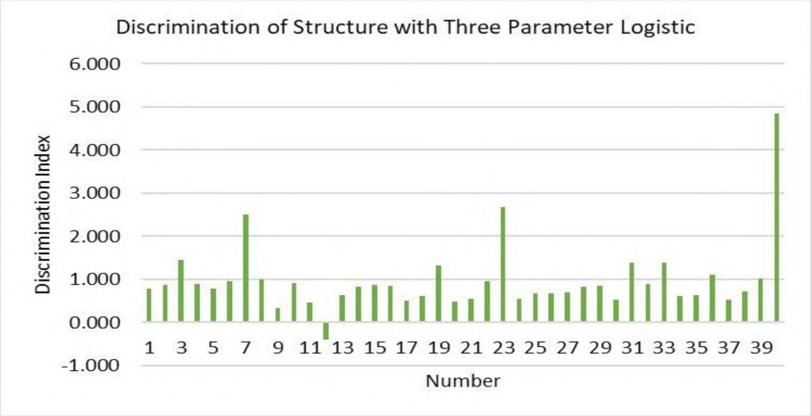


Figure 8. Discrimination Index of EPOT Structure

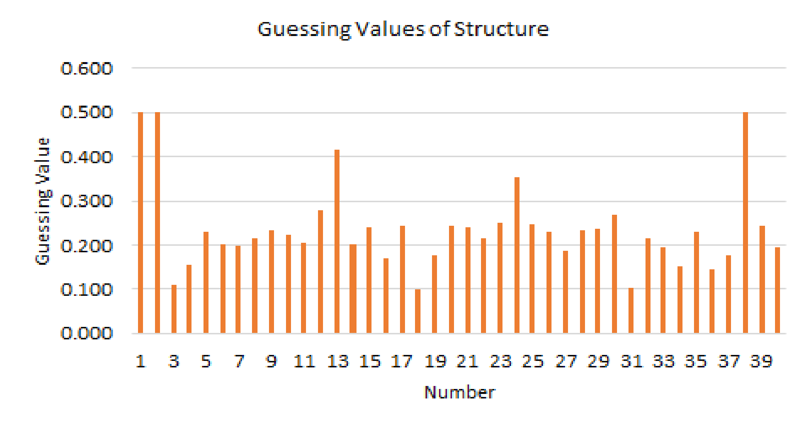


Figure 9. Pseudo Guessing Value of EPOT Structure

Figure 7 shows that all 40 EPOT Structure items have good difficulty level. In Figure 8, the 39 items have a good discrimination index. However, there is one item with a poor discrimination index, that is number 12 with a = -0.395. It shows that number 12 cannot show the difference between the low and high ability of the test takers. Meanwhile, Figure 9 shows that the Structure section has 35 items with good pseudo-guessing. In other words, there are only 12.5% out of all items that can be answered correctly because of the guessing element. The next analysis is about the item fit analysis on Structure, which gives an illustration in the form of ICC, as presented in Figure 10 and Figure 11.

Figure 10 shows the relationship graph of test takers ability and parameter estimation of item number 1 in Structure with b = 0.793; a = -0.746; and c = 0.500. Meanwhile, Figure 11 shows a relationship graph of test takers' ability and parameter estimation of EPOT Structure item number 2 with b = 0.879; a = -0.893; and c = 0.500.

The last section is Reading Comprehension. EPOT Reading section consists of 50 items that are done in 55 minutes. According to the test takers' responses, it can be concluded that 50 items of EPOT Reading also have various difficulty and discrimination index. It can be seen in Figure 12, Figure 13, and Figure 14.

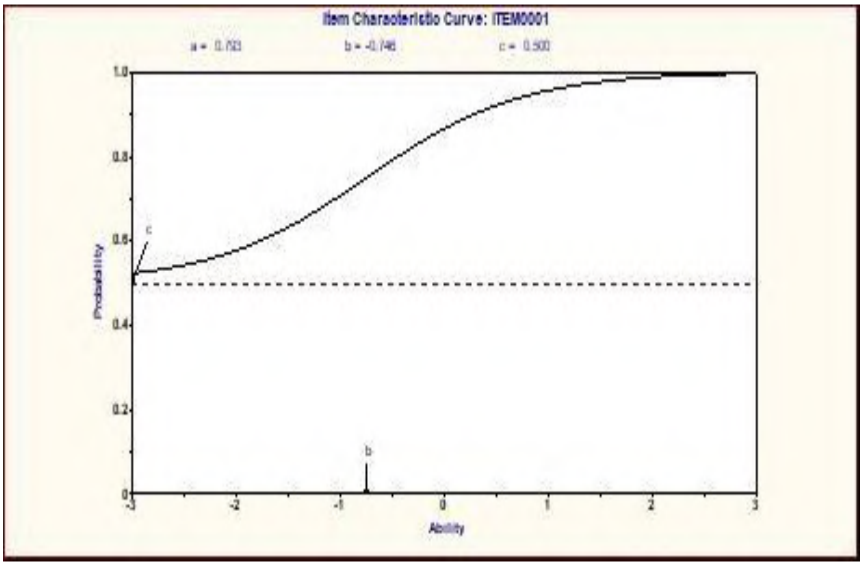


Figure 10. An ICC Example of Structure Item Number 1

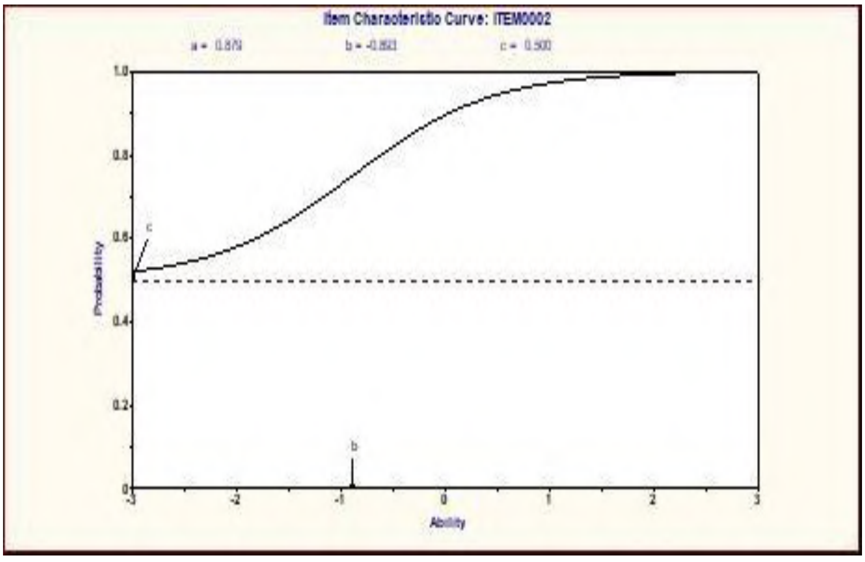


Figure 11. An ICC Example of Structure Item Number 2

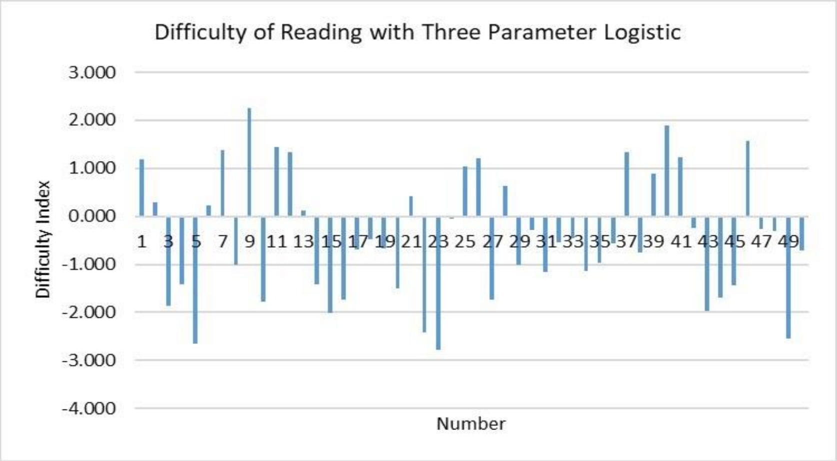


Figure 12. Difficulty Index of EPOT Reading

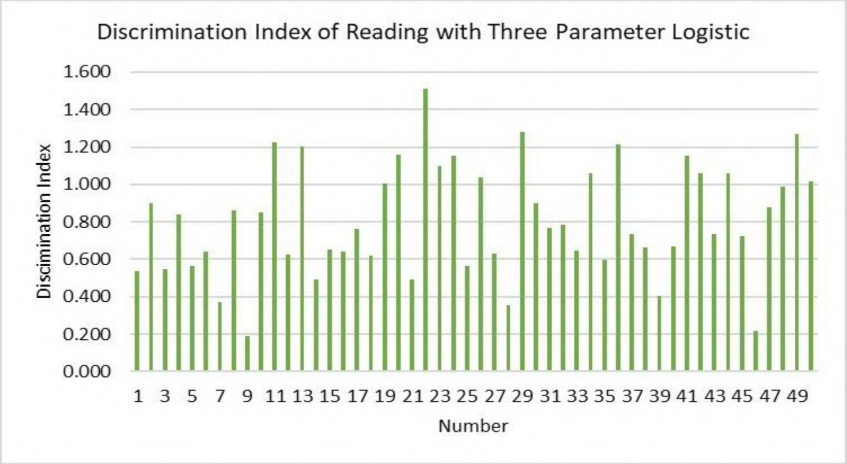


Figure 13. Discrimination Index of EPOT Reading

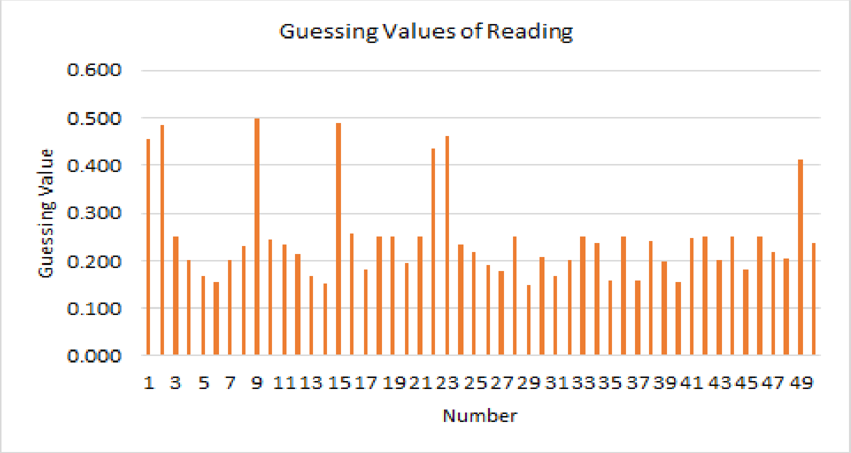


Figure 14. Pseudo Guessing Value of EPOT Reading

Based on Figure 12, 45 items have good difficulty index, and the remaining five items are considered poor. These five items are number 5 (b = -2.657), number 9 (b = 2.264), number 22 (b = -2.407), number 23 (b = -2.771), and number 49 (b = -2.547). The items number 5, 22, 23, and 29 are considered too difficult since the difficulty level is < -2; and number 9 is considered too easy because the difficulty level is > 2. Thus, the test takers' responses tend to be poor, and these items cannot show the difficulty index parameter.

Figure 13 shows that all of the items in the EPOT Reading section have good discrimination index since they are in the range of 0 to 2 so that the test takers' low or high ability can be shown in all EPOT Reading's test items. Meanwhile, Figure 14 shows that the EPOT Reading section only has 43 items with good pseudo-guessing. It means 86% of all items can be answered correctly because of the guessing elements. The next analysis is about items fit in the EPOT Listening section, which gives an illustration in the form of ICC, as shown in Figure 15 and Figure 16.

Figure 15 shows a graph between the test takers' ability and estimated parameter Reading section item number 1 with b = 0.536; a = 0.181; and c = 0.455. In addition, Figure 16 depicts a graph between the test takers' ability and estimated parameter of EPOT Reading section item number 2 with b = 0.899; a = 0.291; and c = 0.484.

The next discussion will be about information function analysis and Standard Error Measurement (SEM). The EPOT information function value will show EPOT's reliability and measurement accuracy. The EPOT information function describes a low curve that increases, reaching the highest score in the middle before falling far from the midpoint. The curve's width shows the extent of the effective capability from the measurement results.

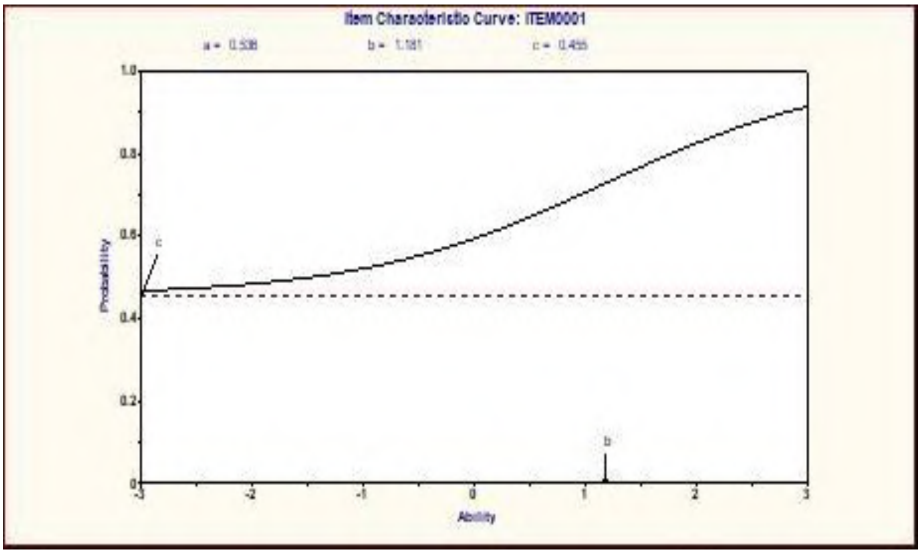


Figure 15. An ICC Sample of Reading Item Number 1

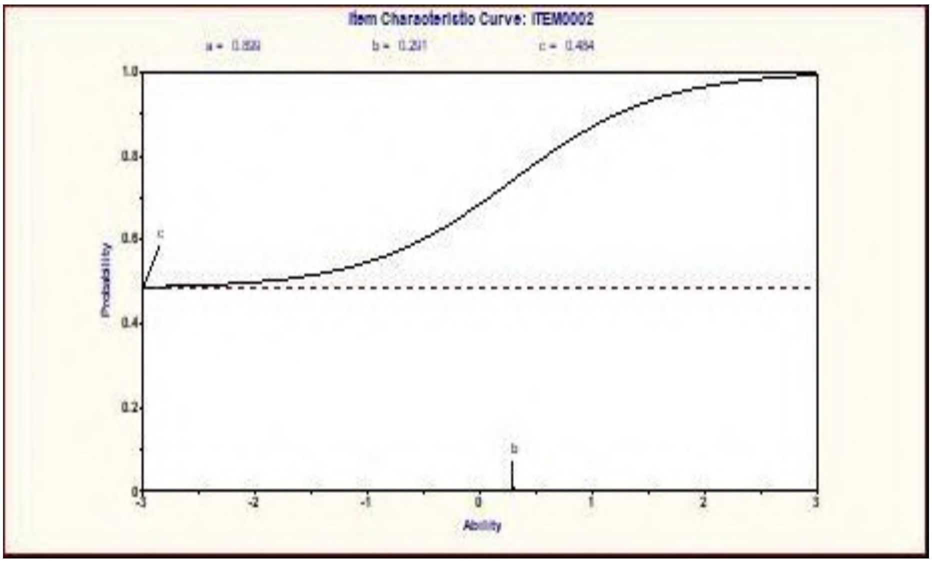


Figure 16. An ICC Sample of Reading Item Number 2

Test Information Function (TIF) will be effective if the curve line extends above the SEM line without having an intersection point. However, EPOT items' analysis yields TIF and SEM curves that have interaction between the two. These are three figures which show the Total Information Curve (TIC) for 1-PL, 2-PL, and 3-PL model.

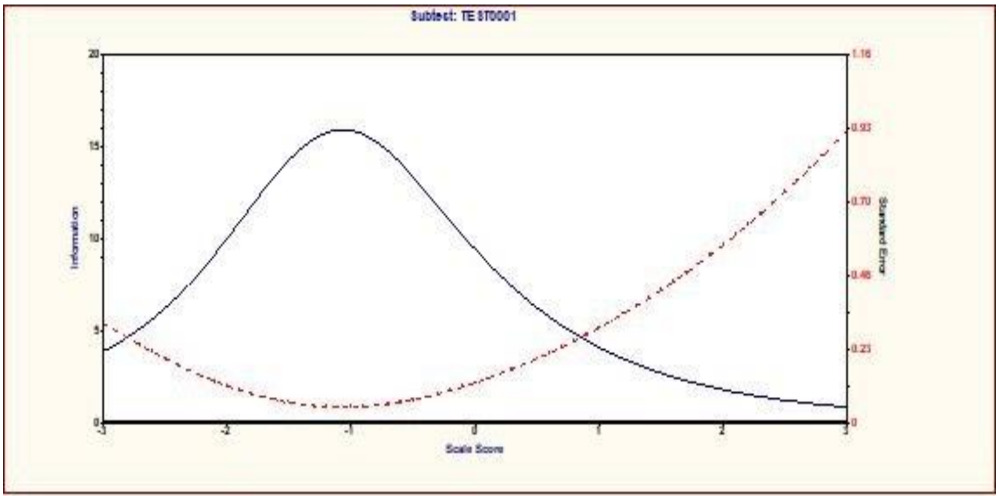


Figure 17. EPOT's TIC for 1-PL Model

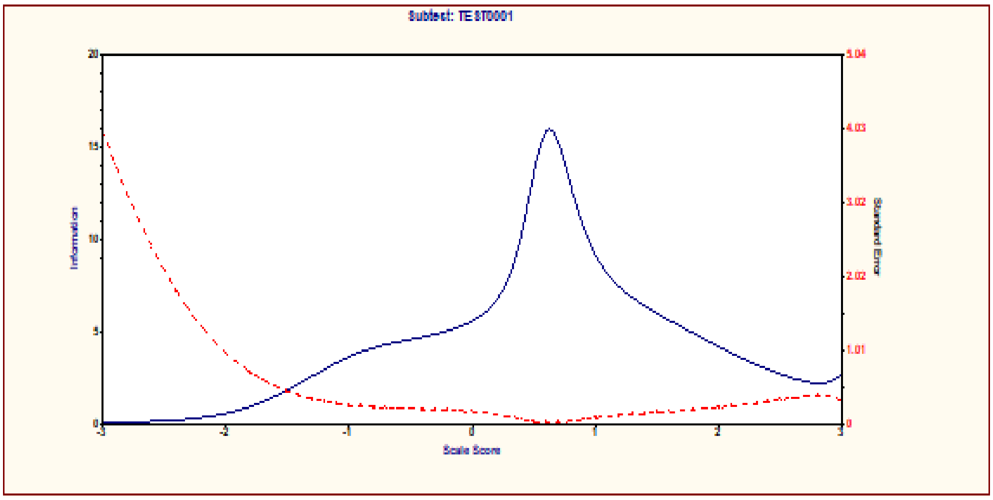


Figure 18. EPOT's TIC for 2-PL Model

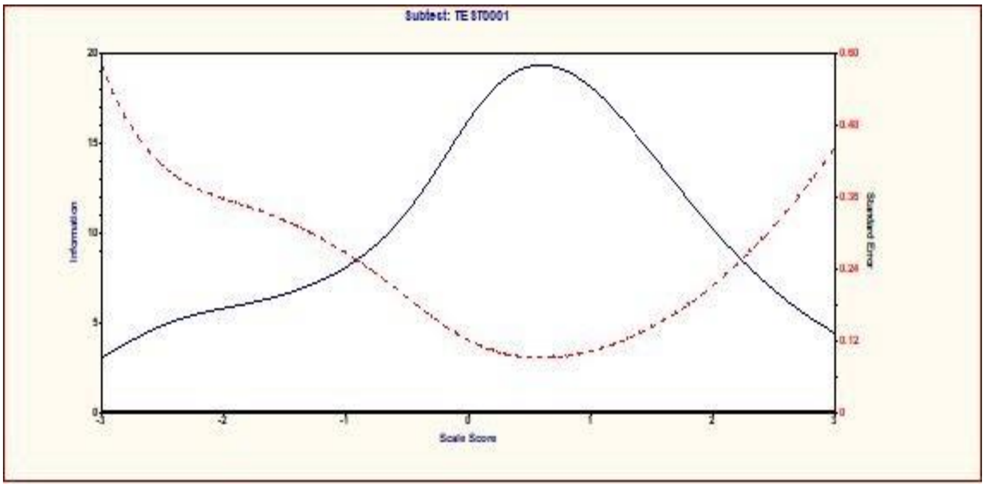


Figure 19. EPOT's TIC for 3-PL Model

Figure 17, Figure 18, and Figure 19 show TIC, which consists of the TIF line, SEM line, and interaction among them. TIC illustrates the total information produced by any level of ability. The dotted line shows SEM, which means the greater the information function, the smaller the measurement error is. The three graphs show the TIF curve above SEM with two intersection points; it means that the information obtained from the measurement results is only accurate on abilities with a certain range. This research's finding shows that the 3-PL IRT model provides the highest TIF compared to the 1-PL and 2-PL models. It is caused by the average of EPOT's items discrimination index with 3-PL model (a = 0.948) higher than the item's discrimination index with 1-PL (a = 0.777) and 2-PL (a = 0.460). In the IRT model that accommodates the presence of discrimination index, if the discrimination index gets bigger, the value of TIF obtained will be greater (Setiawati, Izzaty, & Hidayat, 2018, p. 17; Yang & Kao, 2014, pp. 173–174; Zięba, 2013, p. 96). The presence of this discrimination index causes the item information with 2-PL is higher than 3-PL. As a result, the 1-PL model that becomes the lowest because this model does not accommodate the discrimination index parameter.

Based on the previous analysis, 93% of Listening, Structure, and Reading test item has a good average of difficulty index between -2 to 2. There are 10 test items that were considered poor; they were too difficult or too easy. These items were still used to vary the test items. As stated by Hingorjo and Jaleel (2012), test items with an average difficulty index are more desirable, test items with easy level can be placed in the beginning question as warming up, and the difficult item should be reviewed to avoid language confusion.

In addition, out of the 140 EPOT's test items, one item of Structure test and one item of the Reading test had a discrimination index of > 2. The two items are not modified since the gap between the scores and also the standard score is not significant. Meanwhile, the pseudo-guessing index showed that only 19 test items can be answered correctly by the test takers, which rely solely on guessing. The results of TIF and SEM curved almost perfectly and interacted at two intersection points. The results of the study pointed out that the IRT 3-PL model provides higher test information function than the 1-PL and 2-PL model. The reason was the average of the EPOT's 3-PL discrimination index was higher than the 1-PL and 2-PL model.

Conclusion

Item analysis can give useful information related to the item characteristics of a test set. English Proficiency Online Test (EPOT) is a set of English proficiency test developed by YEC and has gone through several processes of testing and evaluation on its test items. The testing and evaluation are using a 3-PL model to show the characteristics of the test, consisting of difficulty index, discrimination index, and pseudo-guessing index.

Based on the results of EPOT's item analysis using the IRT 3-PL model, it can be concluded that most of the items have a good difficulty index. Several items that have poor difficulty index are still used to vary the test items. Moreover, EPOT's test items are also able to effectively distinguish test takers' ability and improve test takers' reliability (Nelson, 2001; Wells & Wollack, 2003). Several test items that have poor discrimination index are not modified as the gap between the scores, and the standard score is not significant. As for the pseudo guessing index, there are only a few test items that can be answered correctly by the test takers who rely on guessing. In conclusion, EPOT has sufficient quality of effective test items, and it can be employed as a TOEFL Prediction test.

References

Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement* (V. Greaney & T. Kellaghan, Eds.). https: //doi.org/10.1596/978-0-8213-7497-9

Arnani, M. (2019, November 14). CPNS 2019, 9 instansi ini wajibkan TOEFL, berapa skornya? *Kompas.Com*. Retrieved from https://www.kompas.com/tren/ read/2019/11/14/120925265/cpns-

2019-9-instansi-ini-wajibkan-toefl-berapa-skornya?page=all

Azwar, S. (2017). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.

Chung, H. (2005). *Calibration and validation of the body self-image questionnaire using the Rasch analysis*. Master thesis, University of Georgia, Athens, GA.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

de Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York, NY: Guilford Press.

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

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, 82–88. Colombus, OH: The Ohio University.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). New York, NY: McGraw-Hill.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Handayani, S. (2016). Pentingnya kemampuan Bahasa Inggris dalam menyongsong ASEAN Community 2015. *Jurnal Profesi Pendidik*, *3*(1), 102–106. Retrieved from http://ispijateng.org/wp-content/uploads/2016/05/PENTINGNYA-KEMAMPUAN-BERBAHASA-INGGRIS-SEBAGAI-DALAM-MENYONGSONG-ASEAN-COMMUNITY-2015_Sri-Handayani.pdf

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *JPMA: The Journal of the Pakistan Medical Association*, *62*(2), 142–147. Retrieved from https://jpma.org.pk/article-details/3255?article_id=3255

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones- Irwin.

Huriaty, D. (2019). Analisis karakteristik parameter butir berdasarkan model Logistik 3 Parameter. *Lentera: Jurnal Pendidikan*, *14*(2), 33–40. https://doi.org/10.33654/jpl.v14i2.885

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*(4), 304–315. https://doi.org/10.1177/0146621613475471

Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Jakarta: Gunadarma.

Nelson, L. (2001). *Item analysis for test and surveys using Lertap 5*. Perth: Curtin University of Technology.

Olufemi, A. S. (2013). Item Response Theory as a basis for measuring latent trait of interest. *Greener Journal of Social Sciences*, *3*(7), 378–382. https://doi.org/10.15580/GJSS.2013.7.062513691

Pollard, B., Dixon, D., Dieppe, P., & Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: An item analysis using classical test theory and item response theory. *Health and Quality of Life Outcomes*, *7*, 1–20. https://doi.org/10.1186/1477-7525-7-41

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi* *pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Analisis respons butir pada tes bakat skolastik. *Jurnal Psikologi*, *17*(1), 1–17. https://doi.org/10.14710/jp.17.1.1-17

Sharpe, P. J. (2002). *How to prepare for the TOEFL test: Test of English as a foreign language* (10th ed.). Jakarta: Binarupa Aksara.

Sugiyono, S. (2014). *Metode penelitian pendidikan: Pendekatan kuantitatif, kualitatif, dan R&D*. Bandung: Alfabeta.

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. https://doi.org/10.1007/978-1-4757-2691-6 I

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Madison, WI: University of Wisconsin.

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org/10.3969/j.issn.1002-0829.2014.03

Yen, W., & Fitzpatrick, A. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.

Zięba, A. (2013). The item information function in one and two-parameter logistic models – A comparison and use in the analysis of the results of school tests. *Didactics of Mathematics*, *10*(14), 87–96. https://doi.org/10.15611/dm.2013.10.08

REiD (Research and Evaluation in Education), 6(1), 2020, 66-77
Available online at: http://journal.uny.ac.id/index.php/reid

Analysis of factors of students' stress of the English Language Department

***[1]Siwi Karmadi Kurniasih; [1]Nur Hidayanto Pancoro Setyo Putro; [1]Sudiyono**
[1]Faculty of Languages and Arts, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
*Corresponding Author. E-mail: siwikarmadi@uny.ac.id

Submitted: 12 January 2020 | Revised: 2 June 2020 | Accepted: 3 June 2020

Abstract

The study is aimed at further describing psychological factors inducing academic stresses of the students of the English Education Department, Yogyakarta State University (EED-YSU). The study is a continuation of a previous study that identifies sources of academic stress of students of EED-YSU. Data collection is conducted by an on-line survey technique. Confirmatory factor analysis is used for the data analyses. The results show six factors that become sources of students' academic stresses, namely academic demands, parent-child relationship, traumatic experiences during childhood, peer pressures, financial matters, and self-expectancy. It is expected that other studies involving other factors of students' academic stress be conducted to give further information on the topic.

***Keywords**: psychological factors, academic stress, student's productivity*

How to cite: Kurniasih, S., Putro, N., & Sudiyono, S. (2020). Analysis of factors of students' stress of the English Language Department.*REiD (Research and Evaluation in Education), 6*(1), 66-77. doi:https://doi.org/10.21831/reid.v6i1.29567.

Introduction

Evidence suggests that academic stress is among the most important factors that affect university students' success. Demanding academic assignments, practicums, or theories alike, are often alleged as the causes for the elevation of students' stresses. A large number of classes that the students take in one semester also often causes students to not be able to adequately focus in classes so that they are not able to achieve the expected instructional objectives. The level of students' academic is doubly worsened by the present-day advancement of technology (Gabre & Kumar, 2012). The high level of students' academic stresses unconsciously affects the students' physical and mental health as often seen from their complaints of common tiredness. In fact, the phenomenon is like the tip of the ice burgh: students bear a huge haunting burden, unconsciously.

A study by the National College Health Assessment in 2014 shows that 55.5 percent of the students participating in the survey experience depression above the average level of the other students; in the category of high concern (American College Health Association - National College Health Assessment, 2016). Students have difficulty focusing in class and doing instructional tasks since they are too worried about small things that happen in their daily life, causing an increase in their stress level in their academic life. A similar study conducted in 2015 draws a similar conclusion that 20 percent of students seek psychological consultation and treatment for problems related to their academic stresses (Henriques, 2014).

Academic stress is experienced alike in students of the English majors. Differences in the cultures of the native and target languages, learning difficulties, and anxiety of losing self-identities are among the heavy burdens for the students of English as a foreign language (Hashemi, 2011; Horwitz, Horwitz, & Cope, 1986). These problems have been identified as triggers of English students' feeling stresses.

The level of academic stress of university students has long been studied (Yeh & Inose, 2003). The study emphasizes that most international students experienced anxieties, and English fluency is among the predictors of university students' academic stress. It is these anxieties that contribute to the factors that elevate academic stresses (Leyva, 2003; Mezzacappa & Katkin, 2002). These anxieties have caused students' self-confidence to decrease to a low level. Consequently, their performance is not maximal.

On one other side, worries of not being able to teach English in front of the class also contribute to the stress level of these English teacher candidates (McNeil, 2016), in as much as they are demanded to master the instructional materials and manage the classroom. Such anxieties can elevate to a level so high that they are often exaggerated. This anxiety does not give a good condition to students since such over-felt worries raise students' academic stress that, subsequently, give negative influences on their physical and mental health. Possible worse impacts are the high threats of a high drop-out and, in few cases, students' thoughts and intentions to end their life (Ang & Huan, 2006; Robotham & Julian, 2006).

Although extensive research has been carried out on students' academic stress, few writers have been able to draw on any systematic research into the factors affecting the academic stress of students majoring in English Education Department. Thus, this study is aimed at identifying factors causing academic stresses to elevate in students of the English Department of YSU. Levels of academic stresses will also be stress long relevant variables such as gender, Grade Point Average (GPA), years of entrance, parents' education backgrounds, degrees students tend to obtain, and socio-economic status.

Definition of Stress

From the field of educational psychology, a number of definitions of stress are found. As a first definition, Butt, Weinberg, and Horn (2003) define stress as an imbalance or gap between one's demands, physically and psychologically, for achievement and one's abilities to achieve it. The demands can either be internal or external. Failure in the fulfillment of the demands will cause physical or psychological impacts. In the same line of thoughts, Sarafino (2008) states that stress is in an individual's condition wherein there is a perceived gap between demands that come from the inside of the individual, psychologically, biologically, or socially, as a result of his interaction with the environment. Stress is a condition that influences one's physical or psychological states because of pressure from either the inside or outside of the individual.

Another definition is given by Suldo, Shaunessy, Thalji, Michalowski, and Shaffer (2009) who state that stress is an individual's feeling of pressure in responding to demands that come from the inside of the individual or from the environment. This state can be identified from the levels of blood pressures, heart pulses, or neurotransmitter hormones as an individual's physiological response against stress.

Emphasizing on the individual's emotions, Folkman (2013) defines stress as an individual's condition in which the individual experiences an over-sized emotional demand so that he/she finds difficulties in effectively functioning all his competences. This condition may give rise to psychological symptoms like chronic tiredness, depression, anxieties, and anger. Stress can also be defined as an adaptive response to differences in an individual's characteristics from external pressures and demands, leading to his physical and psychological conditions (DeFrank & Ivancevich, 1998).

From the foregoing discussion, it can be summarized that stress can be defined as an individual's mental or psychological disturbances as a result of pressures. These pressures come from the individual's failures in satisfying his demands or desires, internal or external.

Academic Stress

Academic stress can be defined as a student's perception of his over-loaded knowledge, concepts, and skills that he must master against the lack of time that he has to achieve them (Misra & Castillo, 2004). A student's academic stress is mostly related to academic tensions the student faces. This condition causes the emergence of distortion in the student's thoughts that influences his physical, emotional, and behavioral pattern of actions. This distortion can come from the student's own demands or those of the environment. Instances of these demands are daily or weekly assignments, final examinations, and competitions among students in obtaining achievements. A student's academic stress can trigger distress that is manifested in various negative psychological behaviors.

The discrepancies between what needs to be achieved in knowledge, concepts, and skills and the abilities to achieve them (Misra & McKean, 2000) cause the student to feel inadequate or uncomfortable in his interactive activities. Another definition is related to the student's misperception on his academic loads he needs to finish resulting in physical and psychological problems.

In relation to the types of academic stress, Suldo et al. (2009) mention several sources. These sources are, among others, academic requirements, parent-child relations, childhood traumatic experiences, peer pressures, extra-curricular activities, and struggles to achieve high academic standards. Each of them is elaborated as follows.

Academic Requirements

This source of academic stress can be of many forms. These can be (1) fulfillment of academic assignments such as daily quizzes, weekly tasks, and mid- and semester tests; (2) individual time management of individual academic assignments; and (3) over-expectation by self, peer, and lecturers for higher academic achievement.

Relation between Parents and Students

Causes of academic stress are often related to the relationship between parents and students concerning academic matters. This often arises from a variety of conflicts between parents and students, such as time management related to the student's responsibilities in the household.

Unpleasant Early Adulthood Experience

The next source of students' academic stress is related to changes in the lifestyle of the students during young adulthood. Such causes can be in the form of the need for safety, the transition from school life to university life, loss of a family member, awareness of more global environmental problems such as drug abuse, and the community environment which itself experience stress.

Peer Relationship

Often, the source of academic stress comes from matters related to peer relationships. This can be in the form of problems with close friends or partners, uncomfortable atmosphere in peer relation, and pressures or threats from friends.

Domestic or Family Problems

Households are abounding that they have an impact that leads to academic stress. Examples of these are conflicts between parents, parents' divorces, and others. Such problems cause students to feel unable to concentrate well on academic matters.

Extra-curricular Activities

Students' activities outside the classroom can also turn out to cause students' academic stress. These may be in the forms of students' anxieties in their lack of sport or art skills, poor time management concerning curricular and extracurricular commitments, and personal needs such as eating and sleeping patterns.

Academic Endeavours

One source of academic stress is related to students' strains in fulfilling their academic commitments (Suldo et al., 2009). These include low students' proficiency and skills, missing classes and other instructional activities, and students' health problems. Further, McPherson (2009) relates this type of academic stress into the following.

Underachievers and Overachievers

Underachievers are those students who are not yet able to achieve the expected minimal competencies in order to pass classes. In fact, these students are expected to work hard, by their own interest and motivation, and supported by families and friends. It is this demand that gives these underachievers academic stress. On the other hand, the overachievers are those students who make it to the above level of students' average. Often, they need to sacrifice their sleep and fun times to do that. For some of them, the lack of time for having recreative activities can cause academic stress.

Appreciation or Reward

Conferring awards to some students or groups of students may oftentimes induce academic stress, both for the achieving and non-achieving students. Formerly, the award-winning has drained from them too much effort and energy that they do not have time to do other things such as recreation and extracurricular activities. For the latter, over-expectation to gain such achievement is not backed up by their abilities and efforts to do so.

Loss of Rest and Recreation

Loss of time for rest and recreation consequently causes academic stress. Rest and recreation are needed to loosen their thoughts and muscles and recharge their energy for optimum concentration and work.

Expectancy

This is closely related to the reward matter. High demands from self or family to achieve well can become a quick cause of academic stress. For a high proportion of students, a high GPA is a fixed price, mandatory over everything else. Many students stake all their time and energy for high GPAs resulting in burn-outs that trigger academic stress.

Class Assignment or Project

This derives from two possibilities: unclear criteria and over-sized amounts for the assignment or project, and either one can cause academic stress. This condition is often contra-productive as students will feel burdened and frightened when they are not yet able to complete their assignments.

Based on the foregoing discussion, the framework of the study can be proposed. It is presented in Figure 1.

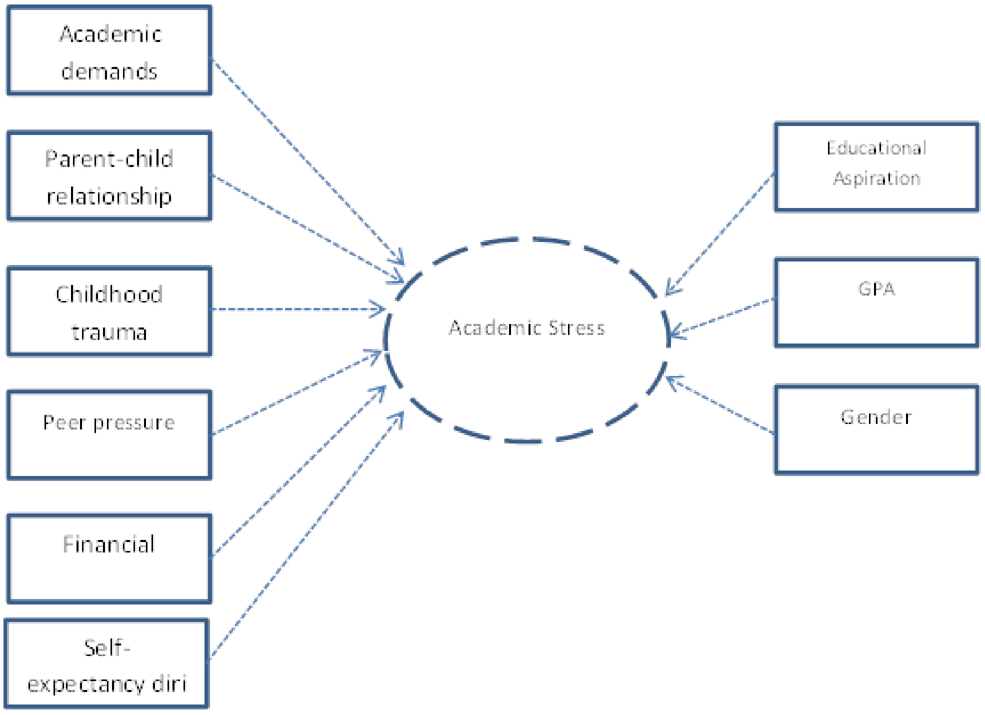


Figure 1. Research Framework

Method

The study was a correlational survey aimed at revealing the factors that lead to academic stress within students of the English Education Department at Yogyakarta State University (EED YSU). The research design was of the quantitative method based on survey data. The research subjects were students of EED YSU of the fourth semester and above. The choice of these students was based on the assumption that the students had taken all the knowledge and skill classes and some of the content classes. The total number of research participants amounted to 135 students coming from semester 4, 6, 8, 10, and 12. They completed the survey instrument online during April and May 2019 using computer sets, cellphones, or other gadgets.

Data collection used the close-type survey technique adapted from Calaguas (2012). The instrument was designed to obtain in-depth information on students' perceptions of the factors causing academic stress during their study in EED YSU. Raw data were subjected to an SPSS software program for statistical analyses. Data were subsequently analyzed using the Confirmatory Factor Analysis (CFA) technique. An ANOVA procedure was used to compare differences in academic stresses in view of the demographic variables of Educational Aspiration, GPA, and Gender. The study was conducted in the vicinity of EED YSU.

Findings and Discussion

Demographic information supplied by the research participants consists of Gender, Educational Aspiration, and GPA. Frequencies and percentages are presented in Table 1.

Table 1 shows that most of the respondents were students of the English Language Education Study Program. The data shows that most of the students of the English Language Education Study Program were female (77.8%). Almost half of the respondents (48.1%) show an educational aspiration of the S2 level (Master or Graduate). Over half of the participants (53.3%) have a current GPA of higher than 3.51.

Results of the CFA

The CFA on MPlus 7.2 software program is to determine the fit of the four criteria of Comparative Fit Index (CFI), Tucker-Lewis index (TLI), Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), and the *chi-square* (Bentler, 1990; Hu & Bentler, 1999; Tabachnick, Fidell, & Osterlind, 2007). A CFI < 0.90 indicates that there is not enough fit, while a CFI ≥ 0.90 indicates there is enough or almost perfect fit for the model (Bentler, 1990; Hu & Bentler, 1995; Hu & Bentler, 1999; Wang & Wang, 2012). The Tucker-Lewis index (TLI) is to know whether the model is less or more than estimated. The same value is used, i.e., a CFI ≥ 0.90 indicates that the model has enough or almost perfect fit (Bentler, 1990; Hu & Bentler, 1995; Hu & Bentler, 1999; Wang & Wang, 2012). Meanwhile, the RMSEA and SRMR is also used to identify whether the model is fit as viewed from the initial EFA model. An RMSEA and SRMR index of <.05 is regarded as evidence that the model is enough or almost perfect (Bentler, 1990; Hu & Bentler, 1995; Hu & Bentler, 1999; Wang & Wang, 2012).

Table 1. Research Respondents

		Frequency	Percentage
Gender	Female	105	77.8
	Male	30	22.2
Educational Aspiration	Undergraduate	19	14.1
	Graduate	65	48.1
	Post graduate	51	37.8
GPA	2.51-3.00	2	1.5
	3.01-3.50	60	44.4
	3.51-4.00	72	53.3

Table 2. Results of the CFA on Mplus 7.2

Item	Factor					
	1	2	3	4	5	6
V 32	0.469					
V 33	0.688					
V 34	0.55					
V 35	0.524					
V 36		0.784				
V 37		0.883				
V 39		0.732				
V 40			0.804			
V 41			0.614			
V 42			0.751			
V 43			0.664			
V 45				0.8		
V 47				0.873		
V 48				0.753		
V 49					0.468	
V 51					0.63	
V 52					0.829	
V 53						0.767
V 54						0.929
V 55						0.753
Alpha	0.751	0.732	0.744	0.850	0.787	0.720

Notes: Factor 1: academic demands; Factor 2: parent-child relationship; Factor 3: childhood trauma; Factor 4: peer pressure; Factor 5: financial; and Factor 6: self-expectancy

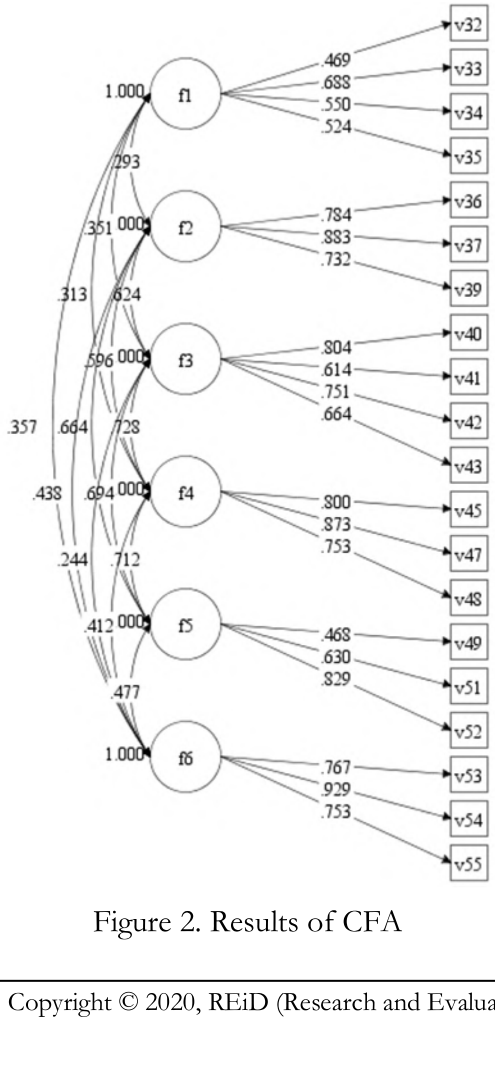


Figure 2. Results of CFA

The results of the CFA by the Mplus 7.2 procedure show that the research data are congruent with the model that is assumed to be based on the theories related to academic stress (Suldo et al., 2009). There are six factors found to be causes of academic stress. The 6-factor/dimension model representing the academic-stress sources is regarded as a good model ($x2$ = 219.128, *df* = 155, RMSEA = .055, SRMR= .067, CFI = .932, and TLI = .916). Factor loadings for each of the survey items are presented in Table 2.

Table 2 shows that all the 20 survey items have good loadings, ranging from 0.468 to 0.929. The reliability measurement on the SPSS 22 application also shows good scores, ranging from 0.720 to 0.850.

These analysis results show that the six factors represent the academic-stress sources factors, as suggested in the earlier study by Suldo et al. (2009). The six factors are elaborated as follows, and the details of these six factors are shown in Figure 2.

Factor 1: Academic Demands

To fulfill academic assignments, students face a number of difficulties. Some of these are related to finding materials and references for classes, finishing out final assignments or projects, preparing materials and media for class presentation, and studying for mid- and final examinations.

Factor 2: Parent-child Relationship

This factor can come in some forms. Among others, some parents or guardians do not give full support to their children's studies, those who lay excessive academic expectations of their children, and those who are not quite open in terms of family relationships.

Factor 3: Childhood Trauma

Some respondents express their traumatic experiences during their childhood. Most of these are related to being bullied by their peer, love affairs, accidents, and guilty feelings of their wrong-doings.

Factor 4: Peer Pressure

Problems with peer pressure can come in identical forms. Some of these are related to personal matters in a peer relationship, quarrel, bullying, being isolated by their close friends, and being ignored.

Factor 5: Financial

Problems in this matter are mostly related to financial management. The common problems are carelessness in spending, unexpected expenses, zero balance in the bank account, and debts to friends or neighbors.

Factor 6: Self Expectancy

This factor may not be seen on the surface; however, it gives most of the heaviest burdens to students. This problem is mainly derived from students' own self-expectation to be best in front of their parents, relatives, and neighbors.

The six factors of students' academic-stress sources analyzed in the study are found to surface as have been expected. However, it is undeniable that there are still many other factors that are possible. The CFA is followed up by analyses of each factor using recorded scores for testing of differences.

Results of *T-test* and *ANOVA*

Testing of differences is done to know whether there are differences in stress sources from the research variables: Gender, GPA, and Educational Aspiration. The GPA is divided into three categories: low (2.51-3.000, medium (3.01-3.50), and high (3.51-4.00). Educational aspiration is divided into three: undergraduate (S1), graduate (S2), and doctorate (S3). The results of the tests of the differences are presented in Table 3, Table 5, and Table 7.

Differences in Gender

Table 3 presents the results of the t-analysis for the equality of means of the six factors against Gender. The results of the t-analyses in Table 3 show significant differences among the five of the six factors of academic stress. Meanwhile, Table 4 shows that female students tend to report more problems than male students in the parent-child relationnship, childhood traumatic experiences, peer pressures, financial problems, and self-expectancy.

Table 3. Results of the *t*-test on Gender

		t-test for Equality of Means		
		t	**df**	**Sig. (2-tailed)**
Factor 1	Equal variances assumed	1.892	133	.061
	Equal variances not assumed	1.608	38.612	.116
Factor 2	Equal variances assumed	2.790	133	.006
	Equal variances not assumed	3.070	54.637	.003
Factor 3	Equal variances assumed	2.274	133	.025
	Equal variances not assumed	2.521	55.370	.015
Factor 4	Equal variances assumed	3.275	133	.001
	Equal variances not assumed	3.692	57.088	.000
Factor 5	Equal variances assumed	2.294	133	.023
	Equal variances not assumed	2.421	50.872	.019
Factor 6	Equal variances assumed	2.671	133	.009
	Equal variances not assumed	2.677	47.020	.010

Table 4. Descriptive Statistics of Each Factor

	Gender	**N**	**Mean**
Factor 1	Female	105	.0262
	Male	30	-.0916
Factor 2	Female	105	.1156
	Male	30	-.4047
Factor 3	Female	105	.0863
	Male	30	-.3021
Factor 4	Female	105	.1096
	Male	30	-.3835
Factor 5	Female	105	.0421
	Male	30	-.1472
Factor 6	Female	105	.0804
	Male	30	-.2813

Table 5. Results of the *F*-test on GPA

ANOVA

		Sum of Square	**df**	**Mean Square**	**F**	**Sig.**
Factor 1	Between Groups	1.409	2	.704	8.446	.000
	Within Groups	10.923	131	.083		
	Total	12.332	133			
Factor 2	Between Groups	1.881	2	.941	1.097	.337
	Within Groups	112.364	131	.858		
	Total	114.245	133			
Factor 3	Between Groups	5.988	2	2.994	4.471	.013
	Within Groups	87.722	131	.670		
	Total	93.710	133			
Factor 4	Between Groups	2.445	2	1.223	2.183	.117
	Within Groups	73.386	131	.560		
	Total	75.832	133			
Factor 5	Between Groups	.181	2	.090	.544	.582
	Within Groups	21.763	131	.166		
	Total	21.944	133			
Factor 6	Between Groups	.629	2	.315	.698	.499
	Within Groups	59.017	131	.451		
	Total	59.646	133			

Results of the F. test on GPA

Table 5 presents the F test results for the six research variables of the students' academic-stress factors against GPA. The results of the ANOVA as shown in Table 5 indicate that there are significant differences that are found in two of the factors, if viewed from the GPA, namely, the academic demands and childhood trauma.

Table 6 shows that students with a GPA of 3.51 and above tend to have higher academic demands and experience more severe childhood traumatic experiences than those with a GPA of 3.01- 3.50. There is no significant difference in the two factors between students with a GPA of 3.51-4.00 and those with a GPA of 3.01-3.50.

Results of the F. test on Educational Aspiration

Table 7 shows the results of the mean differences test among the six variable factors against educational aspirations. ANOVA test results in Table 7 show significant differences are found in two of the six variables: academic demands and parent-child relationship.

Table 8 shows that students who pursue an S3 education level report to have higher academic demands and heavier problems in parent-child relationship than those pursuing for an S2 level. No significant difference is found in these two-factor variables between students with an S1 level than either S2 or S3.

From the data analysis results, especially the CFA and tests of mean differences, the sources of students' academic stresses consist of six factors: academic demands, parent-child relationships, traumatic on childhood experiences, peer pressure, and self-expectancy. There are significant differences in some factors between male and female students, between students with high GPA and medium GPA, and between students who have an educational aspiration of the S2 and S3 levels.

Table 6. Results of the Post-hoc Analyses on Differences against GPA

Dependent Variable	(I) GPA	(J) GPA	Mean Difference (I-J)	Std. Error	Sig.
Factor 1	1.00	2.00	-.05878	.20756	.957
		3.00	.14825	.20700	.754
	2.00	1.00	.05878	.20756	.957
		3.00	.20703*	.05048	.000
	3.00	1.00	-.14825	.20700	.754
		2.00	-.20703*	.05048	.000
Factor 3	1.00	2.00	.92323	.58820	.262
		3.00	1.25432	.58662	.086
	2.00	1.00	-.92323	.58820	.262
		3.00	.33109	.14304	.047
	3.00	1.00	-1.25432	.58662	.086
		2.00	-.33109	.14304	.047

Table 7. Results of the *ANOVA* against Educational Aspiration

ANOVA

		Sum of Square	df	Mean Square	F	Sig.
Factor 1	Between Groups	.984	2	.492	5.706	.004
	Within Groups	11.383	132	.086		
	Total	12.367	134			
Factor 2	Between Groups	5.728	2	2.864	3.483	.034
	Within Groups	108.552	132	.822		
	Total	114.280	134			
Factor 3	Between Groups	1.768	2	.884	1.265	.286
	Within Groups	92.282	132	.699		
	Total	94.050	134			
Factor 4	Between Groups	1.692	2	.846	1.502	.226
	Within Groups	74.315	132	.563		
	Total	76.007	134			
Factor 5	Between Groups	.305	2	.152	.929	.398
	Within Groups	21.651	132	.164		
	Total	21.956	134			
Factor 6	Between Groups	1.850	2	.925	2.101	.126
	Within Groups	58.132	132	.440		
	Total	59.982	134			

Table 8. Results of the Post-hoc Analyses against Educational Aspiration

Dependent Variable	Educational Aspiration		Mean Difference (I-J)	Std. Error	Sig.
Factor 1	1.00	2.00	.04737	.07658	.810
		3.00	.20977*	.07893	.024
	2.00	1.00	-.04737	.07658	.810
		3.00	.16241*	.05493	.010
	3.00	1.00	-.20977*	.07893	.024
		2.00	-.16241*	.05493	.010
Factor 2	1.00	2.00	-.38825	.23650	.232
		3.00	.03225	.24374	.990
	2.00	1.00	.38825	.23650	.232
		3.00	.42050*	.16964	.038
	3.00	1.00	-.03225	.24374	.990
		2.00	-.42050*	.16964	.038

Results of the study have an agreement with the results of previous studies, which show a high level of academic stress experienced by S1-level students (McPherson, 2009; Oon, 2007; Suldo et al., 2009). These studies find that students have high academic stress due to their over-expectation despite their low levels of skills and abilities, for not being too enthusiastic with their academic activities and assignments because of their lack of interest in the subject contents, and because of health problems. If such academic stresses are not given adequate anticipation and solution, they can cause adverse effects on students since academic stresses may induce fatal impacts. One such impact is the high level of student drop-outs. This is in line with the results of previous studies by Rayle and Chung (2007) and Zajacova, Lynch, and Espenshade (2005). This is not intended to say that academic stresses cause drop-outs; this, however, shows that academic stresses make students have low levels of concentration that lead to a decline in their academic achievement. This decline in academic achievement eventually becomes the cause for students to repeat classes, but unable to complete, and, eventually, fail academic requirements and drop out.

Another problem that triggers students' academic stress is their inability to complete the minimal competencies set up in the curriculum. This problem is often doubled by the heavy assignments and expectations imposed by the curriculum either because it is not quite realistic or because it is not quite well-developed. In the same manner, the over-expectation of self and family becomes an acute source of academic stress. For most students, the achievement of a high GPA is a dead target, outweighing any other thing. As a result, students will do anything, including loosening their muscles and brain when they are actually tight, in order to obtain high GPAs bringing about burn-outs and academic stress. This is in agreement with the results of a previous study by McPherson (2009).

Other factors causing students' academic stress are parent-child relationships, traumatic childhood experiences, peer pressure, financial matters, and self-expectancy. This is in agreement with the results of previous studies such as one by Suldo et al. (2009). From observation in the field, it is true that these factors exert a heavy influence on the psychology of people, including that of students.

Another phenomenon to point out is that which suggests that female students, students with high GPAs, and students with high educational aspiration (such as of the S3 level), are found to have high academic demands, parent-child relationship, traumatic childhood experiences, peer pressure, financial problems, and self-expectancy. This is identical with the findings of other previous studies such as by Misra and McKean (2000) and Zajacova et al. (2005).

Conclusion

The study departed from the objective identifying and describing sources of academic stresses of students of the English Education Department, Yogyakarta State University

(EED YSU). A total of 135 English Education Department students participated in the study. Results of the CFA analyses found six factors that were identified as sources of students' academic stress. The six factors were academic demands, parent-child relationships, childhood traumatic experiences, peer pressure, financial matters, and self-expectancy. Further analyses showed that female students, students with high GPAs, and students with high academic aspirations reported higher measures of factors of academic stresses.

References

American College Health Association - National College Health Assessment. (2016). *Data report of undergraduate student reference group*. Retrieved from https://www.acha.org/documents/ncha/NCHA-II SPRING 2016 UNDERGRADUATE REFERENCE GROUP DATA REPORT.pdf

Ang, R. P., & Huan, V. S. (2006). Relationship between academic stress and suicidal ideation: Testing for depression as a mediator using multiple regression. *Child Psychiatry and Human Development*, *37*(2), 133–143. https://doi.org/10.1007/s10578-006-0023-8

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Butt, J., Weinberg, R., & Horn, T. (2003). The intensity and directional interpretation of anxiety: Fluctuations throughout competition and relationship to performance. *The Sport Psychologist*, *17*(1), 35–54. https://doi.org/10.1123/tsp.17.1.35

Calaguas, G. M. (2012). Survey of college academic stressors: Development of a new measure. *Journal of Human Sciences*, *9*(1), 441–457. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/1811

DeFrank, R. S., & Ivancevich, J. M. (1998). Stress on the job: An executive update. *The Academy of Management Executive*, *12*(3), 55–66. Retrieved from https://www.jstor.org/stable/4165477

Folkman, S. (2013). Stress: Appraisal and coping. In *Encyclopedia of Behavioral Medicine* (pp. 1913–1915). https://doi.org/10.1007/978-1-4419-1005-9_215

Gabre, H., & Kumar, G. (2012). The effects of perceived stress and Facebook on accounting students' academic performance. *Accounting and Finance Research*, *1*(2), 87–100. https://doi.org/10.5430/afr.v1n2p87

Hashemi, M. (2011). Language stress and anxiety among the English language learners. *Procedia - Social and Behavioral Sciences*, *30*, 1811–1816. https://doi.org/10.1016/j.sbspro.2011.10.349

Henriques, G. (2014). The college student mental health crisis. Retrieved from Psychology Today website: https://www.psychologytoday.com/blog/theory-knowledge/201402/the-college-student-mental-health-crisis

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Hu, Li-tze, & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 77–99). Thousand Oaks, CA: Sage.

Hu, Li-tze, & 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

Leyva, E. M. R. (2003). The impact of the internet on the reading and information practices of a university student community: The case of UNAM. *New Review of Libraries and Lifelong Learning*, *4*(1), 137–157. https://doi.org/10.1080/1468994042000240287

McNeil, E. (2016). Study: Teacher stress reduction leads to instructional improvement. Retrieved from Education Week website: http://blogs.edweek.org/teachers/teaching_now/2016/05/lessstressforteachers.html

McPherson, S. S. (2009). *Stressed out in school?: Learning to deal with academic pressure*. Berkeley Heights, NJ: Enslow.

Mezzacappa, E. S., & Katkin, E. S. (2002). Breast-feeding is associated with reduced perceived stress and negative mood in mothers. *Health Psychology*, *21*(2), 187–193. https://doi.org/10.1037/0278-6133.21.2.187

Misra, R., & Castillo, L. G. (2004). Academic stress among college students: Comparison of American and international students. *International Journal of Stress Management*, *11*(2), 132–148. https://doi.org/10.1037/1072-5245.11.2.132

Misra, R., & McKean, M. (2000). College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American Journal of Health Studies*, *16*(1), 41–51.

Oon, A. N. (2007). *Seri teaching children: Handling study stress*. Jakarta: Elex Media Komputindo.

Rayle, A. D., & Chung, K.-Y. (2007). Revisiting first-year college students' mattering: Social support, academic stress, and the mattering experience. *Journal of College Student Retention: Research, Theory & Practice*, *9*(1), 21–37. https://doi.org/10.2190/X126-5606-4G36-8132

Robotham, D., & Julian, C. (2006). Stress and the higher education student: A critical review of the literature. *Journal of Further and Higher Education*, *30*(2), 107–117. https://doi.org/10.1080/03098770600617513

Sarafino, E. P. (2008). *Health psychology: Biopsychosocial interactions* (6th ed.). Hoboken, NJ: John Wiley & Sons.

Suldo, S. M., Shaunessy, E., Thalji, A., Michalowski, J., & Shaffer, E. (2009). Sources of stress for students in high school college preparatory and general education programs: Group differences and associations with adjustment. *Adolescence*, *44*(176), 925–948.

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. West Sussex: John Wiley & Sons.

Yeh, C. J., & Inose, M. (2003). International students' reported English fluency, social support satisfaction, and social connectedness as predictors of acculturative stress. *Counselling Psychology Quarterly*, *16*(1), 15–28. https://doi.org/10.1080/0951507031000114058

Zajacova, A., Lynch, S. M., & Espenshade, T. J. (2005). Self-efficacy, stress, and academic success in college. *Research in Higher Education*, *46*(6), 677–706. https://doi.org/10.1007/s11162-004-4139-z

REiD (Research and Evaluation in Education), 6(1), 2020, 78-86
Available online at: http://journal.uny.ac.id/index.php/reid

An analysis of the suitability of students' civic knowledge and disposition in the topic of citizen's rights and obligations

[*]Dwi Riyanti
Institute for Educational Development, Universitas Ahmad Dahlan
Jl. Ringroad Selatan, Kragilan, Tamanan, Banguntapan, Bantul, Yogyakarta 55191, Indonesia
[*]Corresponding Author. E-mail: dwiriyanti.ysu@gmail.com

Submitted: 10 May 2020 | Revised: 26 May 2020 | Accepted: 23 June 2020

Abstract

Civic Education has been taught in primary education, but it has not impacted significantly with no strength and function. It is proven by the number of youths who have not understood and implement the citizen right and obligation (*hak dan kewajiban warga negara*) topic in Civic Education for their daily life. Thus, civic knowledge and civic disposition have not run as expected. This matter has become a great task for Civic Education lecturers to maximize and correlate comprehensively civic knowledge and civic disposition. This study discussed the suitability of civic knowledge and civic disposition in the topic of "*Hak dan Kewajiban Warga Negara*". This study was descriptive research with a qualitative approach conducted in Universitas Negeri Yogyakarta (UNY), Universitas Ahmad Dahlan (UAD), SMKN 1 Kalikajar (Vocational High School), SMPN 1 Magelang (Junior High School), and SD Kembangsongo (Elementary School) and focused on Civic Education subject by using purposeful sampling. The data were collected through interviews and documentation. The data were then analyzed through the triangulation technique. This study resulted that students in their daily life have not fully implemented both civic knowledge and civic disposition. The matter was caused by students and lecturers of the Civic Education, whereas the subject's topic has met the criteria of the curriculum in the level of Elementary School, Junior High School, and Senior High School/Vocational High School although not all topics were taught in these levels.

Keywords: *civic knowledge, civic disposition, civic education*

How to cite: Riyanti, D. (2020). An analysis of the suitability of students' civic knowledge and disposition in the topic of citizen's rights and obligations. *REiD (Research and Evaluation in Education), 6*(1), 78-86. doi:https://doi.org/10.21831/reid.v6i1.31621.

Introduction

Sunarso, Sartono, Dwikusrahmadi, and Sutarini (2016, p. 6) explains the report of the Session of BPUPKI and PPKI stating that education in Indonesia must be able to prepare students to become citizens who have a strong commitment to maintain the Republic of Indonesia unity which has the essence of modern nationalism. It means that in a modern era, the formation of nation and state is based on a sense of nationalism of the people who have a strong determination to build a future with a variety of different populations.

A country that adheres to the concept of democracy will prove the development of the concept of civil society that has a concept of a position to gather the strength of society to maintain the freedom, diversity, and independence of the community against state and government power. Although there is independence, both are having a mutual relationship (Alam, 2014, p. 196). Ferguson, Hume, and Adam Smith began to identify the concept of civil society with a civilized society oriented to the material organization (Jb & Darmawan, 2016, p. 40). Therefore, to realize the civil society concept, a state with a good civic competence is required.

Civic Education has been taught since elementary school education in Indonesia. In addition, citizenship knowledge can be obtained from non-formal education by reading news both from print and electronic media. Therefore, citizenship knowledge is not only theoretically, but also practically looking directly on the current evidence in the scope of citizenship. Even though citizenship knowledge can be obtained through non-formal education, formal education still has a very significant share of students' knowledge of citizenship (Galston, 2007, p. 627).

Currently, Civic Education has not been able to have a significant impact; it is still not functioning properly and powerless, although, in the reformation era, Civic Education demands that it can revitalize itself so that it can carry out its vision and mission. Charles in Print, Ellickson-Brown, and Baginda (1999, pp. 133–135) believes that the contents of Civic Education can be arranged in three models, namely *formal curriculum* implemented in learning, *an informal curriculum* that can be implemented in extracurricular activities, and *hidden curriculum* such as ethical development that can be developed in daily actions. With these three models, it is expected that students can have citizenship knowledge and can internalize it in everyday life.

According to Branson (1999, pp. 8–25), there are three aspects of Civic Education, namely civic knowledge, civic skills, and civic disposition. One of the aspects of Civic Education is civic knowledge. This material/topic is a substance related to rights and obligations that citizens should know it. This must be owned by every individual because it can positively affect and a picture of democratic values in society. When specifically described, the material/topic on citizenship knowledge includes several things, namely knowledge in terms of structure and political system in government, national identity, free and impartial justice, the constitution used, and the values that live in society.

Civic disposition is a citizenship competency. This is a combination of civic knowledge and civic skills. Civic disposition is a component related to a citizen's character in the scope of democracy that can be measured through the level of citizen awareness. This includes how a citizen understands his rights and obligations by complying with applicable laws, thinking critically, expressing opinions, having good morals, being responsible, being a good listener, discipline, and upholding human dignity (Feriandi & Harmawati, 2018, p. 77). It is also similar to the purpose of national education to develop students' potential to be faithful and devoted the almighty, noble, knowledgeable, skilled, creative, independent, and responsible for a democratic country (Ernawati, Tsurayya, & Ghani, 2019, p. 21).

Civic disposition in the formal curriculum can play an important role in shaping the character of students. Moreover, this is supported by Law No. 12 of 2012 of Republic of Indonesia on Higher Education in Article 35, Paragraph 3 for which the higher education curriculum must contain compulsory subjects, one of which is Civic Education. Therefore, through this subject, students can get civic knowledge and civic skills so that they can form civic disposition. In addition to the formal curriculum of civic disposition, it can be formed through activities undertaken by students, such as UKM or *Unit Kegiatan Mahasiswa* (students' extracurricular activities) and state defense training activities for students that have been carried out by Universitas Ahmad Dahlan.

Frailon, Schulz, and Ainley argue that civic disposition research resulted from ICCS on Civic Education situation in five countries such as Indonesia, Hong Kong, Republic of Korea/South Korea, Taiwan, and Thailand, have produced civic knowledge in Indonesia and Thailand on VIII class students is lower as compared to other sample countries in Asia. In addition, there are still many traffic violators derived from students themselves (Ainley, Fraillon, & Schulz, 2012, p. 3).

This opinion is in line with Wardhani, who states that *Operasi Progo Polresta Jogja* has cited and reported 977 violators, and 327 are students. Therefore, there are some points for educators to improve students' civic knowledge in order to produce a good civic disposition (Wardhani, 2019).

According to Setiawan and Suardiman (2018, p. 12), the social attitude can be identified through positive and negative trust in a particular entity's feeling and attitude that has three categories, namely emotional, cognitive, and attitude as it is connected with the previous matter, so that students have not fulfilled themselves in these categories, and that the civic knowledge and civic disposition have not been reached optimally.

The fact indicates that there are still a number of tasks of the lecturers to improvise civic knowledge for students to gain and reach for good civic disposition. As long as the civic disposition has been well developed, the citizens will have good behavior to support their political participation, and that the political system will function proportionally in order to improvise dignity and public's interest (Sunarso et al., 2016, p. 15). Therefore, this study discussed and analyzed the suitability of civic knowledge and civic disposition in the topic of citizen's rights and obligations in the subject of Civic Education.

Method

This research was a descriptive study that used a qualitative approach. This research aimed to find and analyze the suitability of civic knowledge and civic disposition, especially in the topic of rights and obligations of citizens on students of Universitas Negeri Yogyakarta (UNY) and Universitas Ahmad Dahlan (UAD). This study collected various kinds of data and exploited a time effectively in the study's field (Creswell, 2016, p. 254). The procedures undertaken in this study were interviews and documentation. The interview was conducted since February 15 to April 17, 2020, with ten different informants as classified into three categories, namely (a) Civic Education expert and Civic Education lecturer, (b) Civic Education teacher both in Elementary/Junior High School/Vocational School, and (c) Five Students at UNY and UAD.

This qualitative study was started by using an assumption as well as interpretation that could form and influence the studied matter (Creswell, 2015, p. 59). In validating the data, the researcher used source triangulation that involved different information that could coherently build thematic justification (Creswell, 2016, p. 269). Miles and Huberman (1994, pp. 10–12) stated that the technique of data analysis used data reduction, data display, and conclusion.

This study used a purposeful sampling technique by two considerations in determining the subject of research. The two considerations were the decision to choose the subject of research and the sample's specific strategy (Creswell, 2015, p. 215). Therefore, the researcher chose the specific subject of research to obtain a reflection on a problem that was being investigated, namely the suitability of civic knowledge and civic disposition in the subject of Civic Education for students at UNY and UAD. Based on the criteria, the researcher involved some subjects as informants, namely the expert on Civic Education, lecturer on Civic Education, teacher on Civic Education, and students at UNY and UAD who were enrolled in Civic Education class.

The research was conducted at Universitas Negeri Yogyakarta (UNY), Universitas Ahmad Dahlan (UAD), SMPN 1 Magelang, SMK Negeri 1 Kalikajar, and also SD Negeri Kembangsongo. The reason that the researcher chose the Civic Education subject was to know how far the suitability between civic knowledge and civic disposition in the topic of rights and obligations as well as to identify its relationship to the curriculum of the Elementary School, Junior High School, and Senior High School/Vocational High School. Besides, the reason that the researcher conducted the study at UNY and UAD was to know how far the suitability between civic knowledge and civic disposition in the topic of citizen's rights and obligations in the private university, state university, and Islamic university. SMPN 1 Magelang, SMK Negeri 1 Kalikajar, and SD Negeri Kembangsongo were chosen because the teachers were the alumni of the Civic Education study program at UNY.

To get a clear description and information about the suitability of civic knowledge and civic disposition in tertiary institutions, especially at UNY and UAD as well as the compatibility between subject/course with

Civic Education curriculum in Elementary/ Junior High School/ Vocational High School, this study determined the researched subject by using a purposive sampling technique. Creswell (2015, p. 217) states that research that used purposeful sampling technique was to determine specific and qualified subjects who could provide an overview of the investigated problem. The followings are the characteristics of the subject based on their roles classification. (1) Two Civic Education experts and Civic Education lecturers were involved in determining the extent of conformity between civic knowledge and four civic disposition in citizens' rights and obligations. (2) Three Civic Education teachers both in elementary/middle school/vocational school were involved in studying the compatibility among subjects in college with the Civic Education curriculum in elementary/middle school/vocational school. (3) Five students at UNY and UAD, who had taken Civic Education subjects to find out the suitability between civic knowledge and civic disposition, were involved.

In collecting data, this study applied two kinds of techniques, namely interview and documentation. The performed interview was a structured interview for which the issues and questions were previously determined. The results of the interview were called primary data obtained from research subjects. Meanwhile, the documentation technique was to support and supplement primary data, namely, the interview. Documentation was taken from data and records related to the compatibility between civic knowledge and civic disposition in the Civic Education subject.

Findings and Discussion

Civic Knowledge

According to Cogan in Winarno (2013, p. 4), Civic Education is a subject/course that prepares young people to have an active role in the nation's life and state. Civic Education aims to prepare students to become active, critical, rational, and creative in addressing the issue of citizenship. Rosnawati, Kartowagiran, and Jailani (2015, p. 187) also state that critical thinking can ease someone to process and use the information to solve any problem. Besides, the goal of Civic Education is that the younger generation can actively participate as well as intelligently be responsible for social activities in terms of the nation and state (Winarno, 2013, p. 95).

In this study, the learning outcomes are conducted by focusing on one of the aspects of civic knowledge where students can know, understand, and internalize the material/topic in the Civic Education subject. The material/ topic involves the rights and obligations of citizens. Certainly speaking, this is a process of teaching and learning activity that students must achieve because civic knowledge is basically a matter related to rights and obligation that citizens should carry out (Budimansyah, 2010, p. 49).

From an interview on students at UNY and UAD, it was found that they have already known about the rights and obligations of citizens. They have also understood the elements in the rights and obligations as regulated and clearly stipulated in the 1945 Constitution of the Republic of Indonesia. The interview with ADW, a student at the Automotive Study Program of UNY on February 15, 2020, has indicated that he could explain the rights and obligations in detail. In addition, other students could also explain the Articles in the 1945 Constitution of the Republic of Indonesia that regulated Indonesians' rights and obligations.

> *Citizens are those who live in a certain area and people in relation to the state. In their relation to the state, citizens have obligations to the state, and that the citizens also have rights that must be granted and protected by the state. Citizens' rights are everything that citizens must obtain from the state (government). Obligations are all things that must be carried out by citizens of the state. Rights and obligations of citizens are according to the 1945 Constitution on Citizens' Rights in the Article 27 (1,2,3) the Article 28 (A, B, C, D, E, F, G, H, I, J), the Article 29 (2) on the freedom of religion, the Article 30 on Defence and National Security, the Article 31 on Obtaining Education, and Rights and Obligations of Citizens in the Article 27 (1) on establishing the same citizens' rights in law and*

> *government, and the obligation to uphold the law and government. It is also stipulated in Article 27 (2) on establishing the right of citizens to work and a decent living for humanity, and finally, in Article 27 (3) on establishing the rights and obligations of citizens to participate in efforts to defend the state.*

The result of the interview indicated that the student understood the mutual relationship between citizens and the government. The mutual relationship means that both citizen and the government have their own rights and obligations. Thus, the rights and obligations as in the 1945 Constitution of the Republic of Indonesia regulate not only citizens' obligations and rights but also the rights and obligations of the state (government). In the same context, a similar opinion was also expressed by FAR (on February 16, 2020) as one of the students at Universitas Ahmad Dahlan, that between citizen and the state, there is a mutual relationship. He added that a right is something a citizen has a free-access to perform an action or speak a statement (opinion/argumentation) because he has done his duty as a citizen, whereas the oblige-tion of citizens is everything that must and a citizen himself do compulsory.

Both interviews have indicated that students from both UNY and UAD have understood the meaning of citizens and the state's rights and obligation. Therefore, civic knowledge competency is achieved. Similarly, civic knowledge has a relationship to what citizens must know. The content of civic knowledge is also related to the compulsory knowledge for which the citizens should recognize and comprehend (Budimansyah, 2010, p. 29).

According to Feriandi and Hermawati, civic knowledge is not only seen from the cognitive aspect, but also from other aspects such as social services and discussions in lectures on the issues of citizenship. Students are also done in doing social services in orphanages and in communities with low economic level. In addition, students have also been accustomed to be told in a class about current topics in the community. Thus, students can explore the civic knowledge they have gained in the subject they have learned (Feriandi & Harmawati, 2018, p. 78).

Civic Disposition

Civic disposition is a very basic and essential competency. Civic disposition is considered as the spearhead of the development of civic knowledge and civic skills. Quigley, Buchanan Jr., and Bahmueller (1991, p. 11) explain civic disposition as "...those attitudes and habits of mind of the citizen that are conducive to the healthy functioning and common good of the democratic system". It means that citizens' attitudes and habits are conducive to healthy functioning and the same virtue in a democratic order. Similar to this opinion, Branson (1999, p. 23) states that both public and private characters are important in developing a constitutional democratic system.

Branson (1999, pp. 23–25) strengthens that public and private characters can be described as follows. (1) Becoming an independent member of the community. (2) Being able to fulfill the responsibilities of being a citizen in the economic and political fields. (3) Being able to respect the dignity of all individuals regardless of social status and so on. (4) Being able to actively participate in the affairs of citizenship effectively, responsibly, and also wisely. (5) Being able to develop the function of democracy in a healthy way.

As previously explained, it can be seen that the character (private and public character) of citizenship is very important in the survival of the nation and state. In this case, the researcher examines civic disposition's competence in terms of citizens' rights and obligations. This resulted in both students at Universitas Negeri Yogyakarta and Universitas Ahmad Dahlan having implemented civic disposition in their daily lives, although it is not optimally performed.

> *As a citizen, I have the right to get appropriate education services as my choice, and I can get it. As a citizen, I am obliged to help maintaining harmony, and that I perform by always adapting with tolerance against differences, for example being tolerant in the campus environment; remembering that the campus consists of a variety of different religion, culture, ethnicity, and so on. (An interview with SM on February 16, 2020).*

The aforementioned statement indicated that the civic disposition of citizens' rights and obligations has appeared in student's personality by not only claiming rights but also carrying out his obligations as citizens through maintaining tolerance in the campus environment with a diversity of religion, culture, and ethnicity. Thus, any effort to establish students' civic disposition on the material of "citizen's rights and obligations" has been achieved, although not fully performed. The similar result can be also identified through an interview on April 6, 2020 with DA as a student at UNY who stated that taking education in tertiary institution is an implementation that every citizen has the right to receive education, to carry out worship according to religion, to choose their respective religion, and the fulfilment of food and clothing as a form of citizens' right in a decent living, obeying the tax system as a fulfilment of tax and legal obligation, following the Pancasila and Civic Education course as an obligation to defend the country, and avoiding SARA or *Suku, Agama, Ras,* and *Antargolongan* (Ethnic, Religion, Race, and Multi-groups) as an implementation of the obligation to respect the rights of others.

The interview also proves another similar result on April 6, 2020, with DCN as a student at UAD, who explains that students have indeed implemented their rights as citizens through participating in public's opinion, such as participating in demonstration activities. Therefore, it can be concluded that not all students have embedded in civic disposition, because there are still the rest of the students who have not implemented their rights and obligations properly.

The Suitability of Civic Knowledge and Civic Disposition in the Material of "Rights and Obligations"

Somantri (2001, p. 116) states that Civic Education is an effort done scientifically and psychologically in providing easy access for students to learn so that the moral internalization of Pancasila and knowledge of citizenship to realize personal integrity and everyday behavior are based on the national education goals. Thus, Civic Education in higher education is expected to prepare students as young people to participate in the nation's life and state. Students, as the young generation, are given an understanding of national ideals and how to act to overcome problems through Civic Education. Therefore, students can withdraw any decision that is responsible for overcoming private and national problems.

The 1945 Constitution is a foundation of formal values, norms, and moral education in Indonesia and implemented through Civic Education. It is also poured into Law No. 12 of 2012 of Republic of Indonesia on Higher Education in Article 35, which states that universities are required to teach subjects in Religion, *Pancasila* (five pillars of the nation), Citizenship, and Indonesian both at undergraduate and diploma level. Civic Education in nomenclature is always undergoing a transformation. Initially, it is previously a Civic course, and transformed to be Civic Education course, although there are still a number of topics that are typically citizenship, such as the concept of *Pancagatra* and *Trigatra*.

In this context, S as an expert in Civic Education at UNY stated that after the reformation, the transformation of the nomenclature has changed from Gallantry (manliness) to Civic Education. Although it has undergone a change, there are still typical topics of dignity such as *Pancagatra* and *Trigatra*, which indicate that the atmosphere still has an atmosphere and nuance of Gallantry (manliness). Furthermore, S said:

> *Although the material of democracy, human rights, and knowledge on state institutions are included in the Civic Education subject, the metamorphosis of that authority tends to be considered to represent military typology in the current state defence framework. Civic knowledge aspects are related to democracy, human rights, local government in tertiary institutions especially in state universities issued by the General Director of Higher Education in 2004. Because the Civics Education subject in tertiary institutions is still new, it was previously becoming a Gallantry (manliness). The figure is Kunto Wibisono; the development team for the transformation of Civic Education from Gallantry (manliness) in Higher Education. For me, there are actually many things that must be clarified from the epis-*

> *temology both from the scientific framework and from the scope of activities. On the other hand, building the character of citizenship in the University's Civic Education subject on post-citizenship, we can also see the difference after it was regulated in 2012; when there was a circular from the General Director of Higher Education mentioned that one of the concepts of Law No. 12 of 2012 on Tertiary Education required tertiary institutions to teach a minimum of four compulsory subjects in tertiary and diploma colleges. Those subjects were religion, Indonesian, Pancasila, and Civic Education. (An interview with S on March 9, 2020)*

The statement indicates that Civic Education still tends to represent the military typology in the concept of state defense and its civic knowledge aspects related to democracy, human rights, and government. There must also be clarity from the epistemological concept. Another perspective is also given by C as an expert of Civic Education at UNY on March 9, 2020. He states that to make civic disposition matched to civic knowledge, a theory is not the first priority, but rather to the understanding of citizens concerned that each citizen has to associate for the sake of living in a democratic country.

In addition, C emphasized that Civic knowledge and civic disposition should be able to develop citizens' intelligence, so that students as young people can stick the values of virtue with a good character. The attitude can be controlled through intelligence, although there are two ways to control attitudes, namely habit (obtained since Early Childhood Education to junior high school) and the level of intelligence understanding of the importance of attitude that compulsory to be realized by high school and college students. He also argued that that civic disposition presently is only introduced to concepts and theories, so it will not be formed at any time. Thus, civic knowledge should gain a high level of understanding than just a theory.

In the view of the material aspects of citizens' rights and obligations, the suitability between civic knowledge and civic disposition is still not fully internalized, although many students actively participate in student's organizations on the campus. This evidence has addressed that they have internalized their rights as citizens of an association (community). In the same context, the interview on March 1, 2020, with HH, as one of the lecturers of Civic Education at UAD stated that the material rights and obligations are appropriate, although students are still not aware of the importance of the material so that any experience in the field is needed.

Civic Education has not been very effective in shaping citizens' character because it still needs to be strengthened with other supplements outside of Civic Education. In line with it, S, another informant being interviewed on March 9, 2020, also stated that there was a need to reorganize the substance of Civic Education subject in higher education with current issues in defending the country, deradicalization, and efforts to minimize intolerance. One of the values built on Civic Education was how to live together and have responsibility for the nation and the state with dynamic challenges.

Further, S argued that state defense material also had an impression of denying the existing Civics Education models. Defending the state is similar to defending the nationality model by Gallantry (manliness). Any physical activity is how to differentiate it. In this context, the topic of state defense in the National Defense Institute is like revitalizing the spirit of dignity that once exists in the Civic Education subject. It is similar to the view that Civic Education merely reaches the cognitive domain, and there are some studies found in state defense materials in the candidates for civil servants through education and training.

Further, civic knowledge and disposition in the material of rights and obligations have not been fully successful, although it is caused by individual factor. The interview on April 17, 2020, with SYT, a lecturer of Civic Education at UNY, found that civic knowledge and disposition on citizen's rights and obligations were caused by individual factors. It is also reinforced by the opinion of FFH, a lecturer of Civic Education at UNY who stated that the students' factors have become a cause of the entirely appropriateness of civic knowledge and disposition (Resulted from an interview on March 4, 2020).

Apart from the students' factors', C additionally pointed out that it was influenced by factors of lecturers who did not necessarily have a Civic Education background; and that the lecturers from the graduate program of Civic Education were also similarly considered not to have the same perspective in terms of attitude (Resulted from an interview on March 9, 2020). The attitude perspective has varied, and there should be a powerful test for it. Besides, the course was not conducted intensively. Another weakness was that it was rare for someone to pursue a field of expertise during the course (Resulted from an interview on March 9, 2020). Thus, it is the time for lecturers to be engaged in a truly ingrained field so that they can get more actual and relevant views to be effectively substantive and productive in the development of science.

The compatibility of Civic Education subject in tertiary institutions and elementary/ junior high school/senior high school/vocational high school must also be assessed because the assessment needs to be conducted to verify and validate teachers' competency. On the other hand, TM, a teacher at State Junior High School 1 Magelang and the alumnus of Civic Education at UNY, stated that most were appropriate and in class VII put more emphasis on constitutionality (Resulted from an interview on March 5, 2020). In line with this, WW as a teacher at Kembangsongo Elementary School and alumnus of Civic Education at UNY, also argued that there was conformity even though the curriculum for elementary school still applied for K13 curriculum were elaborated in the form of theme, sub-theme, and basic competency (Resulted from an interview on April 6, 2020).

Based on the interviews, it is identified that most of the assumptions have met the criteria of analysis, although not all materials/ topics obtained from university's lecture have been taught in the elementary/junior high school/senior high school/vocational high school. EP positively confirms this argumentation as a teacher at State Vocational High School 1 Kalikajar that not all materials/ topics obtained from university's lecture have been already taught, and that the teachers have already delivered any material/topic by using a variety of methods, such as Discovery Learning, Project-based Learning, and Problem-based Learning (Resulted from an interview on March 3, 2020).

Conclusion

The suitability between civic knowledge and disposition in the subject of Civic Education in the topic of "citizen's rights and obligations" has not been fully implemented as expected. This is proven by the fact that the students have understood the mutual relationship between citizens and government, but they have not fully implemented their comprehension into their daily life. A field study like democratic learning is necessary in order that students do not only understand but also implement it.

Apart from the students and lecturers factor, the lecturers of the subject Civic Education do not certainly have the background knowledge on civic education and those who come from the Civic Education study program with a few of attitude and perspective enrichment. This evidence can make civic knowledge, and civic disposition have not been optimally conducted and matched. It means that the Civic Education has not been effective in forming citizenship character for youths, like the supplementary program the candidates for civil servants that are also expected to form the character of the citizen. A study on the suitability of Civic Education in higher education with the curriculum on elementary school, junior and senior high school, and vocational high school has indicated the existence of suitability although not all topics discussed in higher education are previously taught in elementary school, junior and senior high school, and vocational high school.

References

Ainley, J., Fraillon, J., & Schulz, W. (2012). *ICCS 2009 Asian report: Civic knowledge, attitudes, and engagement among lower-secondary students in five Asian countries*. Retrieved from https://research.acer.edu.au/civics/17

Alam, B. (2014). Antropologi dan civil society: Pendekatan teori kebudayaan.

Antropologi Indonesia, *30*(2), 193–200. https://doi.org/10.7454/ai.v30i2.3564

Branson, M. S. (Ed.), Syarifudin, S. (Trans.). (1999). *Belajar civic education dari Amerika*. Yogyakarta: Lembaga Kajian Islam dan Sosial (LKIS).

Budimansyah, D. (2010). *Pembelajaran pendidikan kesadaran kewarganegaraan multidimensional*. Bandung: Genesindo.

Creswell, J. W. (2015). *Penelitian kualitatif & desain riset: Memilih di antara lima pendekatan* (A. L. Lazuardi, Trans.). London: SAGE Publications.

Creswell, J. W. (2016). *Research design: Pendekatan metode kualitatif, kuantitatif, dan campuran* (A. Fawaid & R. K. Pancasari, Trans.). London: SAGE Publications.

Ernawati, E., Tsurayya, H., & Ghani, A. R. A. (2019). Multiple intelligence assessment in teaching English for young learners. *REiD (Research and Evaluation in Education)*, *5*(1), 21–29. https://doi.org/10.21831/reid.v5i1.23376

Feriandi, Y. A., & Harmawati, Y. (2018). Analisis penguasaan kompetensi kewarganegaraan pada mahasiswa PPKn Universitas PGRI Madiun. *Jurnal Citizenship: Media Publikasi Pendidikan Pancasila Dan Kewarganegaraan*, *1*(2), 76–83. https://doi.org/10.12928/citizenship.v1i2.13620

Galston, W. A. (2007). Civic knowledge, civic education, and civic engagement: A summary of recent research. *International Journal of Public Administration*, *30*(6–7), 623–642. https://doi.org/10.1080/01900690701215888

Jb, M. C., & Darmawan, L. (2016). Wacana civil society (masyarakat madani) di Indonesia. *Jurnal Sosiologi Reflektif*, *10*(2), 35–64. https://doi.org/10.14421/jsr.v10i2.1157

Law No. 12 of 2012 of Republic of Indonesia on Higher Education. , (2012).

Miles, M. B., & Huberman, M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Print, M., Ellickson-Brown, J., & Baginda, A. R. (Eds.). (1999). *Civic education for civil society*. London: ASEAN [Association of South East Asian Nations] Academic Press.

Quigley, C. N., Buchanan Jr., J. H., & Bahmueller, C. F. (1991). *CIVITAS: A framework for civic education*. Calabasas, CA: Center for Civic Education.

Rosnawati, R., Kartowagiran, B., & Jailani, J. (2015). A formative assessment model of critical thinking in mathematics learning in junior high school. *REiD (Research and Evaluation in Education)*, *1*(2), 186–198. https://doi.org/10.21831/reid.v1i2.6472

Setiawan, A., & Suardiman, S. P. (2018). Assessment of the social attitude of primary school students. *REiD (Research and Evaluation in Education)*, *4*(1), 12–21. https://doi.org/10.21831/reid.v4i1.19284

Somantri, M. N. (2001). *Menggagas pembaharuan pendidikan IPS*. Bandung: Remaja Rosdakarya.

Sunarso, S., Sartono, K. E., Dwikusrahmadi, S., & Sutarini, Y. C. N. (2016). *Pendidikan Kewarganegaraan: PKn untuk Perguruan Tinggi*. Yogyakarta: UNY Press.

Wardhani, C. M. (2019). Operasi Keselamatan Progo, Polresta Yogya tilang 977 pelanggar, 327 merupakan pelajar dan mahasiswa (A. Nugroho, Ed.). Retrieved from Tribun Jogja website: https://jogja.tribunnews.com/2019/05/13/operasi-keselamatan-progo-polresta-yogya-tilang-977-pelanggar-327-merupakan-pelajar-dan-mahasiswa

Winarno, W. (2013). *Pembelajaran pendidikan kewarganegaraan: Isi, strategi, dan penilaian*. Jakarta: Bumi Aksara.