



Jurnal

Penelitian dan Evaluasi Pendidikan



Jurnal

Penelitian dan Evaluasi Pendidikan

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

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



9 772338 606001



9 772685 711007

Jurnal Penelitian dan Evaluasi Pendidikan

Volume 23, No 1, June 2019

- Intan Febry Sulasini, Supriyono Koes Handayanto, Wartono* Development of self-rating scale instrument of self-directed learning skills for high school students
- M. Dwirifqi Kharisma Putra, Wardani Rabayu, Jahja Umar* Indonesian-language version of general self-efficacy scale-12 using Bayesian confirmatory factor analysis: A construct validity testing
- Dwi Agus Kurniawan, Astalini, Deti Kurnia Sari* An evaluation analysis of students' attitude towards physics learning at senior high school
- Aulia Ninda Haryoni, Istiana Hermawati* An educational-evaluation study for street children in Rumah Impian Foundation
- Ikhyia Ulumudin, Sisca Fujianita* The implementation of attitude assessment in Curriculum 2013 at elementary schools
- Herwin, Sophak Phonn* The application of the Generalized Lord's Chi-Square method in identifying biased items
- Nur Ichsanuddin A. Kurniawan, Sudji Munadi* Analysis of the quality of test instrument and students' accounting learning competencies at vocational school
- Dani Surya Lee* China's K-12 teacher qualification system
- Nursanti Dwi Yogawati, Widibastuti* Evaluating the implementation of English communication therapy (ECT): An objective structured clinical assessment (OSCA) approach
- Kusaeri, Ali Ridho* Learning outcome of mathematics and science: Features of Indonesian madrasah students



Jurnal

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

Publisher

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA
In Cooperation With
PROGRAM PASCASARJANA UNIVERSITAS NEGERI YOGYAKARTA
(MOU Nomor 195 B/J.35.17/LK/04)

Director of Publication

Djemari Mardapi, *Universitas Negeri Yogyakarta*

Editor in Chief

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

Associate Editors

Nur Hidayanto Pancoro Setyo Putro, *Faculty of Languages & Art, Universitas Negeri Yogyakarta*

Editors

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

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

Sudiyatno, *Faculty of Engineering, Universitas Negeri Yogyakarta*

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

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

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

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

Board of Reviewers

Suratno

Universitas Lambung Mangkurat

Moch Alip

Universita Negeri Yogyakarta

Budiono

Universitas Sebelas Maret Surakarta

Maria Josephine Wantah

Universitas Negeri Manado

Kartono

Universitas Negeri Semarang

Rasmuin

Universitas Dayanu Ikhsanuddin Bau-Bau

Purwo Susongko

Universitas Pancasakti Tegal

Nurul Choyimah

LAIN Tullungagung

Zurqoni

STAIN Samarinda

Amir Syamsudin

Universitas Negeri Yogyakarta

Wasis

Universitas Negeri Surabaya

Sugeng Eko Putro Widoyoko

Universitas Muhammadiyah Purworejo

Nanik Estidarsani

Universitas Negeri Surabaya

Ekohariadi

Universitas Negeri Surabaya

Mansyur

Universitas Negeri Makassar

Undang Rosidin

Universitas Lampung

Lilik Sabdaningtyas

Universitas Lampung

Primardiana Hermilia Wijayati

Universitas Negeri Malang

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

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

e-mail: jurnalhepi@uny.ac.id

FOREWORDS

We are very pleased that *Jurnal Penelitian dan Evaluasi Pendidikan* is releasing its issue **Volume 23, No 1, June 2019**. We are also very excited that the journal has been attracting papers from many institutions in Indonesia and many foreign countries. *Jurnal Penelitian dan Evaluasi Pendidikan* was first published in **1998** and since then regularly published online and in print twice a year: June and December.

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

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

Yogyakarta, June 2019

Editor in Chief

Table of Content

<i>Intan Febry Sulasini, Supriyono Koes Handayanto, Wartono</i>	Development of self-rating scale instrument of self-directed learning skills for high school students	1-11
<i>M. Dwirifqi Kharisma Putra, Wardani Rahayu, Jahja Umar</i>	Indonesian-language version of general self-efficacy scale-12 using Bayesian confirmatory factor analysis: A construct validity testing	12-25
<i>Dwi Agus Kurniawan, Astalini, Deti Kurnia Sari</i>	An evaluation analysis of students' attitude towards physics learning at senior high school	26-35
<i>Aulia Ninda Haryoni, Istiana Hermawati</i>	An educational-evaluation study for street children in Rumah Impian Foundation	36-45
<i>Ikhyia Ulumudin, Sisca Fujianita</i>	The implementation of attitude assessment in Curriculum 2013 at elementary schools	46-56
<i>Hernin, Sopbak Phonn</i>	The application of the Generalized Lord's Chi-Square method in identifying biased items	57-67
<i>Nur Ichsanuddin A. Kurniawan, Sudji Munadi</i>	Analysis of the quality of test instrument and students' accounting learning competencies at vocational school	68-75
<i>Dani Surya Lee</i>	China's K-12 teacher qualification system	76-86
<i>Nursanti Dwi Yogawati, Widihastuti</i>	Evaluating the implementation of English communication therapy (ECT): An objective structured clinical assessment (OSCA) approach	87-94
<i>Kusaeri, Ali Ridho</i>	Learning outcome of mathematics and science: Features of Indonesian madrasah students	95-105

DEVELOPMENT OF SELF-RATING SCALE INSTRUMENT OF SELF-DIRECTED LEARNING SKILLS FOR HIGH SCHOOL STUDENTS

Intan Febry Sulasivi

Graduate Program, Universitas Negeri Malang

Supriyono Koes Handayanto

Department of Physics FMIPA, Universitas Negeri Malang

Wartono

Department of Physics FMIPA, Universitas Negeri Malang

Abstract

This study aims to develop a valid and reliable self-rating scale instrument for measuring Self-Directed Learning (SDL) skills. This DDI study follows the steps of Hinkin's development (1995) which consists of five stages: creating an item pool, expert conclusion, implementation, confirmatory factor analysis, and reliability analysis. The self-rating scale developed in this study consisted of sixty statements accompanied by a 1-5 Likert scale. Based on the factors analysis, 16 items were still in the draft and 44 items were declared valid and reliable. Five factors that are determined are: awareness (8 items, $\alpha = 0.717$), learning strategies (9 items, $\alpha = 0.806$), and learning activities (7 items, $\alpha = 0.777$), evaluation (8 items, $\alpha = 0.790$), and interpersonal skills (12 items, $\alpha = 0.907$). The reliability coefficient (Cronbach Alpha) of the self-rating scale is $\alpha = 0.933$, with the required reliability criteria is 0.5. On a scale conversion of 1-100, the student's highest score of SDL skills is 93, and the lowest score SDL skills are 31 (SD = 20.334).

Keywords: *self-rating scale, self-directed learning skills, high school students*

Permalink/DOI: <http://dx.doi.org/10.21831/>

Contact *Supriyono Koes Handayanto*

 *supriyono.koesbandayanto.fmipa@um.ac.id*

 *Department of Physics FMIPA, State University of Malang*

*Jalan Semarang 5, Sumbersari, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145,
Indonesia*

Introduction

Skills for self-directed learning among students and their role in improving lifelong learning skills have been emphasized lately (Taqipour, Abbasi, Naeimi, Ganguly, & Miandashti, 2016). With the development of information and technology today, students need to be equipped with the skills to navigate unexpected challenges in the future such as those contained in self-directed learning skills (SDLS) (Scott, 2015). Self-directed learning (SDL) skills allow students to know what they are learning about their learning pathways and freely choose how they will learn (Acar, Kara, & Taşkin Ekici, 2016). Thus, the SDL skills are key factors that affect the ability of lifelong learning (Shen, Chen, & Hu, 2014). Additionally, SDLS is one factor that contributes to higher achievement in science (Kan'an & Osman, 2015).

Mastery of SDLS allows independently student-centered learning to occur more optimally. This is in accordance with the definition of SDLS offered by Knowles (1975), namely description of process in which individuals take the initiative, with or without the assistance of others, in diagnosing their learning needs, formulating learning goals, identifying human resources and learning materials, choosing and implementing appropriate strategies, and evaluating learning outcomes. Merriam & Caffarella (1999) (in Huang, 2008) added the definition of SDL as a process where a person can take the initiative to plan, implement, and evaluate his learning process as a "personal attribute" which includes initiative, ability, and willingness of students to control essential self-directed learning process. This shows that SDL skills have a shared commitment to changes in position and role of students where students hold more significant control over themselves in terms of conceptualization, design, implementation, and evaluation of learning and the application of ways to use learning resources for further learning.

Seeing the importance of SDL skills, SDLS research on physics is still limited. One of the studies that have been done is to

find the relationship between problem-based learning and SDLS to the problem of global warming and power plants (Malan, Ndlovu, & Engelbrecht, 2014). Besides, previous research carried out was limited to straight motion and work and energy (Yasa, 2014). While research related to the development of a self-rating scale to measure the SDL Skills is still not done.

The most widely used instrument in educational research to measure SDL skills is the Self-Directed Learning Readiness Scale (SDLRS) developed by Guglielmino (1977). The problem with this instrument is effectiveness and practicality. Field (1991) examined the relationship among item scales, factors and SDLRS total scores of respondents. Some items do not reach 0.35 on one of eight factors. Correlation between items and a total score of SDLRS is 0.30. These results indicate that the scale does not measure SDL skills. Based on the problems with the validity test of the instrument, Field (1991) suggested stopping using this instrument. Bonham (1991) also reported concerns about the validity of SDLR constructs by questioning the meaning of low scores. It was concluded that low scores did not measure low SDL readiness, thus construct validity is questionable for low SDLRS values. Although measuring devices such as the Guglielmino SDLRS have been developed, this self-rating scale is not available and is subject to fees for its use.

Gündüz & Selfi (2016) developed a Self-Directed Learning Implementation Skills Scale that is intended for elementary school students. In addition, most of the instruments developed are specifically for the world of medical or medical education such as the Self-Directed Learning Readiness Scale (SDLRS) developed by Guglielmino (1977), Self-Rating Scale of Self-Directed Learning (SRSSDL) developed by Williamson (2007), The Self-Rating Scale Self-Directed Learning Italian version (SRSSDL-Ita) developed by Cadorin, Ghezzi, Camillo, & Palese (2017), Self-Directed Learning Instrument (SDLI) developed by Shen et al. (2014) and Self-Directed Learning Readiness

Scale developed by Williams & Brown (2013).

Related to Science education, a Self-Directed Learning Skills Scale has been developed for students of prospective science teachers by Acar et al. (2015). Whereas for research in physics education as done by Yasa (2014) adapted the Williamson instrument (2007). Related to this, the development of a self-rating scale for self-directed learning skills for high school students especially those who are taking physics learning refers to the Self-Rating Scale of Self-Directed Learning (SRSSDL) developed by Williamson (2007). SRSSDL of sixty statements accompanied by a Likert (1932) scale of 1-5. The sixty statements are evenly distributed on five factors. These five factors are awareness, learning strategies, learning activities, evaluation, and interpersonal skills.

The purpose of this study is to develop an instrument self-rating scale of self-directed learning skills for high school students, especially MIPA majors. Another goal is to find out the validity and reliability of self-rating scales for advanced self-directed learning skills. Besides that as a result of the implementation phase, this study also aims to determine the Self-Directed Learning skills of high school students in class X, XI, and XII MIPA

Research Method

The DDI (Design, Development, and Implementation) research method was implemented and adapted from the development steps of Hinkin (1995). There are five primary stages in developing the Hinkin scale, namely: creating an item pool, expert conclusion, application, factor analysis, and reliability analysis. This research followed all the steps needed to develop the scale as stated by Hinkin.

The first step is to create an item group. At this stage, an assessment of the self-rating scale for existing self-directed learning skills was carried out. The most widely used instrument in educational research to measure SDL skills is the Self-Directed Learning Readiness Scale (SDLRS)

developed by Guglielmino (1977). The problem with this instrument is the validity and reliability and the cost of the use of this instrument.

In physics education research as done by Yasa (2014) adapted the Williamson instrument (2007). Related to this, the development of the self-rating scale of self-directed learning skills for high school students especially those who are taking physics learning refers to the Self-Rating Scale of Self-Directed Learning (SRSSDL) developed by Williamson (2007).

Based on the literature, 60 items will be used. The first thing to do is to translate the language, which is to translate from English to Indonesian. Then it provides remarks for some terms that are not familiar to high school students. Also, this statement is intended to equalize student perceptions regarding the statements on each item so that it is more focused on a particular purpose. In each item, it accompanied by a Likert rating scale in the form of 1-5. In the end, adjustments are made for instructions and student data.

The second stage is an expert conclusion. At this stage language, construct, and content validation is done. Language validation to assess the structure of the sentence and the meaning contained in it as well as the ease of students in accessing the meanings contained in the items under physics. Construct validation to assess the suitability of the item with its construct and the completeness of the self-rating scale instrument with scoring guidelines. Content validation is done to assess the statement on the item so that it corresponds to learning activities on physics. Validation was carried out by two expert lecturers. The results of this validation need to be revised on 1 item. The self-rating scale instrument in this stage has been revised and still consists of 60 items.

The third stage is implementation. According to the literature, the sample size is an essential problem in obtaining valid results from factor analysis. Kline (1994) (in Gündüz & Selfi, 2016) states that 200 participants are sufficient to identify the decisive

factors in factor analysis. Thus, at this stage, the final form of the self-rating scale draft was given to convenience samples of 216 high school students majoring in MIPA of classes X, XI, and XII. The participants' demographics in detail can be seen in Table 1.

Table 1. Demographics of Participants

Demographics of Participants		Fre- quency	%
Gender	Man	73	33.8
	Women	143	66.2
Class Level	Class X	121	56.0
	Class XI	77	35.7
	Class XII	18	8.3
Age Range	14 years	3	1.4
	15 years	52	24.1
	16 years	106	49.1
	17 years	49	22.7
	18 years	6	2.7
School	SMAK Yos Sudarso	54	25.0
	SMAN 1 Talun	162	75.0

The fourth stage is factor analysis. Factor analysis was performed with SPSS 17.0. This analysis is done to determine the construct validity of the self-rating scale. The fifth stage is the analysis of reliability. Reliability analysis was carried out with SPSS 17.0.

Findings and Discussion

Unidimensionality item

The unidimensional scale is one, where each item measures the same basic concept, in this case, the SDL skill. To test unidimensionality, which is whether the response to a particular item reflects a response to another item, the item-total correlation coefficient is done and the results obtained as shown in Table 2. The higher the coefficient for each item the more the item is included in the scale. Generally, the coefficient of less than 0.30 indicates that the item should be removed from the scale. Five of the sixty items produce coefficients less than 0.30 and are therefore excluded from the scale. The five deleted scales are 1.3, 1.9, 1.10, 2.2, and 3.12.

Factor Analysis of Awareness Indicators

The first condition for analyzing a factor is if the KMO (Kaiser Meyer Olkin) value is high, which is more than 0.5. The second condition is the value of Approx. Chi-Square on Bartlett's Test for correlation between variables more than 0.5 with Sig. on Bartlett's Test less than 0.05. At SPSS output the awareness indicator shows the KMO value is 0.818, the value of Approx. Chi-Square on Bartlett's Test is 425.056, and Sig. at Bartlett's Test is 0.000. Based on the three conditions, it can be said that the variables and samples used to allow further analysis.

Table 2. Statistics of Item-Total Correlations

Item	Mean	SD	Corrected item-total correlation	Cronbach's Alpha if items deleted
1.1	3.54	0.746	0.469	0.941
1.2	3.72	0.726	0.339	0.941
1.3	3.86	0.859	0.270	0.942
1.4	3.25	0.859	0.380	0.941
1.5	3.57	0.912	0.477	0.941
1.6	3.70	0.871	0.428	0.941
1.7	3.38	0.870	0.330	0.942
1.8	3.53	0.846	0.487	0.941
1.9	3.97	0.940	0.292	0.942
1.10	3.37	0.889	0.284	0.942
1.11	3.60	0.964	0.534	0.940
1.12	2.93	1,030	0.334	0.942
2.1	3.66	0.869	0.552	0.940
2.2	3.70	0.897	0.240	0.942
2.3	3.15	0.869	0.357	0.941
2.4	4.09	0.890	0.430	0.941
2.5	3.96	0.856	0.395	0.941
2.6	3.63	1,070	0.421	0.941
2.7	3.92	0.973	0.482	0.941
2.8	3.77	1,022	0.531	0.940
2.9	3.33	0.872	0.489	0.941
2.10	3.39	0.953	0.462	0.941
2.11	3.77	1,035	0.474	0.941
2.12	3.50	0.935	0.509	0.941
3.1	3.23	0.862	0.566	0.940
3.2	3.73	0.946	0.378	0.941
3.3	3.22	0.937	0.574	0.940
3.4	3.54	0.939	0.526	0.940
3.5	3.26	0.983	0.451	0.941
3.6	3.29	1,066	0.491	0.941
3.7	3.34	0.858	0.335	0.941
3.8	3.49	0.852	0.409	0.941
3.9	3.28	0.914	0.511	0.941
3.10	3.34	0.796	0.520	0.941

Item	Mean	SD	Corrected item-total correlation	Cronbach's Alpha if items deleted
3.11	3.76	0.794	0.395	0.941
3.12	4.04	0.870	0.104	0.943
4.1	3.55	0.770	0.423	0.941
4.2	3.56	0.861	0.417	0.941
4.3	3.40	0.919	0.430	0.941
4.4	3.72	0.969	0.504	0.941
4.5	3.95	0.755	0.415	0.941
4.6	4.26	0.830	0.496	0.941
4.7	4.01	0.792	0.343	0.941
4.8	3.44	0.828	0.483	0.941
4.9	3.13	0.980	0.367	0.941
4.10	3.41	0.906	0.516	0.941
4.11	3.35	0.902	0.617	0.940
4.12	4.35	1,010	0.469	0.941
5.1	3.99	0.972	0.399	0.941
5.2	3.64	0.888	0.583	0.940
5.3	3.93	0.912	0.565	0.940
5.4	3.68	0.871	0.570	0.940
5.5	3.98	0.937	0.452	0.941
5.6	4.09	0.933	0.529	0.940
5.7	3.73	0.976	0.424	0.941
5.8	3.62	0.908	0.609	0.940
5.9	3.45	0.856	0.549	0.940
5.10	3.42	0.952	0.464	0.941
5.11	3.65	0.922	0.574	0.940
5.12	3.91	0.989	0.467	0.941

Furthermore, what is seen is the Component Matrix Table. In the Component Matrix Table there should only be one component, but in this case, there are three components. If in the table there are more than one component this indicates that there is an invalid statement. Thus an item reduction is needed. To reduce items can be seen based on the MSA (Measures of Sampling Adequacy) contained in the Anti-Image Matrices Table. The lowest MSA value item must be discarded. The lowest MSA value found in the table is 0.774 for statement 1.9.

Then repeated analysis by removing item 1.9 and still found three components. Then the removal of item 1.10 was carried out, and there were still three components. Followed by deleting item 1.3 and there are still two components. Finally, after deleting item 1.6, one component is obtained in the Component Matrix Table as a factor value of each item. The factor value for each item of awareness indicator after reduction can be seen in Table 3.

Factor Analysis of Learning Strategies Indicators

At the SPSS output, the learning strategy indicator shows the KMO value is 0.861, the value of Approx. Chi-Square on Bartlett's Test is 433.430, and Sig. Bartlett's Test is 0.000. Of the three conditions, it can be said that the variables and samples used to allow further analysis. Furthermore, with the same analysis previously items were successively reduced between 2.2, 2.9 and 2.8. The factor value for each item indicator of learning strategy after reduction can be seen in Table 3.

Factor Analysis of Learning Activity Indicators

In the SPSS output, the learning activities indicator shows the KMO value is 0.808, the value of Approx. Chi-Square on Bartlett's Test is 551.069, and Sig. at Bartlett's Test is 0.000. Of the three conditions, it can be said that the variables and samples used to allow further analysis. Furthermore, with the same analysis as before, items were reduced in succession, including 3.12, 3.2, 3.7, 3.9 and 3.8. The factor value for each item in the learning activity after reduction can be seen in Table 3.

Factor Analysis of Evaluation Indicators

At the SPSS output, the indicator of learning activities shows the KMO value is 0.846, the value of Approx. Chi-Square on Bartlett's Test is 693.458, and Sig. Bartlett's Test is 0.000. Of the three conditions, it can be said that the variables and samples used to allow further analysis. Furthermore, with the same analysis as before, item reduction was carried out in succession including 4.12, 4.7, 4.3, and 4.2. Factor value for each item of evaluation indicator after the reduction can be seen in Table 3.

Factor Analysis of Interpersonal Capability Indicators

At the SPSS output, the indicator of learning activity shows the value of KMO is 0.926, the value of Approx. Chi-Square on

Bartlett's Test is 1147.279, and Sig. Bartlett's Test is 0.000. Of the three conditions, it can be said that the variables and samples used to allow further analysis to be carried out. In the indicator of interpersonal ability is found in one component in the Component Matrix Table, so item reduction does not need to be done. The factor values for each item indicator of interpersonal skills can be seen in Table 3.

Table 3. Results of Analysis After Reduction

Item	Results of Factor Analysis After Reduction				
	Aware- ness	Learning Strategy	Learning Activities	Evalu- ation	Interperson- al ability
1.1	0.594				
1.2	0.511				
1.4	0.522				
1.5	0.683				
1.7	0.548				
1.8	0.705				
1.11	0.596				
1.12	0.465				
2.1		0.559			
2.3		0.520			
2.4		0.698			
2.5		0.662			
2.6		0.655			
2.7		0.572			
2.10		0.660			
2.11		0.744			
2.12		0.553			
3.1			0.625		
3.3			0.731		
3.4			0.660		
3.5			0.627		
3.6			0.653		
3.10			0.608		
3.11			0.478		
4.1				0.575	
4.4				0.574	
4.5				0.504	
4.6				0.595	
4.8				0.697	
4.9				0.645	
4.10				0.745	
4.11				0.742	
5.1					0.675
5.2					0.749
5.3					0.772
5.4					0.694
5.5					0.715
5.6					0.747
5.7					0.674
5.8					0.776
5.9					0.656
5.10					0.572
5.11					0.733
5.12					0.682

Table 4 presents sample sizes of central tendency and dispersions for the total scale and subscale. The total score for this sample is not normally distributed as indicated by the Sig. on the Kolmogorov-Smirnov normality test is 0.000. It can be concluded that the number of scores higher than 150 which is 158.12, cannot indicate the readiness of students for SDL.

Table 4. Reliability Coefficient After Item Reduction

	Item Amount	Mean	Std. Deviation	Cronbach's α
Consciousnes (Awareness)	8	27.51	4,031	0.717
Learning Strategy	9	33.08	5,360	0.806
Learning Activities	7	23.63	4,023	0.777
Evaluation	8	28.81	4,433	0.790
Interpersonal ability	12	45.08	7.819	0.907
Total	44	158.12	20,334	0.933

The internal consistency of each component is estimated using the Cronbach alpha coefficient. The values calculated for each item included: awareness (8 items, $\alpha = 0.717$), learning strategies (9 items, $\alpha = 0.806$), learning activities (7 items, $\alpha = 0.777$), evaluation (8 items, $\alpha = 0.790$), and interpersonal abilities (12 items, $\alpha = 0.907$). The overall internal reliability coefficient (Cronbach Alpha) of the self-rating scale is $\alpha = 0.933$. According to DeVaus (Fisher, 2001), a scale with an alpha calculation greater than 0.70 he thinks has an acceptable level of internal consistency (although consistency for other types of scale, such as achievement tests, is generally estimated to be at or above 0.80).

The primary objective of this study was to develop a valid and reliable scale to understand self-directed learning skills for high school students who will become life-long learners both now and in the future. According to the factor analysis carried out, the scale is grouped into five factors that is awareness, learning strategies, learning activities, evaluations, and interpersonal skills.

The findings indicate that the scale developed has appropriate qualifications to determine students' self-directed learning abilities. The scale can help both students and teachers to understand self-directed learning skills in the five factors, especially high school students.

The developer of the Self-Rating Scale of Self-Directed Learning (SRSSDL), Williamson (2007), states that awareness is the ability to detect learning needs. Acar et al. (2015) in their study confirmed that students who have the skills of self-directed learning have awareness towards their responsibility in learning, acting independently without the help of others, having a high sense of curiosity, enthusiasm, confidence, they have the ability to manage time and make plans to complete the work they have set goals. Furthermore, Taqipour et al. (2016) mentioned that the dimension of consciousness shows that students are aware of their responsibility to learn. This awareness is reflected in the following behaviors, namely students able to identify of their own learning needs, customize their learning goals, can choose the best method of learning for them, balancing learning with their daily activities, updating their learning and independent learning. As mentioned by Istiyani (2009), students' awareness is needed in order to maximize their learning.

The factor of learning strategies state strategies used in a variety of different situations to develop student learning (Williamson, 2007). Seifert (1993) defines learning strategies as mental events carried out by students to achieve some desired goals. Deshler & Schumaker (1986) reinforces the assertion that the student will become independent learners and players in their learning when they begin to produce their learning strategies that apart from the help of teachers. Students' mastery of learning strategies enables them to successfully analyze and solve the new problems they face both at the academic and non-academic environment. Overall, mastery of the learning strategy has to do with self-directed learning skills because the mastery of learning strategies is not only im-

mediate but also generalizes learning strategy skills for different situations and settings from time to time. Things like this are commonly known as a life long learning.

The factor of learning activities covers a range of activities carried out an individual in their study (Williamson, 2007). Learning activities are defined by Eurostat (2006) as an activity of individuals organized with the purpose to improve their knowledge, skills, and competencies. The character of this learning activity is done intentionally for specific purposes. Besides, another important character is a form of organized activity, in self-directed learning skills of learning activities are organized by the students who usually involves the transfer of information in the sense that more general (may be an idea, a message, knowledge, strategy). For less organized activities (in the sense that each student will be different) such as self-directed learning, Eurostat (2006) states that determining whether student activities include learning activities or not learning activities must be decided more careful, this depends on presence or absence of intention to study. This is slightly different from more organized activities which are designed based on the students' desire or effort to learn, which will not change the nature of student activities into non-learning activities.

The ability to effectively evaluate has long been known as a cognitive process and is an essential educational goal (Airasian & Miranda, 2002; Krathwohl, 2002). The factor of evaluation by Williamson (2007) describes student's ability of learning evaluation in different situations and students get feedback from their learning. If students can evaluate their work, then their learning will not depend entirely on external evaluators (Warren, 2010). Thus students will be able to identify and correct their own mistakes, allowing them to learn better themselves. Such is the hope of mastering skills in self-directed learning.

According to Williamson (2007), the factor of interpersonal ability refers to the communication ability of students with others to expand their learning volume.

These factors are essential forms of interpersonal interaction (Babonea & Munteanu, 2012), which can strengthen the bond between students and their colleagues and between students and teachers. Past research on communication in learning has identified several variables of interpersonal positively related to learning. These variables include the relationship or closeness between the teacher and students (Andersen, 1979; Frymier & Houser, 2000; Hughes, 2012), attractiveness and communication style (Norton & Pettegrew, 1977), adjustments to the school associated with progress and challenges encountered by students (Hughes, 2012), student motivation and beliefs (Koca, 2016), humor (Wanzer & Frymier, 1999; Ziyaeemehr & Kumar, 2014), and caring (Teven & McCroskey, 1997) contribute to an understanding of student relations in class dynamic. Whereas in science education, a sufficiently emphasized variable has an open mind, that is, with open-minded students can change their beliefs (Burns & Norris, 2009). As confirmed by Hare & McLaughlin (1998), even though students have certain beliefs they can still accept rationally.

All of the variables included in the interpersonal skills affect how student learning takes place. For example, when students believe that their colleagues and teachers like and respect them, they will tend to be successful during learning activities (Goodenow, 1993; Ryan, Pintrich, & Midgley, 2001). In academic settings, interpersonal skills help engagement and student learning (Lindsey & Rice, 2015). By understanding and managing interpersonal skills, students can manage personal intellectual growth and social growth. Increasing interpersonal skills will help students enrich individual relationships, do better learning activities because they can overcome work or tasks that they have designed better.

Conclusion

Based on the results of research and data analysis, it is known that, of the sixty self-rating scale items that have been developed, there are forty-four items which fall

into the valid category. The forty-four self-rating scale items consist of five indicators and have a Cronbach Alfa reliability level of $\alpha = 0.933$. This shows that the forty-four self-rating scale items which are divided into five indicators have high reliability so that it can be used to measure self-directed learning skills for high school students, especially majors in Mathematics and Natural Sciences both class X, class XI, and class XII. The self-rating scale developed to measure self-directed learning skills with SDL skill indicators in the form of awareness, learning strategies, learning activities, evaluation, and interpersonal skills.

Tests that have been declared valid and reliable are then used to measure self-directed learning skills. The number of respondents was 216 from two different schools namely SMAN 1 Talun and SMAK Yos Sudarso Kepanjen. The total score for this sample is not normally distributed as indicated by the Sig. on the Kolmogorov-Smirnov normality test is 0.000. It can be concluded that the number of scores higher than 150 which is 158.12 cannot indicate the readiness of students for SDL. On a scale conversion of 1-100, the highest SDL student skill score was 93, and the lowest SDL skill score was 31 (SD = 20.334).

The self-rating scale developed to measure students' self-directed learning can be used as a cost-effective research or educational tool. This self-rating scale will help educators, especially in the MIPA department for all levels in diagnosing student learning needs so that educators can apply teaching strategies that are appropriate to student needs. In connection with this development paucity of research that cannot be indicated the readiness of student's self-directed learning skills. The suggestion for the next researcher is to conduct case study research related to the students' self-directed learning skills who are studying physics to examine individuals in depth.

References

- Acar, C., Kara, I., & Taşkin Ekici, F. (2016). Development of self directed learning

- skills scale for pre-service science teachers. *International Journal of Assessment Tools in Education*, 2(2), 3–13. <https://doi.org/10.21449/ijate.239562>
- Airasian, P. W., & Miranda, H. (2002). The role of assessment in the revised taxonomy. *Theory Into Practice*, 41(4), 249–254. https://doi.org/10.1207/s15430421tip4104_8
- Andersen, J. F. (1979). Teacher immediacy as a predictor of teaching effectiveness. *Annals of the International Communication Association*, 3(1), 543–559. <https://doi.org/10.1080/23808985.1979.11923782>
- Babonea, A., & Munteanu, A. (2012). Towards positive interpersonal relationships in the classroom. In *International Conference of Scientific Paper*.
- Bonham, L. A. (1991). Guglielmino's self-directed learning readiness scale: what does it measure? *Adult Education Quarterly*, 41(2), 92–99. <https://doi.org/10.1177/0001848191041002003>
- Burns, D. P., & Norris, S. P. (2009). Open-minded environmental education in the science classroom. *Paidousis*, 18(1), 35–42.
- Cadorin, L., Ghezzi, V., Camillo, M., & Palese, A. (2017). The self-rating scale of self-directed learning tool: findings from a confirmatory factor analysis. *Journal of Nursing Education and Practice*, 7(2). <https://doi.org/10.5430/jnep.v7n2p31>
- Deshler, D. D., & Schumaker, J. B. (1986). Learning strategies: an instructional alternative for low-achieving Adolescents. *Exceptional Children*, 52(6), 583–590. <https://doi.org/10.1177/001440298605200611>
- Eurostat. (2006). *Classification of learning activities-manual*. Luxembourg: Office for Official Publications of The European Communities.
- Field, L. (1991). Guglielmino's self-directed learning readiness scale: should it continue to be used? *Adult Education Quarterly*, 41(2), 100–103. <https://doi.org/10.1177/000184819104100204>
- Frymier, A. B., & Houser, M. L. (2000). The teacher-student relationship as an interpersonal relationship. *Communication Education*, 49(3), 207–219. <https://doi.org/10.1080/03634520009379209>
- Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationships to motivation and achievement. *Journal of Early Adolescence*, 13(1), 113–126.
- Guglielmino, L. M. (1977). *Development of the self-directed learning readiness scale*. Doctoral dissertation. University of Georgia.
- Gündüz, G. F., & Selfi, K. (2016). Developing a “self-directed learning implementation skills scale for primary school students”: validity and reliability analysis. *Agathos: An International Review of the Humanities and Social Sciences*, 7(1).
- Hare, W., & McLaughlin, T. (1998). Four anxieties about open-mindedness: reassuring Peter Gardner. *Journal of the Philosophy of Education*, 32(2), 283–292. <https://doi.org/10.1111/1467-9752.00093>
- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21(5), 967–988. <https://doi.org/10.1177/014920639502100509>
- Huang, M. B. (2008). *Factors influencing self-directed learning readiness among Taiwanese nursing students*. Thesis. The Queensland University of Technology. Retrieved from https://eprints.qut.edu.au/20709/1/Mei-hui_Huang_Thesis.pdf

- Hughes, J. N. (2012). Teacher–student relationships and school adjustment: progress and remaining challenges. *Attachment & Human Development, 14*(3), 319–327. <https://doi.org/10.1080/14616734.2012.672288>
- Istiyani, D. (2009). Kesadaran dan self-directed learning sebagai model pembelajaran alternatif dalam era neoliberalisme. *Forum Tarbiyah, 7*(2), 131–142.
- Kan'an, A., & Osman, K. (2015). The Relationship between self-directed learning skills and science achievement among Qatari students. *Creative Education, 06*(08), 790–797. <https://doi.org/10.4236/ce.2015.68082>
- Knowles, M. S. (1975). *Self-directed learning: a guide for learners and teachers*. Chicago: Follett Publishing Company.
- Koca, F. (2016). Motivation to learn and teacher-student relationship. *Journal of International Education and Leadership, 6*(2), 1–20.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Practice, 41*(4), 212–218.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5–55.
- Lindsey, N. S., & Rice, M. L. (2015). Interpersonal skills and education in traditional and online classroom environments. *Journal of Interactive Online Learning, 13*(3), 126–136.
- Malan, S. B., Ndlovu, M., & Engelbrecht, P. (2014). Introducing problem-based learning (PBL) into a foundation programme to develop self-directed learning skills. *South African Journal of Education, 34*(1), 1–16. <https://doi.org/10.15700/201412120928>
- Norton, R. W., & Pettegrew, L. S. (1977). Communicator style as an effect determinant of attraction. *Communication Research, 4*(3), 257–282. <https://doi.org/10.1177/009365027700400302>
- Ryan, A. M., Pintrich, P. R., & Midgley, C. (2001). Avoiding seeking help in the classroom: who and why? *Educational Psychology Review, 13*(2), 93–144. <https://doi.org/10.1023/A:1009013420053>
- Scott, C. L. (2015). The Futures of Learning 2: what kind of learning for the 21st century. Retrieved from <http://unesdoc.unesco.org/images/0024/002429/242996e.pdf>
- Seifert, T. (1993). Learning strategies in the classroom. Retrieved from <https://www.mun.ca/educ/faculty/mwatch/vol2/seifert.html>
- Shen, W., Chen, H., & Hu, Y. (2014). The validity and reliability of the self-directed learning instrument (SDLI) in mainland Chinese nursing students. *BMC Medical Education, 14*(1), 108. <https://doi.org/10.1186/1472-6920-14-108>
- Taqipour, M., Abbasi, E., Naeimi, A., Ganguly, S., & Miandashti, N. Z. (2016). An investigation of self-directed learning skills among the Iranian agricultural students (case of agricultural college, Tarbiat Modares University). *Journal of Agricultural Science and Technology, 18*(1), 15–26. Retrieved from http://mcej.modares.ac.ir/browse.php?a_id=6847&sid=23&slc_lang=en
- Teven, J. J., & McCroskey, J. C. (1997). The relationship of perceived teacher caring with student learning and teacher evaluation. *Communication Education, 46*(1), 1–9. <https://doi.org/10.1080/03634529709379069>
- Wanzer, M. B., & Frymier, A. B. (1999). The relationship between student perceptions of instructor humor and students' reports of learning. *Communication Education, 48*(1), 48–62. <https://doi.org/10.1080/03634529909379152>

- Warren, A. R. (2010). Impact of teaching students to use evaluation strategies. *Physical Review Special Topics-Physics Education Research*, 6(2), 1–12.
- Williams, B., & Brown, T. (2013). A confirmatory factor analysis of the Self-Directed Learning Readiness Scale. *Nursing & Health Sciences*, 15(4), 430–436. <https://doi.org/10.1111/nhs.12046>
- Williamson, S. N. (2007). Development of a self-rating scale of self-directed learning. *Nurse Researcher*, 14(2), 66–83. <https://doi.org/10.7748/nr2007.01.14.2.66.c6022>
- Yasa, P. (2014). Model belajar pemecahan masalah berbasis konteks untuk pengembangan kompetensi generik siswa kelas X SMA Negeri 3 Singaraja. In *Seminar Nasional FMIPA UNDIKSHA IV Tahun 2014*. Universitas Pendidikan Ganesha.
- Ziyaeemehr, A., & Kumar, V. (2014). The relationship between instructor humor orientation and students' report on second language learning. *International Journal of Instruction*, 7(1), 91–105. Retrieved from <https://eric.ed.gov/?id=EJ1085255>

INDONESIAN-LANGUAGE VERSION OF GENERAL SELF-EFFICACY SCALE-12 USING BAYESIAN CONFIRMATORY FACTOR ANALYSIS: A CONSTRUCT VALIDITY TESTING

Muhammad Dwirifqi Kharisma Putra

Universitas Islam Negeri Syarif Hidayatullah Jakarta

Wardani Rabayu

Universitas Negeri Jakarta

Jahja Umar

Universitas Islam Negeri Syarif Hidayatullah Jakarta

Abstract

The General Self-Efficacy Scale 12 (GSES-12) is a brief measure for assessing self-efficacy. This study aimed to revise an Indonesian language version of the GSES-12 that was translated and adopted from previous research. The revision conducted by following the Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures, and the final version was administered to 303 (132 male, 171 female) Indonesian students, with a mean age of 19.56 years (SD: 1.20). This study is presented to establish the construct validity of this instrument further. The results of Bayesian CFA revealed a higher-order structure of factor representing constructs of self-efficacy. Considering the theoretical background and the best model fit indices (PPP-value = 0.549 and BRMSEA = 0.001), it is concluded that the Indonesian version of GSES-12 appears to be a valid instrument in assessing self-efficacy in Indonesian speaking students and is expected to facilitate the examination of self-efficacy in Indonesian speaking populations.

Keywords: *Bayesian, confirmatory factor analysis, general self-efficacy scale-12, self-efficacy*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.20008>

Contact *Muhammad Dwirifqi Kharisma Putra*

 *dwirifqi@icloud.com*

 *Fakultas Psikologi UIN Syarif Hidayatullah Jakarta
Jl. Kertamukti No. 5 Cirenden Jakarta 15412, Indonesia*

Introduction

Self-efficacy has become a commonly studied variable in education, psychology, health, and also organizational field. Recent developments show that theoretically, self-efficacy has undergone many revisions by the early developers of the theory (Bandura, 2012), on which this theoretical development indicates that the use of self-efficacy is increasingly widespread. The ongoing development also shows that the measurements made on self-efficacy should be developed as well to adjust the theoretical developments.

However, in fact, Bandura did not develop a measuring instrument based on the theory of self-efficacy that he developed. Therefore, various studies on the measurement of self-efficacy have produced many alternative theories describing the self-efficacy itself.

Self-efficacy is commonly understood as task-specific or domain-specific, but some researchers also conceptualize it as a common generalization, a concept (Luszczynska, Gutiérrez-Doña, & Schwarzer, 2005). Currently, research on self-efficacy, by and large, focuses on generalizations such as trait from the dimension of self-efficacy known as general self-efficacy (Chen, Gully, & Eden, 2001). Research conducted in the context of self-efficacy generally requires other variables to explain one's self-efficacy in certain behaviors which ultimately creates general self-efficacy that can be useful as an explanatory value in describing one's self-efficacy (Bosscher & Smit, 1998). On this basis, a research was developed exploring the factor structure of self-efficacy as an alternative theory.

Exploring the latest research related to general self-efficacy, we found a variety of recent studies in 2018 that also measured the construct of self-efficacy with general self-efficacy. These studies found that high self-efficacy in the context of research made students' academic performance even higher (Tiyuri et al., 2018), that self-efficacy is an important factor that can make students successful in facing exam (Willson-Conrad & Kowalske, 2018), and that in the health field,

self-efficacy increases motivation in recovering someone's illness (Klompstra, Jaarsma, & Strömberg, 2018). These studies show that until now self-efficacy is a construct that is still developing and commonly used in various fields, not only in the fields of psychology or education but other disciplines such as the health field.

In Indonesia, similar developments also occur regarding studies on self-efficacy, as traced in various journals in Indonesia published in the range of 2016-2018. These findings indicate that self-efficacy, by and large, has been studied in Indonesia either in the fields of psychology, education or health over the past two years. However, none of these articles focused on adaptation and validation of the measurement of self-efficacy carried out based on the guideline of adaptation of the measuring instruments, so that the measurements taken were independent between one researcher to another.

The rapid development of general self-efficacy in research in the field of psychology and education was initially caused by the availability of instruments that can be used. Based on the search of researchers for measuring self-efficacy, researchers obtain measuring instruments that can be used in measuring self-efficacy, which is entirely focused on general self-efficacy. Those instruments are called General Self-Efficacy Scale Sherer (Sherer et al., 1982), General Self-efficacy Scale (Schwarzer & Jerusalem, 1995), and also General Self-Efficacy Scale 12 (GSES-12) (Bosscher & Smit, 1998). For studies in Indonesia, these measuring instruments are used with adaptations that are independent of each other.

In the previous study, Putra and Tresniasari (2015) adapted the GSES-12 instrument into Indonesian and attached it to the publications conducted. However, the research had not followed the available guidelines in adapting measuring instruments (e.g., Beaton, Bombardier, Guillemin, & Ferraz, 2000), so that even though it has been used in the research, there are the concerns about the psychometric aspects of the measuring instruments that have been

adapted. Therefore, the objective of this research was to adapt the GSES-12 measuring instrument into the Indonesian language, but the adaptation was conducted based on the guidelines proposed by Beaton et al. (2000) and the reporting of the analysis results was carried out based on the guidelines proposed by Schreiber, Nora, Stage, Barlow, and King (2006). This research will focus on construct validity to confirm the structure of the factors underlying the GSES-12 measurement model. Construct validity is defined as the extent to which the scale measured the intended construct, where the method commonly used in construct validity is CFA (Kaplan, 2000).

This research uses CFA with Bayesian approach, known as Bayesian CFA as a special case of Bayesian SEM, where the results of data analysis would be compared with previous studies (see, Putra & Tresniasari, 2015) to compare the quality of psychometric aspects obtained from this adapted measuring instrument. Various advantages of Bayesian CFA use would be obtained, for example, the flexibility of this approach to diagnose models that have specification errors, models whose estimates have deadlocks, and analysis with small sample sizes. However, the biggest advantage is that the resulting score in the form of an estimate of the true score in the form of the highest quality score, which is plausible values, so that when used in a further analysis like regression analysis, it would produce a very good estimate. Other advantages of using the Bayesian approach can be seen in various literature (see, van de Schoot et al., 2014; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). Speaking of which, in this research, the Bayesian CFA used to test construct validity is the Indonesian version of GSES-12.

Research Method

Participants

This study included a sample of 303 (132 males, 171 females), Indonesian students. All of the participants were under-

graduate students in various departments of the Syarif Hidayatullah State Islamic University Jakarta. The mean age of the sample was 19.56, with a range of 18-22 years. The willingness of the respondents to participate in research is available in the form of informed consent. The sample size of 303 had met the minimum sample size in using the CFA method, which is the criteria that a minimum sample size is 200 (Hoogland & Boomsma, 1998) and 265 (Muthén & Muthén, 2002) so that in this research, the use of CFA was not interrupted by insufficient sample size problems to obtain optimal estimation results.

GSES-12 and Adaptation Process

The Indonesian version of the General Self-efficacy Scale-12 (GSES-12; Bosscher & Smit, 1998) was used to assess self-efficacy. The GSES-12 consists of 12 items with the following subscale: initiative (item 1, 2, 4, 12), effort (item 3, 5, 7, 8) and persistence (item 6, 9, 10, 11), rated across a 5-point Likert-type scale. In adapting the GSES-12 instrument, researchers referred to the procedures described in the Guidelines for the Process of the Cross-Cultural Adaptation of Self-Report Measures (Beaton et al., 2000). The process of adaptation conducted consisted of five stages: Initial Translation, Synthesis of Translations, Back Translation, Expert Committee and Test of the Prefinal Version. The GSES-12 items have gone through adaptation processes in stages 1-4. Stage 5 was not applied as it is not necessary since the method used has produced a plausible value as a true score.

The translation process was carried out by experts at a professional institution of UIN Syarif Hidayatullah Jakarta Language Center. The items that were the result of adaptation from GSES-12 were modified on the Likert scale, on which the original scale using a Likert scale model with a modified five-point range was changed into a 4-point scale range, namely "SS" (strongly agree), "S" (agree), "TS" (disagree) and "STS" (strongly disagree). This was done based on suggestions from various previous studies suggesting that the existence of a response in the

middle position (for example, neutral) would cause respondents to tend to choose that option, and consequently, it affected the validity of the measurement model (Moors, 2008). Then, the response of the respondents' answers was given a predetermined score as follows: SS = 4, S = 3, TS = 2, STS = 1, and for unfavorable items, the scoring was done otherwise. The data analysis performed with Bayesian CFA was analyzed using the Mplus 8.1 program (Muthén & Muthén, 2017). Nevertheless, due to the fit model index that had just been found in the Bayesian context, the Bayesian Root Mean Square Error Approximation (BRMSEA), which is not yet available in Mplus, the computation was done with the 'Blavaan' package (Merkle & Rosseel, 2018) in the R version 3.5.1 program.

Bayesian CFA (Confirmatory Factor Analysis)

To test the construct validity of the General Self-Efficacy Scale-12 (GSES-12) instrument, the researchers used the CFA method (confirmatory factor analysis). As mentioned in the introduction, in the research of applied science fields, factor analysis is the most commonly used method for evaluating psychometric aspects of measuring instruments with a large number of items (e.g., questionnaires). The basic CFA equation derived from the common factor model in the form of a matrix can be written as in Equation (1) (Cai, 2013; Kaplan, 2000):

$$\Sigma = \Lambda\Phi\Lambda' + \Theta_{\epsilon} \quad (1)$$

In which, Σ is a symmetric correlation matrix with $p \times p$ dimension from indicators of as many as p , Λ is the λ factor load matrix of $p \times m$ dimension, Φ is a symmetric correlation matrix with $m \times m$ dimension from the correlation between factors, and Θ_{ϵ} is a diagonal matrix with $p \times p$ dimension from ϵ unique variances. Referring to the matrix algebra, the matrix used in the factor analysis and SEM is denoted by the Greek letter capital (e.g., Λ , Ψ , Θ) and more specific elements of the matrix are represented by

Greek letters that are not capital (e.g., λ , ψ , ϵ) (Brown, 2015). Equation (1) is a CFA model commonly known as the "first-order". But in this research, the CFA model used was a higher-order model, also known as "second-order". This model was first introduced by Joreskog (1971), where Equation (1) was added to be Equation (2):

$$\Sigma = B(\Lambda\Phi\Lambda' + \psi^2)B' + \theta^2 \quad (2)$$

In which $B_{((pk))}$ is a factor load of items in the first-order factor of as many as k , $\Theta_{((pxp))}$ is a diagonal matrix containing error variance from the first-order level factor, $\Lambda_{((kxr))}$ contains a load factor from the first-order level factor to the second-order level factor as many as r , $\Phi_{((rxr))}$ is a correlation matrix between the factors of the second-order level and ψ^2 is a diagonal matrix which contains error variance from the second-order level factor (Joreskog, 1971).

This model can be used when: (1) CFA at the first-order level is conceptually valid; (2) testing the amount and pattern of the correlation between factors in the first-order model; and (3) testing the fitness of second-order models based on conceptual and theoretical foundations. The higher-order model itself can free up factors to correlate with each other and propose a model on which these factors are part of one main factor of the construct commonly used in testing the theory. Unlike the first-order model, this model must have a metric reference unit that is generally done by standardizing higher-order parts, but it is also possible to do so with a model that has not been standardized by using indicators whose scales are referred to for higher factors (Brown, 2015).

It should be noted that this research used the Bayesian Approach applied to the CFA model. Therefore the estimation method used is no longer the maximum likelihood but the Bayesian-based estimation method. Technical explanations regarding the application of the Bayesian approach in the social sciences have been well summarized in the available literature (e.g., Kaplan, 2014). With the Bayesian approach, the fit model index

of the commonly used classical approach like Chi-square and RMSEA has differences, both in terms of philosophical, computational, and interpretation. Hence, in the next sub-chapter, we will explain the fit model index used in this research.

Model Fit Indices

The goodness-of-fit classical statistical test is not available in the analysis using the Bayesian approach (Brown, 2015). As quoted from van de Schoot et al. (2014), when the researchers use SEM in analyzing data to answer research questions, the researchers do not only test one hypothesis, but they do an overall evaluation of the model. The fit model test with the use of the Bayesian approach is by and large related to how to measure the prediction accuracy against a model known as posterior predictive checking. The basic idea about posterior predictive checking is that there should be small differences between the data generated from the actual model and data. All differences or deviations between the two indicate a possible specification error with the model. Thus, it can be briefly explained that the posterior predictive checking is a method used to assess the quality of the specified model from the point of view of the accuracy of the predictions made. Posterior predictive checking itself was developed by Gelman, Meng, and Stern (1996).

One approach that can be used to quantify the fit model is to calculate the Bayesian posterior predictive p-value (PPP value). The statistical test of the model, the chi-square value, is calculated based on the data compared with the same statistical test, then the generated data is determined. Thus, PPP value is defined as the proportion of the chi-square value obtained from the generation of data that matches the actual data. The amount of PPP value, which is in the range of 0.50, indicates that the model is well fit (van de Schoot et al., 2014). The same criteria are also explained by Muthén and Asparouhov (2012) who stated that the criteria for model fit are: (1) PPP value close to 0.50, and (2) in 95% confidence intervals the

lower limit is negative and the difference is 0 which falls in the middle of the interval.

It should be noted that PPP values should not be interpreted in the same way as the p-value for χ^2 model using the classical approach. Unlike p-value in the classical approach, PPP does not depend on asymptotic theory. In addition to PPP values, the posterior predictive checking results in a 95% confidence interval from the difference between the statistical tests on the sample data and the generated data (Brown, 2015). The absence of the recommended lower limit as a minimum limit makes the writer use strict standards where the value of 0.50 which is proven to be optimal is used as the lower limit to interpret the PPP value used in this research. In addition to PPP values, recent developments indicate that RMSEA commonly used in CFA in the classical approach is available in the Bayesian CFA context, which is Bayesian root mean square error approximation (BRMSEA; Hoofs, van de Schoot, Jansen, & Kant, 2018) where criteria of <0.05 indicates that the model fits really well. Therefore, these two model fit indices are used in this research.

Bayesian Estimation (BAYES Estimator)

Unlike the ML estimator which focuses on the computation of point estimation from parameters in models that have asymptotic properties, the purpose of the analysis using the Bayesian approach is to estimate the features of the posterior distribution (which does not depend on the large-sample theory). In Bayesian analysis, a numerical algorithm called Markov Chain Monte Carlo (MCMC) is used to estimate the posterior distribution containing parameters in the model produced by $P(\theta|y)$ data (Brown, 2015). In the Bayesian approach the posterior distribution is the result of estimation of the values on the features of the population from the things studied which are obtained by combining empirical data with the existing and previous expectations and based on the existing knowledge or previous opinions (prior distribution) (van de Schoot & Depaoli, 2014; van de Schoot et al., 2014,

2017). If modeled, posterior distribution can be described by Equation (3):

$$\begin{aligned} \text{posterior} &= \text{parameter}|\text{data} \\ &= \frac{\text{data}|\text{parameter} \times \text{parameters}}{\text{data}} \\ &= \frac{\text{likelihood} \times \text{prior}}{\text{data}} \\ &\propto \text{likelihood} \times \text{prior} \quad (3) \end{aligned}$$

In which \propto means ‘proportional to’ and not included in data. The prior distribution is modified by likelihood to get the posterior distribution. The Bayesian estimation method produces the average, mode, or median of the posterior distribution. At the same time, the posterior distribution is obtained through the MCMC algorithm (Muthén & Asparouhov, 2012). Although in conducting the simulation method there are large numbers of random draws, the MCMC tries to make an approximation of the joint parameter distribution in the model (posterior distribution) based on the random draws of parameter values according to the conditional distribution from a set of parameters, when another set of parameters is known. In other words, in the Markov chain, large numbers of samples are in "picture"/created from conditional distributions, and the distribution created is summarized (Brown, 2015). There are several different types of algorithms in MCMC, one which is the Gibbs sampler that is a basic algorithm in the Mplus program when analyzing with the Bayesian approach (Brown, 2015; Kaplan, 2014). However, in this research, the algorithm used was Metropolis-Hastings. This algorithm has been known for its use in the CFA method in numerous studies (see, Asparouhov & Muthén, 2012; Bashkov, 2015; Cai, 2008; Cai, 2010b, 2010a; Yang & Cai, 2014) where efficiency is proven until its use in high-dimensional models.

For prior informative testing, one that needs to be considered in Bayesian analysis is the quantification and operationalization of MCMC convergence. It is certainly quite difficult because the MCMC aims to con-

verge on the posterior distribution compared to point estimate (unlike the ML estimator). The use of parallel chains and having different initial values will allow us to measure the level of convergence. MPLUS employs the Gelman-Rubin convergence criteria to determine the convergence level of the Bayesian estimation method. These criteria measure the convergence by considering the variability that exists within or between chains in parameter estimation carried out by the name of potential scale reduction (PSR) on the factor. Using the estimated variance between chains (B) and within-chains (W), PSR is calculated (Brown, 2015; Kaplan, 2014), as presented in Equation (4):

$$PSR = SQRT(W + B)/W \quad (4)$$

In which the PSR value around or fits 1.0 indicates that it has converged. The ratio of variance close to 1.0 shows that convergence has been successfully achieved when variations between chains are small compared to within-chain variations. Gelman, Carlin, Stern, and Rubin (2004) recommended a PSR value of 1.10 for all parameters as an illustration of convergence. Except for models with a small number of parameters, the value of 1.10 is used as the default by MPLUS to determine the convergence in the Bayesian approach. In addition, we can also check the convergence of MCMC in a more subjective way by studying convergence plots formed from chains on each parameter (this is often referred to as trace plots or history plots) and by looking at the prior distribution of the parameters as well as the autocorrelation of the chain (Brown, 2015).

Findings and Discussion

In addition to conducting the first-order CFA test on 12 items, the researcher wanted to test whether the 12 items came from three dimensions, which are unidimensional initiative, effort and persistence, meaning that they only measured self-efficacy. The results of the CFA analysis conducted with the second-order model obtained a model fit with PPP value of 0.549 (95%

CI = -33.065, 30.422). Like the limit that the PPP value has been explained previously, which is equal to 0.50, it can be stated that the higher-order model of GSES-12 is a model that is well fit. It is indicated by how the proposed model does not experience specification and convergence errors. To be able to see an overview of the CFA higher-order model in this study, the path diagram of the higher-order CFA model of GSES-12 is presented in Figure 1.

At first glance, it can be seen in Figure 1 that there is no metric scaling in each dimension. It occurs because the CFA solution described is a solution that uses a standardized unit of measurement. After obtaining a PPP-value of 0.549 > 0.50, it can be stated that the second-order model with one factor (ksi) and three dimensions (eta) can be accepted and fit very well. It means that all items derived from three dimensions, which are initiative, effort, and persistence really

only measure one factor that is self-efficacy. In Table 1, the convergence of the model with the Bayesian approach, which contains iteration and PSR information from the analysis carried out is presented.

Although it has been previously explained that the convergence criterion of the model when PSR is at the value of 1.10, from the data analysis carried out the number of iterations of 20000 is determined in advance so that it can be seen in Table 1 that in the 8000th iteration the model has actually converged. However, when the iteration is set to be greater than 8000, it can be seen that in the 20000th iteration, the lowest PSR is 1.039. Thus, when it is compared to the 8000th iteration, there is a difference in the model index fit that is better than the analysis using 20000 iterations compared to cases when we did not determine the number of iterations first.

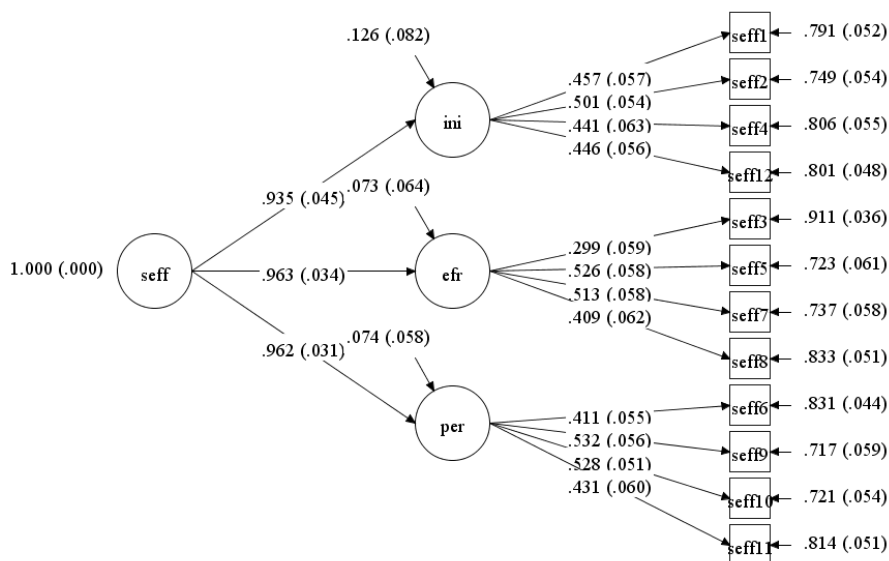


Figure 1. Higher-Order CFA Model CFA of GSES-12

Table 1. Iteration in the Parameter Estimation of the CFA Higher-Order Model of GSES-12

Iteration	PSR	Iteration	PSR	Iteration	PSR	Iteration	PSR	Iteration	PSR
100	3.959	4500	1.306	9000	1.124	13500	1.050*	18000	1.041*
500	2.151	5000	1.440	9500	1.121	14000	1.047*	18500	1.040*
1000	1.687	5500	1.198	10000	1.089*	14500	1.051*	19000	1.051*
1500	1.716	6000	1.118	10500	1.080*	15000	1.080*	19500	1.048*
2000	1.905	6500	1.154	11000	1.085*	15500	1.059*	19600	1.048*
2500	1.430	7000	1.155	11500	1.084*	16000	1.062*	19700	1.051*
3000	1.323	7500	1.124	12000	1.072*	16500	1.070*	19800	1.049*
3500	1.195	8000	1.088*	12500	1.051*	17000	1.072*	19900	1.045*
4000	1.232	8500	1.116	13000	1.070*	17500	1.056*	20000	1.039*

*convergence

If the model tested with the Bayesian CFA method is not correctly specified, the most common thing is iterations can be to tens of thousands, but the model is not convergent. The results of the analysis with the Bayesian Approach generally do not report the results of parameter estimation that are not convergent. Thus, in this research, it can be seen that the tested model has been correctly specified. Then, the researchers looked at whether the item measures the factors to be measured significantly and simultaneously determines whether the item needs to be dropped or not. The test was conducted by looking at the value of Est./S.E. for each factor load coefficient, as in Table 2.

In Table 2, it is clear that the statistics of all items are significant, and there is no negative direction so that all valid items measure self-efficacy as theorized. It means the higher-order fit model with the data is in accordance with the hypothesis that there are three dimensions of self-efficacy at the second-order level which are tested for unidimensionality and proven to be fit to the item level as evidenced by the significance of all statistics for each parameter

Then, as stated earlier about how we see the fit picture of each parameter in the model, Figure 2 clearly present the prior distribution of each parameter and also the trace plot. As can be seen in the trace plot of

each item (see Figure 2), it can be seen that the parameter estimation performed on the CFA higher-order model employs a convergent Bayesian approach, meaning that the estimation made has generated results that can be accepted and interpreted because the model is correctly specified. This can be seen in the form of a trace plot known as "good mixing", where the analysis conducted is convergent without experiencing an autocorrelation disorder that exceeds the limit. Thus, the posterior distribution for each item illustrated through the process is presented in Figure 3.

Based on Figure 3, the posterior distribution of each item has been created in a form that follows the normal curve. It is what causes the compatibility of the CFA model to be very good where Bayesian CFA is very optimally used in estimating the CFA model on the GSES-12 measuring instrument. Generally, if we test the construct validity with the CFA method without the Bayesian approach and when we find an invalid item then we retry the analysis with Bayesian CFA, the posterior distribution generated will be non-optimal like positive or negative skewed. Thereupon, it can be concluded that all GSES-12 items adapted to the Indonesian language have been shown to have very good features based on the analysis carried out using the Bayesian CFA method.

Table 2. Analysis Results of Higher-order CFA model

Parameter	Est.	Posterior S.D.	One-tailed P-value	95% C.I.		Sig.
				Lower 2.5%	Upper 2.5%	
Initiative						
Item 1	0.457	0.057	0.000	0.348	0.571	✓
Item 2	0.501	0.054	0.000	0.391	0.596	✓
Item 4	0.441	0.063	0.000	0.316	0.552	✓
Item 12	0.446	0.056	0.000	0.322	0.542	✓
Effort						
Item 3	0.299	0.059	0.000	0.187	0.415	✓
Item 5	0.526	0.058	0.000	0.416	0.636	✓
Item 7	0.513	0.058	0.000	0.380	0.613	✓
Item 8	0.409	0.062	0.000	0.285	0.530	✓
Persistence						
Item 6	0.411	0.055	0.000	0.295	0.509	✓
Item 9	0.532	0.056	0.000	0.413	0.630	✓
Item 10	0.528	0.051	0.000	0.423	0.621	✓
Item 11	0.431	0.060	0.000	0.300	0.540	✓
Self-efficacy						
Initiative	0.935	0.045	0.000	0.828	0.995	✓
Effort	0.963	0.034	0.000	0.870	0.998	✓
Persistence	0.962	0.031	0.000	0.884	0.998	✓

Comparison to Previous Research

As previously explained, the previous study using the same measurement tool is the study of Putra and Tresniasari (2015). The analysis results using Bayesian CFA in this study were compared to that study. The data from the study using the classical approach with the Robust Maximum Likelihood (MLR) estimation method were re-analyzed using the Bayesian CFA method aiming to compare the results of this study. Likewise, the data of this study were also analyzed with traditional CFA as additional information and comparison. The results of the re-analysis of the two studies are summarized in Table 3.

The comparison results that can be seen in Table 3 provide at least a variety of important information, including that when traditional CFA is used, we focus on the available index of goodness of fit as two of the most common, which are, χ^2 and the RMSEA. In the case of Putra and Tresniasari (2015) research, when faced with a condition where the CFA model has not been fit for example in the higher-order model ($\chi^2=160.468$, p-value = 0.000 and RMSEA = 0.086), then the modification of the model will be done, in which what is generally done is by freeing the error correlation between indicators to correlate with each other. However, when the same data are analyzed by

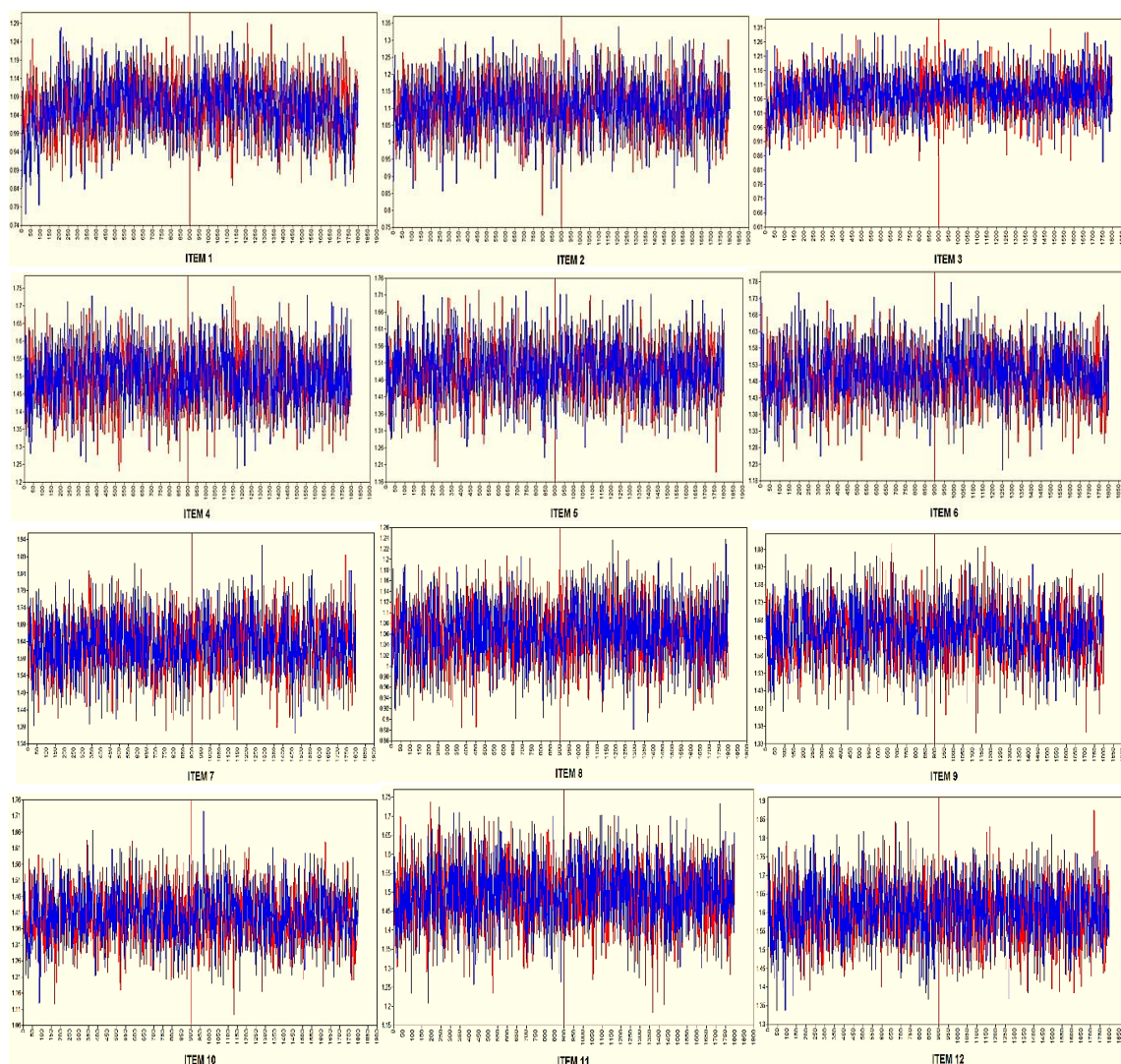


Figure 2. Trace Plot of Each Item of GSES-12

Bayesian CFA, the PPP value is still far from the expected value, but there is no option to free the error correlation to correlate as no modification index is available (Sorbom, 1989). That is why when the model has been specified and will be tested with the Bayesian CFA, it requires an in-depth examination of the model so that there is no specification error.

The research results also show that improvements to the GSES-12 adaptation process into the Indonesian language compared to previous research by Putra and Tresniasari (2015) indicated far better re-

sults. It can be seen when comparing the models and approaches used. The measuring instrument adapted in this research always produces better results. Classic and Bayesian approaches also produce results that are in line with PPP-values, and BRMSEA which is the latest development in the Bayesian SEM field shows results that are in line with RMSEA with the classical approach, according to what is theorized (Hoofs et al., 2018). This result also illustrates that the higher-order model in this research fits very well.

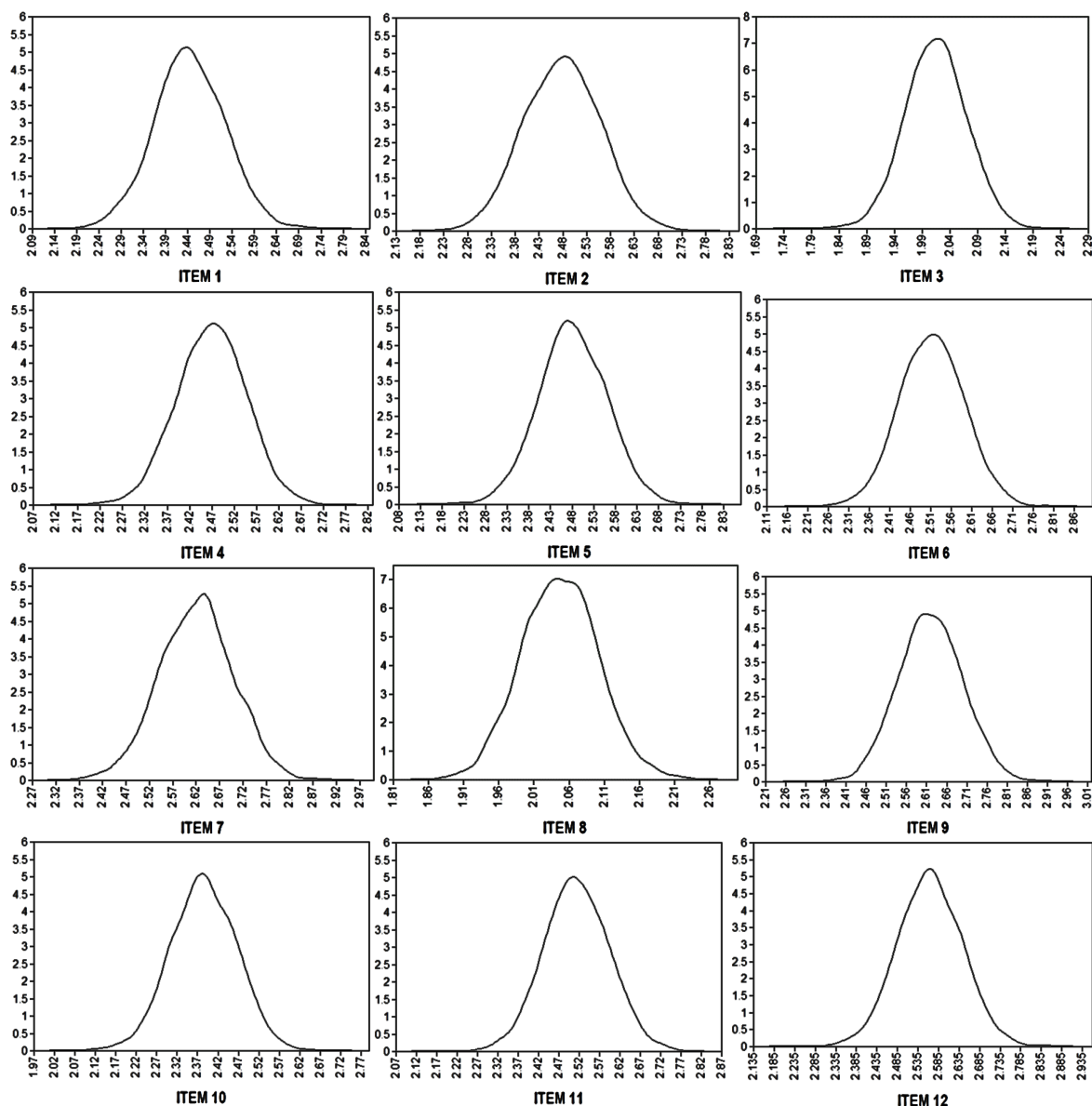


Figure 3. Posterior Distribution of Each of GSES-12 Items

This study shows that the Hastings Metropolis algorithm can work well when applied to the CFA model even though its use has not been commonly found in articles in Indonesia. It is certainly an introduction to its use which is in line with the latest developments in the field of Psychometric research (e.g., Asparouhov & Muthén, 2012; Bashkov, 2015; Cai, 2008; Cai, 2010b, 2010a; Yang & Cai, 2014). However, unfortunately, there is no comparison between the Gibbs sampler and MH used in this study even though the MH efficiency in this study is undeniable.

Regarding the structure of factors, the results of testing construct validity found that higher-order models are more suitable in describing the theoretical framework of GSES-12 compared to the first-order model. This also confirms the structure of factors from previous studies (Bosscher & Smit, 1998; Woodruff & Cashman, 1993). These findings certainly can make the measurement of self-efficacy accommodate different dimensions of other measurements to enrich theoretical understanding of measured aspects such as initiative, effort, and persistence. In practice, the use of Bayesian CFA also benefits in terms of the resulting score where the score is the best estimation of the true score known as a plausible value, so this study shows a bit more about how truly representative scores can be obtained. Therefore, this research is the starting point for more and more people in the future to

become familiar with the application of the Bayesian approach in the social science field of research.

However, an important note that can be considered is that before implementing the Bayesian CFA method, the distributional assumptions of the data need to be explored further because even if the model fits the traditional CFA, the optimal value of PPP value is quite difficult to obtain and BRMSEA is still in the initial development phase. Therefore, further studies need to be conducted on the features of BRMSEA and computation that is fairly complex requires an understanding of the prior distribution available in the Bayesian approach (e.g., informative and non-informative).

Conclusion and Suggestions

Based on the results of the construct validity test on the General Self-Efficacy Scale-12 (GSES-12) instrument using the Bayesian method of confirmatory factor analysis (CFA), it can be concluded that the research shows that the construct validity test with the second-order model fits very well. After the fit model, further information is obtained that all items are unidimensional, meaning that only measuring one factor and all items are valid in measuring self-efficacy as theorized. Comparison with the previous study shows the improvement of psychometric quality of the Indonesian version of

Table 3. Comparison of Fit Model Indices with Previous Research

Study	Model	MLR			Bayesian MH*			
		χ^2 (<i>p</i> -value)	df	RMSEA	PPP value	95% CI	BRMSEA	
Putra & Tresniasari (2015)	<i>First order</i>	<i>Baseline</i>	184.642 (0.00)	54	0.086	0.000	134.871, 208.282	0.084
		<i>Fit</i>	56.891 (0.110)	45	0.028			
	<i>Higher order</i>	<i>Baseline</i>	160.468 (0.00)	51	0.081	0.000	71.449, 133.654	0.079
		<i>Fit</i>	57.510 (0.099)	45	0.041			
This study	<i>First order</i>	<i>Baseline</i>	47.970 (0.704)	54	0.000	0.522	-33.299, 33.380	0.003
	<i>Higher order</i>	<i>Baseline</i>	43.685 (0.756)	51	0.000	0.549	-33.065, 30.422	0.001

GSES-12 items, where this measuring instrument is expected to be used in various other studies in the future. Based on the results of this research, future research is expected to be able to conduct a comparative study between the measuring tools of self-efficacy based on general self-efficacy where the measuring instruments are commonly used and tested for construct validity but the comparative studies need to be conducted to determine which measuring instruments are better used in future research. Furthermore, a measurement invariance test can also be used to obtain information about whether invariance occurs, or the applicable items are different to certain sexes or other conditions that have not been tested in this research.

References

- Asparouhov, T., & Muthén, B. (2012). Comparison of computational methods for high dimensional item factor analysis. In *Mplus technical report*. Los Angeles, CA: Muthen & Muthen.
- Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management*, 38(1), 9–44. <https://doi.org/10.1177/0149206311410606>
- Bashkov, B. M. (2015). *Examining the performance of the Metropolis-Hastings Robbins-Monro algorithm in the estimation of multilevel multidimensional IRT Models*. Unpublished doctoral dissertation, James Madison University, Harrisonburg, VA.
- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186–3191.
- Bosscher, R. J., & Smit, J. H. (1998). Confirmatory factor analysis of the General Self-Efficacy Scale. *Behaviour Research and Therapy*, 36(3), 339–343. [https://doi.org/10.1016/S0005-7967\(98\)00025-4](https://doi.org/10.1016/S0005-7967(98)00025-4)
- Brown, T. (2015). *Confirmatory factor analysis for applied research: Second edition*. New York, NY: The Guilford Press. <https://doi.org/10.1680/geot.8.B.012>
- Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill, NC.
- Cai, L. (2013). Factor analysis of tests and items. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*. Washington, DC: American Psychological Association.
- Cai, Li. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, 75(1), 33–57. <https://doi.org/10.1007/s11336-009-9136-x>
- Cai, Li. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307–335. <https://doi.org/10.3102/1076998609353115>
- Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, 4(1), 62–83. <https://doi.org/10.1177/109442810141004>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall CRC.
- Gelman, A., Meng, X. L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4), 733–807.
- Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, Ij. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the

- BRMSEA. *Educational and Psychological Measurement*, 78(4), 537–568. <https://doi.org/10.1177/0013164417709314>
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26(3), 329–367. <https://doi.org/10.1177/0049124198026003003>
- Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36(2), 109–133.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage Publications.
- Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York, NY: Guilford Press.
- Klompstra, L., Jaarsma, T., & Strömberg, A. (2018). Self-efficacy mediates the relationship between motivation and physical activity in patients with heart failure. *The Journal of Cardiovascular Nursing*, 33(3), 211–216. <https://doi.org/10.1097/JCN.0000000000000456>
- Luszczynska, A., Gutiérrez-Doña, B., & Schwarzer, R. (2005). General self-efficacy in various domains of human functioning: Evidence from five countries. *International Journal of Psychology*, 40(2), 80–89. <https://doi.org/10.1080/00207590444000041>
- Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85(4), 1–30. <https://doi.org/10.18637/jss.v085.i04>
- Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, 42(6), 779–794. <https://doi.org/10.1007/s11135-006-9067-x>
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. <https://doi.org/10.1037/a0026802>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide: Statistical analysis with latent variables* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, Linda K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4), 599–620. https://doi.org/10.1207/S15328007EM0904_8
- Putra, M. D. K., & Tresniasari, N. (2015). Pengaruh dukungan sosial dan selfefficacy terhadap orientasi masa depan pada remaja. *TAZKIYA Journal of Psychology*, 3(1), 71–82. Retrieved from <http://journal.uinjkt.ac.id/index.php/tazkiya/article/view/9194>
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). Windsor, UK: NFER-NELSON.
- Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports*, 51(2), 663–671. <https://doi.org/10.2466/pr0.1982.51.2.663>
- Sorbom, D. (1989). Model modification. *Psychometrika*, 54(3), 371–384.
- Tiyuri, A., Saberi, B., Miri, M., Shahrestanaki, E., Bayat, B., & Salehiniya, H. (2018). Research self-efficacy and its

- relationship with academic performance in postgraduate students of Tehran University of Medical Sciences in 2016. *Journal of Education and Health Promotion*, 7(1), 11. https://doi.org/10.4103/jehp.jehp_43_17
- van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, 16(2), 75–84.
- van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, 85(3), 842–860. <https://doi.org/10.1111/cdev.12169>
- van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2), 217–239. <https://doi.org/10.1037/met0000100>
- Willson-Conrad, A., & Kowalske, M. G. (2018). Using self-efficacy beliefs to understand how students in a general chemistry course approach the exam process. *Chemistry Education Research and Practice*, 19(1), 265–275. <https://doi.org/10.1039/C7RP00073A>
- Woodruff, S. L., & Cashman, J. F. (1993). Task, domain, and general efficacy: A reexamination of the self-efficacy scale. *Psychological Reports*, 72(2), 423–432. <https://doi.org/10.2466/pr0.1993.72.2.423>
- Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis–Hastings Robbins–Monro algorithm. *Journal of Educational and Behavioral Statistics*, 39(6), 550–582. <https://doi.org/10.3102/1076998614559972>

AN EVALUATION ANALYSIS OF STUDENTS' ATTITUDE TOWARDS PHYSICS LEARNING AT SENIOR HIGH SCHOOL

Dwi Agus Kurniawan
Universitas Jambi

Astalini
Universitas Jambi

Deti Kurnia Sari
Universitas Jambi


Abstract


This research is about students' attitude analysis towards Physics learning conducted at senior high schools in Batanghari Regency. The purpose of this research is to know and evaluate students' attitude toward Physics learning. This study is quantitative research employing Survey Research Design strengthened by the result of an interview to support quantitative data. This research involved 926 students of state senior high school in Batanghari. The instruments used by the researchers were in the form of questionnaires consisting of 54 items of statements by using 5-scale Likert. The results of this research show that students' attitude is at good category and quite good based on the indicators used to investigate the attitude.

Keywords: *attitude, social implication, investigation, scientific attitude, interest*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.20821>

Contact *Dwi Agus Kurniawan*

 *dwiagus.k@unja.ac.id*

 *Department of Physics Education, Faculty of Education and Teacher Training, Universitas Jambi
Kampus Pinang Masak, Jl. Raya Jambi-Ma. Bulian Km. 15, Mendalo Darat, 36361, Jambi,
Indonesia*

Introduction

Education is a pivotal activity in the learning process. Developing the students' potency so that they can think critically and creatively is one of the educational goals in Indonesia (Law of Republic of Indonesia No. 20 of 2003). Giving education maximally to the students will create high-quality students. Education is known as the process of obtaining and training the skill done by the students (Wood, 2011, p. 4). In education, learning activity is one of the important factors in influencing the result of learning.

The relation of learning implication that has achievement at three areas such as cognitive, affective, and psychomotor is not apart from the assessment which must be conducted by teachers after the learning process (Riscaputantri & Wening, 2018, p. 233). Education in Indonesia is categorized into some levels, and one of the higher educational levels in Indonesia is senior high school. In the level of senior high school, students learn a various field of knowledge; one of them is science. Physics is one subject which becomes part of science in senior high school level. Physics is considered essential to be taught as a separate subject because it gives provision of knowledge to the students. Furthermore, Kaniawati, Samsudin, Hasopa, Sutrisno, and Suhendi (2016, p. 1) state that Physics is a branch of scientific knowledge which can explain each natural phenomenon in daily life.

Physics is an abstract subject until it needs high order thinking skill which causes the students difficult in understanding Physics topics. The difficulty encountered by the students in comprehending Physics during school time causes deeper difficulty when they are at college (Guido, 2013, p. 2089). In spite of the abstract characteristic, students' internal factor like the attitude towards Physics subject is also one of the difficulties occurs in Physics learning as known that this attitude refers to the behavior and emotions of someone (Martin & Briggs, 1986, p. 2). Attitude will be seen during the learning process. Furthermore, Fatonah (2014) explains that attitude is the tendency to act,

achieve, think, and feel in facing objects, ideas, situation, and values. By knowing students' attitude in learning gives a positive effect on the improvement of learning achievement.

Nordin and Ling (2011) argue that students' attitude is the key to achieving success in the mastery and achievement improvement of the students. If the students have a positive attitude towards a certain learning process, it will give a big impact on the learning itself. Veloo, Nor, and Khalid (2015) explain that the students who have a positive attitude towards learning, it can improve their learning achievement. Likewise, students' attitude towards Physics; if the students have a positive attitude towards Physics learning, then their Physics achievement or science achievement will improve too. Students' positive attitude towards Physics occurs when the students can understand deeper about the concept of Physics and make the learning more effective in their environment. Students' negative attitude in Physics learning causes students' achievement becomes bad. Erdemir (2009) explains that good or bad Physics learning achievement is influenced by students' attitude during the learning process. If the students have a negative attitude towards Physics learning, it will make the present learning, and future learning in Physics will be more difficult. Students' attitude towards Physics can be known by identifying the social indicators of Physics, attitude towards investigation in Physics, adoption of scientific attitude, and interest to enhance the duration of learning Physics.

The social implication of Physics learning, which relates to the natural phenomenon, affects social life. As a result of science and technology development, it will generate some advantages and social impacts that will occur. It can be in the form of attitude towards social advantages, progress case, and scientific research (Welch, 2010). In the learning at school, the social implication of Physics can be seen in how students' independency and teamwork in the group. Students can be active and motivated

during the learning process if they are demanded to share information, work collaboratively in a group, and respect other people (Yance, Ramli, & Mufit, 2013). The form of students' independency can be seen from their completed assignment or homework given by the teacher and believe in their own abilities and cooperative attitude. Students' social ability will be more prominent when working collaboratively than working individually (Iwan & Sani, 2015). Beside the social implication of Physics, the attitude towards investigation in Physics also affects the learning of Physics.

The attitude towards investigation in Physics is the students' point of view or actions in solving the problem occurs in Physics. Lederman, Lederman, and Antink (2013) insist that investigation refers to the combination of a scientific process with knowledge and scientific reasoning and critical thinking. Investigation in Physics can be seen on how the students solve the problem happens in the class. Welch (2010) believes that there are some ways of solving the scientific problem; they are; measuring, identifying, and experimenting in its scientific scale to find accurate information. One of the investigation activities in Physics conducted by the students is experimenting. In addition, Moeed (2013) argues that experiments conducted by students can develop their belief during their time at school that to obtain conclusion then it requires the steps that must be followed in a scientific method to know the result of new knowledge. Students' attitude in Physics investigation helps to improve students' activeness in learning. Students' activeness in learning correlates with the scientific attitude owned by the students.

Scientific attitude is a crucial attitude to be acquired. Ali et al. (2012) explain that in the educational world, especially in the world of science, a scientific attitude is a pivotal aspect because this attitude can improve good learning achievement. By nurturing a positive scientific attitude in the students, they will experience learning achievement improvement (Chiappetta, Koballa, & Collette, 1998). In addition, by owning a sci-

entific attitude, the students can think rationally and critically. Complicated and abstract learning of Physics truly needs the students to behave scientifically. Students with high scientific attitude will have a positive attitude towards Physics. Scientific attitude can be described as the expression or reaction shown in accordance with the ethics of science (Pitafi & Farooq, 2012). The importance of having a scientific attitude for a student is to obtain knowledge from various fields of discipline (Osman, Iksan, & Halim, 2007). In order to obtain knowledge from various fields of knowledge and especially in Physics learning, it requires students' interest to extend the duration of Physics learning.

Students' interest in extending the duration of Physics learning is influenced by themselves. Students' attitude towards science learning refers to their happy feeling or interest in it (Agunbiade, Ngcoza, Jawahar, & Sewry, 2017). The students who consider Physics is difficult because they are not interested in extending the duration of learning Physics. The students who are interested in extending the time for learning Physics will influence their achievement, learning results, and career in Physics (Bybee, McCrae, & Laurie, 2009). Students' interest in extending the time for learning Physics can be seen from their achievement in the field of science or Physics. One of the factors which cause students' failure in Physics learning achievement is the lack of interest in extending the amount of time for learning Physics (Visser, 2007). Therefore, more frequent the students extend their time for learning, then the better their scores or achievements will be. Thus, students' interest in extending the amount of time for learning Physics is vital to achieve good learning results.

The purpose of this research is to know students' attitude towards Physics learning at senior high schools, especially at senior high schools in Batanghari Regency. The attitude indicators are used to identify students' attitude towards Physics learning encompassing the social implication of Physics, attitude towards investigation in Physics, adoption from scientific attitude, and the in-

terest in extending the amount of time in learning Physics. The findings of this research can contribute to improving students' attitude towards Physics learning. This research is pivotal to conducted, especially at senior high school level because attitude can influence learning results or achievement.

Research Method

The research design used in this research was survey research. This survey research design is the design which is used to collect data of questionnaires spread to some samples or to all population used to describe the attitude, opinion, behavior, or particular traits of the population (Creswell, 2012). The research samples were in the amount of 926 students. The samples were designed and determined in line with the selection of the sample through purposive sampling technique.

Research Sample

The samples of this research were five schools located at Batanghari Regency; they are State Senior High School 10 Batanghari, State Senior High School 5 Batanghari, State Senior High School 8 Batanghari, State Senior High School 1 Batanghari, and State Senior High School 6 Batanghari. Total samples consisted of grade X, XI, and XII at each school. The total number of male students were 353 people (38.12%); the total number of female students were 573 people (61.7%).

Instruments and Procedures

Instruments and procedures in this research were in the form of questionnaires and interview. The indicators used in this research can be seen in Table 1.

Table 1. Indicators of Students' Attitude Questionnaire in Learning Physics

No	Attitude Indicators
1	Social Implication of Physics
2	Attitude Towards Investigation in Physics
3	Adoption of Scientific Attitude
4	The Interest of Extending The Amount of Time for Learning Physics

The questionnaires contained four indicators. Then, during the interview, the question items are arranged referring to those indicators. The questionnaires instrument contained 54 items by using 5-scale Likert (consisting of: 1-really disagree, 2-disagree, 3-neutral, 4-agree, and 5-really agree). The questionnaires were given to 926 students. The interview instrument was in the form of questions which were given to 35 students.

Data Analysis

Data analysis employed in this research was quantitative data analysis by using SPSS program to find out descriptive statistics. Data in this research used quantitative data analysis by using SPSS program to find out descriptive statistics. Descriptive statistics is a description or presentation of data in a big number, consisting of modus, mean, median, minimum, maximum, and standard of deviation (Cohen, Manion, & Morrison, 2007). Then, the analysis was continued with an interview for strengthening the results of quantitative data.

Findings and Discussion

Findings

Social Implication of Physics

Table 2. Social Implication of Physics

Range	Classification		%
	Attitude	Total	
5 – 8	Very Bad	1	0.1
9 – 12	Bad	19	2.1
13 – 16	Fair	218	23.5
17 – 20	Good	475	51.3
21 – 25	Very Good	213	23.0
Total		926	100%

The social implication indicator of state senior high school students towards Physics is described in Table 2. In Table 2, there are 54.3% of students in the good category with a maximal score from the whole statements at indicator 1 is 25. It shows that most of the students admit that there is a social implication of Physics towards their

social life. Then, 23.5% of the students are in the fair category, which means the students are still confused about the existence of Physics science role in advance technology. Meanwhile, 2.2% students are at bad category which shows that they do not understand about the existence of good implication of Physics in their social life.

Attitude towards Investigation in Physics

The results of data analysis on the attitude towards investigation in Physics can be seen in Table 3. Table 3 also shows that the attitude towards investigation in Physics categorized into very bad is 2%. Meanwhile, the attitude at bad category is 3.7%, and attitude at fair category is 45%. Then, 46% is in a very good category, and the last very bad attitude is at 5.1%. The attitude towards investigation in Physics has a better attitude in learning Physics compared to other categories.

Table 3. Attitude towards Investigation in Physics

Range	Classification		%
	Attitude	Total	
9 – 16.2	Very Bad	2	2 %
16.3–23.5	Bad	34	3.7 %
23.6–30.8	Fair	417	45 %
30.9–38.1	Good	426	46 %
38.2–45.5	Very Good	47	5.1 %
Total		926	100%

Adoption of Scientific Attitude

Table 4. Adoption of Scientific Attitude

Range	Classification		%
	Attitude	Total	
7 - 12.6	Very Bad	1	0.1%
12.7-18.3	Bad	5	0.5%
18.4 - 24	Fair	338	36.5%
25- 30.6	Good	519	56%
30.7- 36.3	Very Bad	63	6.8%
Total		926	100%

The results of score data analysis on scientific attitude is presented in Table 4. In Table 4, it can be seen the data of questionnaires which have been processed by using SPSS program obtained from 926 respond-

ents from senior high school educational level which more focus on the indicator adoption of scientific attitude. The results show that the most dominant category is good (56%), then followed by fair (36.5%), very good (6.8%), bad (0.5%), and the least is really bad (0.1%).

The Interest to Extend the Duration of Learning Physics

The results of score data analysis on the interest to extend the amount of time for learning Physics is presented in Table 5. Based on Table 5, it shows that the interest to extend the amount of time for learning Physics at very good category is 3.9%, while at good category is 22%, the fair category is 57.9%, bad category 14.3%, and the last very bad category is 1.9%. The most dominant attitude at category the interest to extend the amount of time for learning Physics is a fair attitude.

Table 5. The Interest to Extend the Amount of Time for Learning Physics

Range	Classification		%
	Attitude	Total	
8 – 14.4	Very Bad	18	1.9%
14.5–20.9	Bad	132	14.3%
30 – 36.4	Fair	536	57.9%
36.5– 42.9	Good	204	22%
43 – 49.4	Very Good	36	3.9%
Total		926	100%

The Problem Faced in Improving Students' Attitude in Batanghari Regency

Table 6. The Problem Faced in Improving Students' Attitude in Batanghari Regency

Indicators	Respondents (926)
Social implication of Physics	2.2% (20)
Attitude towards investigation in Physics	3.9% (36)
Adoption of scientific attitude	0.6% (6)
The interest in extending the amount of time for learning Physics	16.2% (150)

The problem faced in improving students' attitude in Batanghari Regency is obtained at the indicator of Physics social implication, the normality of science, happiness in learning Physics and the interest to extend the time for learning Physics. These can be seen in Table 6.

From the results that have been obtained from the respondents who had filled in the questionnaires spread by the researchers, the researchers obtained four obstacles encountered by the students related to their attitude towards Physics learning (Table 6): Social Implication of Physics (2.2%), Attitude towards investigation in Physics (3.9%), Adoption of Scientific Attitude (0.6%), Interest in Expanding Time for Learning Physics (16.2%).

Discussion

Attitude comes from someone's feeling towards an object which is reflected in the feeling of like or dislike. Attitude can be observed in the learning process, perseverance, and also consistency towards an object (Basuki & Hariyanto, 2014). Good learning result is influenced by students' positive attitude. Veloo et al. (2015) believe that the students who have a positive attitude towards learning can improve their achievement. Complicated Physics learning causes the students to have less interest in learning Physics. Yara (2009) states "The attitude towards science or Physics shows interest or feeling towards knowledge, in which the feeling meant here is the disposition of the students towards like or dislike science." Therefore, noticing students' attitude during the learning process can increase either the learning result or achievement of students. The following is the exploration of indicators used to recognize students' attitude towards Physics.

Social Implication of Physics

The results of questionnaires' data analysis on indicator social implication of Physics at senior high school in Batanghari Regency, it shows dominant students at the good category with a percentage in the a-

mount of 51.3%. The results of the interview show that although Physics is complicated, they realize that the concept and formula of Physics can be applied in daily lives. The concept and principles of Physics are mostly applied in life and contribute a lot in life nowadays (Veloo et al., 2015). The students who respect the roles of Physics in daily life are the students with good achievement in Physics at Senior High School, talented in science and Mathematics. Kaniawati et al. (2016) explain the concept which has important roles in learning as the foundation in learning the natural phenomenon. Until in nurturing the concept, it is better to correlate the learning to daily life problems so that the students recognize the importance of learning Physics.

Attitude towards Investigation in Physics

The results of data analysis on indicator the attitude towards the investigation in Physics at senior high school in Batanghari Regency show students' dominant attitude is at category fair with the percentage of 45%. Based on the results of the interview, the students at good category have an active attitude in finding the things opposed to the results of the experiment then the students respond critically, having high curiosity, and never give up. The students who have a good attitude like experimental activities which indicates that they love to think critically, finding new interesting things in Physics through investigation they conduct. Civelek, Ucar, Ustunel, and Aydin (2014, p. 566) explain that one of the hindrances in Physics learning process is the students lack scientific thinking skill towards the science of Physics based on the concepts in learning abstract things.

By experimenting, the abstract science of Physics becomes easier to comprehended and liked by the students. Moeed (2013, p. 539) states that, through experiment, the students can develop their ability to think critically and to obtain the conclusion, step by step is required and must follow the scientific method. Experimental activities conducted can increase the activeness and self-

confidence of the students either in experimenting or in learning activities. Stefan and Ciomos (2010, p. 8) assert that the improvement of investigation activity is influenced by self-confidence towards their own ability (the students) in learning science. This experimental activity can also improve students' curiosity; it can be seen through the students love to ask and search for a solution when doing the investigation. The form of appreciation and support for students' scientific investigation is by showing them respect scientifically by collecting, thinking creatively, thinking rationally, responding critically, communicating, and taking conclusion because they face life situation related to science (Bybee et al., 2009).

Adoption of Scientific Attitude

Based on the results of data analysis of the questionnaires at senior high school in Batanghari Regency on scientific attitude, the most dominant attitude is at the good category with the percentage in the amount of 56%. Thus, it can be said generally that students' scientific attitude at senior high school in Batanghari Regency is categorized into positive. The results of the interview on the students categorized into the good category. It is recognized that they like different opinions in the class. If there is a different opinion in solving the problem, for instance, in conducting discussion, they can obtain some solution to overcome it until the students do not only focus on one solution only.

The students who have a good attitude towards scientific attitude can think objectively. Olasehinde and Olatoye (2014) stress that scientific attitude is the ability to behave consistently, rationally, and objectively in solving the problem. And the students' willingness to solve the problem occurs in the learning process can improve scientific attitude. Osborne, Shirley, and Collins (2013) assert that "Scientific attitude is the desire to know and understand, searching for verification and questioning in science." The students who have dominant attitude seen during the learning process becomes the

center of attention of the teachers and becomes more active. Mukhopadhyay (2014) explains that scientific attitude is one of the main fields which becomes the attention of a teacher in a class situation in general. A good scientific attitude will help the students to improve their learning achievement. Further, Olasehinde and Olatoye (2014) add that the better the students' scientific achievement, the better their attitude towards science.

The Interest in Extending the Time for Learning Physics

From the results of data analysis on the questionnaires about the indicator of the interest in extending the time for learning Physics at senior high school in Batanghari Regency, it is clear that students' dominant attitude is at the fair category with percentage 57%. The category of students' attitude at that indicator proves that only a part of students is interested in extending the time for learning Physics. Based on the results of the interview on the students at the fair category, it is recognized that when they finished school time, they learn Physics at home individually or in a group and ask their friends about the topics that they do not understand yet.

In order to improve students interest in extending the time for learning Physics, it requires students' positive attitude and students' love for the learning process. Osman et al. (2007) state that the students will feel interested and happy in learning a certain subject if they like that subject. The interest in spending time for learning Physics is important because the interest or talent in learning Physics can make the students serious in learning Physics. Students' interest in science learning at school is pivotal in continuing education to the next level. Highly positive attitude towards the interest in extending the time for learning Physics will influence learning achievement. Bybee et al. (2009) argue that the students who are interested in extending the time for learning Physics will affect their achievement and learning results in Physics.

Hindrances

Based on the results of data analysis, there are some hindrances on each indicator. At indicator social implication of Physics, the results of analysis obtained are; 22 students or 2.2% out of 926 students have attitude at bad category. Students' attitude at bad category is because they do not think that Physics is something complicated and abstract until it is difficult for them to solve the problem related to Physics. At the attitude towards the investigation in Physics, there are 36 students at percentage 3.9% out of 926 students in bad attitude category. The hindrance encountered by the students is that they do not know how to conduct an experiment and less active when the activity is happening. Their less activity and incapability of experimenting is because they do not like Physics and consider Physics as a difficult subject.

At the indicator of adoption of scientific attitude, six students with a percentage of 0.6% out of 926 students were categorized having a bad attitude. This bad attitude is seen in learning; the students lack curiosity and thinking critically. On the indicator of the interest in extending the time for learning Physics, 177 students or 16.2% out of 926 students have a bad category. The hindrance occurred in this indicator is because the students do not really understand the topics, and they consider that Physics is difficult until they do not have interest in extending the time for learning Physics.

Conclusion

Based on the results of the research, it can be concluded that the indicator social implication of Physics and adoption of scientific attitude have good results. Meanwhile, the attitude towards investigation in Physics and interest to extend the time for learning Physics have sufficient results. The average students have a good attitude on the indicator of social implication and adoption of scientific attitude. It is because there is a relevant correlation between these two indicators. If the students feel the implication of

Physics towards their social life, then they will have a high scientific attitude. However, in general, not all students feel the implication of Physics in their social life, as a consequence, they do not like the life pattern of a scientist who always conducts investigation towards Physics, and they are not interested in extending the time for learning Physics.

Therefore, the teachers need to recognize how students behave during the learning process and fix the learning design in the class in line with students' ability. Based on the explanation on the hindrance that occurred on the indicators of students' attitude, it can be seen that the students have a bad attitude because the model or the teaching strategies of the teacher cannot improve students' positive attitude towards Physics. The learning method which provides an integrated learning environment with laboratory measurement can help the students to solve the problem in Physics and increase their attitude to be more critical. Beside the teaching method which can increase students positive attitude, the teacher must conduct learning by using science skill. The students who have science skill will cause them to have a positive attitude towards science. Improving students' scientific attitude also can increase students' positive attitude towards the learning of science or Physics.

References

- Agunbiade, E., Ngcoza, K., Jawahar, K., & Sewry, J. (2017). An exploratory study of the relationship between learners' attitudes towards learning science and characteristics of an afterschool science club. *African Journal of Research in Mathematics, Science and Technology Education*, 21(3), 271–281. <https://doi.org/10.1080/18117295.2017.1369274>
- Ali, K., Shafqat, Shah, A., Makhdoom, S., Mahmood, Z., & Zareen, R. (2012). Scientific attitude development at secondary school level: A comparison between methods of teaching. *Language in India*, 12(9), 439–454.

- Basuki, I., & Hariyanto. (2014). *Asesmen belajar*. Bandung: Remaja Rosdakarya.
- Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching*, 46(8), 865–883. <https://doi.org/10.1002/tea.20333>
- Chiappetta, E., Koballa, T., & Collette, A. (1998). *Science instruction in the middle and secondary schools* (Prentice H). Upper Saddle River, NJ.
- Civelek, T., Ucar, E., Ustunel, H., & Aydin, M. K. (2014). Effects of a haptic augmented simulation on K-12 students' achievement and their attitudes towards physics. *Eurasia Journal of Mathematics, Science and Technology Education*, 10(6), 565–574. <https://doi.org/10.12973/eurasia.2014.1122a>
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London and New York, NY: Routledge Falmer.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.
- Erdemir, N. (2009). Determining students' attitude towards physics through problem-solving strategy. *Asia-Pacific Forum on Science Learning and Teaching*, 10(2), 1–19.
- Fatonah, S. (2014). *Pembelajaran sains*. Yogyakarta: Ombak.
- Guido, R. M. (2013). Attitude and motivation towards learning physics. *International Journal of Engineering Research & Technology*, 2(11), 2087–2094. <https://doi.org/10.1093/nar/gkn1085>
- Iwan, N., & Sani, R. A. (2015). Efek model pembelajaran kooperatif tipe group investigation dan teamwork skills terhadap hasil belajar fisika. *Jurnal Pendidikan Fisika*, 4(1), 3.
- Kaniawati, I., Samsudin, A., Hasopa, Y., Sutrisno, A. D., & Suhendi, E. (2016). The influence of using momentum and impulse computer simulation to senior high school students' concept mastery. *Journal of Physics: Conference Series*, 739(1), 1–4. <https://doi.org/10.1088/1742-6596/739/1/012060>
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).
- Lederman, N. G., Lederman, J. S., & Antink, A. (2013). Nature of science and scientific inquiry as contexts for the learning of science and achievement of scientific literacy. *International Journal of Education in Mathematics, Science and Technology*, 1(3), 138–147.
- Martin, B., & Briggs, L. J. (1986). *The affective and cognitive domains: Integration for instruction and research*. Englewood Cliffs, NJ: Educational Technology.
- Moeed, A. (2013). Science investigation that best supports student learning: Teachers understanding of science investigation. *International Journal of Environmental and Science Education*, 8(4), 537–559. <https://doi.org/10.12973/ijese.2013.218a>
- Mukhopadhyay, R. (2014). Scientific attitude – Some psychometric considerations. *Iosr Journal Of Humanities And Social Science (Iosr-Jhss) Osr-Jhss*, 19, 98–100.
- Nordin, A., & Ling, L. H. (2011). Hubungan sikap terhadap mata pelajaran sains dengan penguasaan konsep asas sains pelajar tingkatan dua. *Journal of Science & Mathematics Educational*, 2(June), 89–101.
- Olasehinde, K. J., & Olatoye, R. A. (2014). Scientific attitude, attitude to science and science achievement of senior secondary school students in Katsina State, Nigeria. *Journal of Educational and Social Research*, 4(1), 445–452. <https://doi.org/10.5901/jesr.2014.v4.n1p445>

- Osborne, J., Shirley, S., & Collins, S. (2013). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 37–41. <https://doi.org/10.1080/0950069032000032199>
- Osman, K., Iksan, Z. H., & Halim, L. (2007). Sikap terhadap sains dan sikap saintifik di kalangan pelajar sains. *Saintifik Jurnal Pendidikan*, 32, 39–60.
- Pitafi, A. I., & Farooq, M. (2012). Measurement of scientific attitude of secondary school students in Pakistan. *Academic Rearch International*, 2(2), 379–392. <https://doi.org/10.1002/mrm.24433>
- Riscaputantri, A., & Wening, S. (2018). Pengembangan instrumen penilaian afektif siswa kelas IV sekolah dasar di Kabupaten Klaten. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 22(2), 231–242. <https://doi.org/10.21831/pep.v22i2.16885>
- Stefan, M., & Ciomos, F. (2010). The 8th and 9th grade students' attitude towards teaching and learning physics. *Acta Didactica Napocensia*, 3(3), 7–14.
- Veloo, A., Nor, R., & Khalid, R. (2015). Attitude towards physics and additional mathematics achievement towards physics achievement. *International Education Studies*, 8(3), 35–43. <https://doi.org/10.5539/ies.v8n3p35>
- Visser, Y. L. (2007). *Convergence and divergence in children's attitudes toward the sciences and science education*. Boca Raton, FL: Learning Development Institute, Florida Atlantic University.
- Welch, A. G. (2010). Using the TOSRA to assess high school students' attitudes toward science after competing in the first robotics competition: An exploratory study. *Eurasia Journal of Mathematics, Science and Technology Education*, 6(3), 187–197. <https://doi.org/10.12973/ejmste/75239>
- Wood, K. (2011). *Education is basic*. New York, NY: Taylor & Francis.
- Yance, R. D., Ramli, E., & Mufit, F. (2013). Pengaruh penerapan model project based learning (PBL) terhadap hasil belajar fisika siswa kelas XI IPA SMA Negeri 1 Batipuh Kabupaten Tanah Datar. *Pillar of Physics Education*, 1(1), 48–54. Retrieved from <http://ejournal.unp.ac.id/students/index.php/pfis/article/view/490/279>
- Yara, P. O. (2009). Relationship between teachers' attitude and students' academic achievement in mathematics in some selected senior secondary schools in southwestern Nigeria. *European Journal of Social Sciences*, 11(3), 364–369.

AN EDUCATIONAL-EVALUATION STUDY FOR STREET CHILDREN IN RUMAH IMPIAN FOUNDATION

Aulia Ninda Haryoni

Graduate School, Universitas Negeri Yogyakarta

Istiana Hermawati

Center for Research and Development of Social Welfare Service, Ministry of Social Affairs


Abstract


This research aims to evaluate the education of street children at Rumah Impian Foundation in Yogyakarta. This is an evaluation research using a responsive evaluation model with a naturalistic qualitative approach. The evaluation stages carried out in this study consist of Rational, Antecedent, Transaction, and Outcome. The sampling technique used was purposive sampling with snowball sampling because the data were obtained from the community of the subjective sample, or in other words, the sample used is very rare and is grouped in a set. The data were analyzed using qualitative analysis techniques by Milles and Huberman. The results of the study show that, at the rational stage, the background of the educational concern for street children starts with a sense of caring about the street children's future that is worth fighting for. In the antecedent stage, there is a conformity input between the volunteer, apprenticeship membership, and the street children's education needs, in the form of policies and recruiting volunteers to be a companion to street children. At the transaction stage, the process between the foundation and street children education is appropriate, meaning that the foundation has facilitated the education needed by street children both formally and non-formally, in the form of increasing their skills through courses. Thus, there is a match between what was done in the previous stages and its results, showing that there are street children who reach their dreams, as a result of the foundation's efforts to continue knitting their dreams through intensive activities and assistance.

Keywords: *street children, responsive evaluation, Rumah Impian Foundation*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.22573>

Contact *Aulia Ninda Haryoni*

 *auliabaryoni13@gmail.com*

 *Department of Educational Research and Evaluation, Graduate School of Universitas Negeri Yogyakarta*

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

Introduction

Education is a future investment. It is always seen as something useful for the advancement of humanity. Indeed, education is an important reference for the advancement of national civilization. The importance of education is seen as the basis for growing the future of humankind, from children to adulthood. It is, of course, inseparable from the meaning of education according to the father of education in Indonesia, that is, Ki Hadjar Dewantara, which refers to the effort to advance the character, mind, and body of the child, to advance the perfection of life, which is, living in harmony with nature and society (Neolaka & Neolaka, 2017).

In line with the meaning of education above, education itself can be used as a foundation to obtain noble values that can make children develop according to their abilities and also the skills they should get. Education will give meaning to life for children to achieve their dreams, ideas, and hopes. Education is very broad, and can be obtained anywhere, not always obtained from the school. In practice, education has its own axiological foundation, which is described in Figure 1.

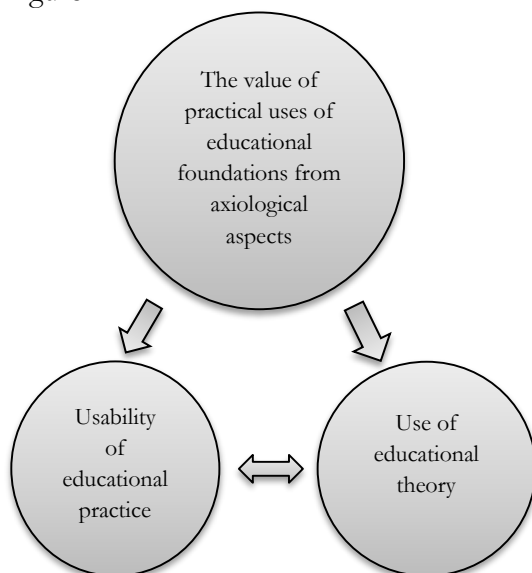


Figure 1. The Foundation of the Educational System Axiology (Neolaka & Neolaka, 2017)

In other words, education must be able to determine the values that will be used in the education process, both in theory and practice. The values in education will then affect how the learning process and learning atmosphere of the students themselves. As explained in the Act, education is a conscious and planned effort to create a learning atmosphere and learning process so that students actively develop their potential to possess the power of spiritually religious, self-control, personality, intelligence, noble character, and the skills needed by each person and the community (Law of Republic of Indonesia No. 20 of 2003).

Departing from the educational foundation, it is known that the right to obtain education has been stated in Law No. 31 paragraph 1 which says "every citizen has the right to education" (Article 31 of the 1945 Constitution). It means that all people, including children who are Indonesian citizens, have the right to procure education. However, in reality, this is not the case. Not all citizens, especially children, get these rights, one of which is street children.

Education polemic for all cannot be felt fully by some people, especially street children. In this case, education for street children has also been regulated in the Child Protection Act, which also regulates the rights of children. Law No. 23 of 2002 concerning child protection states that a child, someone who has not reached the age of 18 years, has the right to obtain education and teaching in the context of his personal development and the level of intelligence according to his interests and talents. It is compulsory education, in which the government gives room for all children, but there is no penalty when they do not go to school, might be due to the economic problems. Then, the paragraph also explains how children have the right to obtain protection in the education unit from acts of crime and violence committed by educators, education personnel, fellow students, and or other parties, so that it can be seen that education is a right that must be given to children, including street children.

Street children are a social phenomenon that cannot be separated from the factor of social value, which takes effect in the community. The phenomenon of street children is often triggered by economic conditions (Ray, 2017). Before discussing further in this article, street children have their own description. Not that all children who are on the streets are street children, but, deeper than that, Lusk (1992) developed four categories of children found on the road. A research conducted in 2012 in Semarang gave an explanation of the characteristics of street children, including: (1) being in public places (roads, markets, intersections, shops, and entertainment venues) for three to 24 hours per day; (2) having low education (most of them drop out of school or are too late for school age); (3) coming from families with low or poor economy condition (there are some who are not clearly family members and come from urban groups) (Rochatun, 2015).

Each group has its own psychological characteristics. Each characteristic is elaborated as follows. First, some poor children work and return to their families at night. They tend to go to school regularly and are not naughty. Second, there are some independent street workers whose family relationships are not harmonious, their school activities have decreased, and their delinquency has tended to increase. Third, there are children from street families who live and work with their families on the streets. Their condition is tied to poverty. In India, they are referred to as sidewalk dwellers (Patel, 1983), while in the United States, they are children from homeless families. And fourth, there are children who break up with their families. They are on the streets full time and are "real" street children (Aptekar & Stoecklin, 2014).

Living as a street child does not mean that a child chooses this without any reason. Viewed from this category, the condition that causes children to be more interested in taking to the streets is the condition of their family and economy. However, it does not mean that they forget about education for

themselves. Their right to get a decent life in their environmental community has also been regulated in the Law, protection of them has also been regulated in it. However, we cannot prevent children from taking to the streets just like that.

Such phenomenon of street children has developed in many regions throughout Indonesia. Big cities are the center of growth for street children, which makes their mobility very high. The existence of street children in big cities that are turning into metropolitan cities is certainly not a strange view. According to Malia (2016), 16 provinces are targeted by street children, including, North Sumatra, West Sumatra, Riau Islands, South Sumatra, Lampung, Jakarta, Banten, West Java, Central Java, Yogyakarta, East Java, East Nusa Tenggara, West Nusa Tenggara, South Kalimantan, West Kalimantan, and South Sulawesi. The data are obtained from data on the number of street children handled by the Ministry of Social Affairs of the Republic of Indonesia. Out of those provinces, Yogyakarta is one of the provinces prone to the growth of street children. The data on the number of street children in Yogyakarta handled until 2017 is presented in Table 1.

Table 1. Number of Street Children in Yogyakarta

Years	Number of Street Children
2013	212
2014	220
2015	219
2016	327
2017	348

Source: (Regional Development Planning Agency, 2018)

The number of street children which tend to increase will certainly become a vital urgency to discuss. It certainly concerns the fate of education that can accommodate the collection of children. We know Yogyakarta as a city of student because many schools are growing and developing well in this city, in addition, Yogyakarta is also a destination for the wider community to take their education

both from within or outside Yogyakarta area itself (Goenawan & Harnoko, 1993). However, it would be ironic if education for street children was not taken into account even though Yogyakarta is one of the provinces with a steady growth in street children.

Education for street children is still seen as a formal matter, which means that when children take part in learning activities at school, they are considered to have already received education, but it cannot ascertain how they will stop taking to the streets to complain about their fate. More than that, the condition of street children is different from children in general who still have intact, caring, and always available families to take care of their development. Street children have their own views about how they live their lives and try to survive amid the difficulties of their families and economic conditions (Ray, 2017).

Treating street children is not just about sending them to the 12-year compulsory education level, but more than that, it is about giving a value of life that is beneficial to the wider community. Of course, this is in line with the meaning and bases of education that was proposed by Ki Hadjar Dewantoro before. Education does not only come from formal schools but can also come from many sources. One of them is humanizing education like the one at Rumah Impian Foundation, Yogyakarta. Thus, this research evaluates how education for street children is carried out at the Rumah Impian Foundation and provides appropriate recommendations in order to reduce the number of street children in Yogyakarta, and broadly throughout Indonesia.

Research Method

This study is an evaluation research using a responsive evaluation model with a qualitative approach which was developed by Stake (Stake, 2005). This evaluation is called a client-centered evaluation. According to Stake, evaluation is called a response if it meets three criteria: (1) more oriented directly to program activities, than program objectives; (2) respond to the requirements

of the audience's information needs; and (3) perspectives of different values from people served are reported in the success and failure of the program. This model was used because the researchers want to understand more in the activities of implementing education for street children (Stufflebeam, Madaus, & Kellaghan, 2000). Furthermore, the information needed in this evaluation model is rational, antecedent, transaction, and outcome. Rational is intended to find out the background of the philosophy and the basic purpose of a program. The antecedent is used to see the input or initial conditions of the program. Transaction regarding the process was carried out by Rumah Impian Foundation in conducting education for street children, while the outcome explains the results of the program (Worthen & Sanders, 1973).

The principle used in processing data from the evaluation results is the descriptive naturalistic principle by finding the contingency or possibility that can occur as well as the congruency or conformity with the expected goals. Conformity was obtained by comparing the results of interviews, study documents, and observations in the field. Data collection was done using direct observation, in-depth interviews, and document studies.

The sampling technique used in this research was purposive sampling with snowball sampling, in which the researchers interviewed the chairperson of the foundation as the primary informant, one staff member who also works at the Rumah Impian Foundation, and two street children. Besides, the researchers also examined the documents that were used as a reference for implementing the educational services program, such as the certificates of legality and letters of cooperation with other parties. The research took place at Rumah Impian Foundation Yogyakarta, located in Purwomartani, Sleman, Yogyakarta. In data analysis, the techniques used were qualitative analysis, data collection, data reduction, data presentation, and also conclusion drawing (Miles & Huberman, 1994).

Findings and Discussion

General Description of the Location

Rumah Impian Foundation is a non-profit foundation that is engaged in the social sector that is concerned with the fate of street children, especially in Yogyakarta. The Rumah Impian Foundation began its approach to street children in 1999. In 1999, the Foundation had not yet been established but still as a movement to care for street children in Yogyakarta containing students who were studying in Yogyakarta and named Rumah Impian (Source: Interview data, 12 December 2018).

Developing the potential of street children is certainly not easy, so following the development, the community of the movement began to expand in 2006 and formed a foundation named Rumah Impian Foundation. The choice of the name is due to the existence of regulation from the government that the name of an established foundation must use an Indonesian-termed name, and if a foreign language is used, it must have an acronym or an abbreviation in Indonesian. In 2006, the foundation first owned a half-way house located in Jetis Yogyakarta.

In 2008 the foundation continued to grow to have several shelters to accommodate street children. In the process of its founding, many obstacles were experienced, such as rejection from the communities at some areas in Yogyakarta. However, the Rumah Impian Foundation itself has a belief that every street child has equal rights in all fields, especially in the field of education. They have the right to reach their aspirations, and they need the right support to live an independent life in achieving their aspirations, being responsible for themselves and having a caring attitude towards others and their environment. The activities done in the foundation is presented in Figure 2.

In approaching street children, the Rumah Impian Foundation continues to develop holistic methods to assist children at risk. Funds used by the foundation in assisting the implementation of mentoring activities come from regular donors, as well as general donors spread across several regions

in Indonesia. The children at risk are categorized as follows: (1) Children who cannot reach their dreams or be kept away from their dreams; (2) children living on the streets aged 0-15 years; (3) children with marginal or low income; (4) children in accordance with the age range determined by the child convention which is 0-17 years old; and (5) children whose communities and families face the law.



Figure 2. Documentation of the Activities of the Rumah Impian Foundation (Source: Field Documentation)

Rumah Impian Foundation has the belief that every street child has their own dreams. They are then fostered and accompanied by using a family approach so that there are no differences that can hinder their dreams.

This foundation has a vision, mission, and values that must be adhered to by street children guided by them, besides that it is also used as a companion as a stepping stone to accompany street children (see Figure 3). The vision of the Rumah Impian Foundation is "transforming the lives of children at risk through dreams that affect others" (Document study, December 2018). Meanwhile, the mission of the Rumah Impian Foundation is contained in three points: (1) accompanying children at risk as friends; (2) facilitating children at risk of realizing impactful dreams; (3) building a network to care about children's dreams. Then, for the values that exist in the Rumah Impian Foundation that they believe, (1) children have the right to dream and realize their dreams in a

supportive community - the road is not a good place for children; (2) children need to grow into individuals who have an impact on others - the transformation of children's lives begins with dreams.



Figure 3. Vision and Mission of the Dream House Foundation. Source (Field Documentation)

Until now, the foundation continues to move and build the dreams of street children and help them to make them come true by providing many things for the street children, one of which is the Education Center program at the stage of intervention in street children. Rumah Impian Foundation believes that education is one of the rights that street children must get. Through education, they can actualize themselves into what they dream of.

The education center is a dream house built for street children (Sajiwo, 2018). The purpose of this education center is for street children so that they can actualize themselves through learning, reading, drawing, to playing, and so on. The location of the Education Center is also available in several regions, meaning that the Rumah Impian Foundation does not only have one Education Center, but they are also spread in several regions. The areas targeted by the Education Center are scattered at several points, namely Sidomulyo, Tukangan, Ngemplak, Wonocatur, and Jogoyudan (Sajiwo, 2018).

This education center program portrays how the foundation cares about education for street children. However, it is necessary to conduct an in-depth study of how

education is carried out in it so that it can be known more deeply about the feasibility of the Education Center program at Rumah Impian Foundation.

Evaluation of Street Children Education

The evaluation of street children education is based on data on the number of street children in Yogyakarta in the last five years, as shown in Figure 4. Based on these data, it is necessary to conduct an initial analysis of the possible growth in the number of street children.

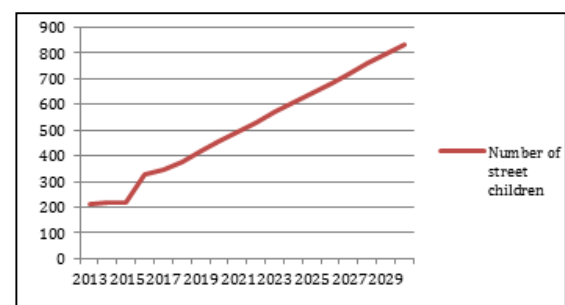


Figure 4. Projection of the Number of Street Children in Yogyakarta (Regional Development Planning Agency, 2018)

Based on the data in Figure 4, it is known that there is a significant increase in the number of street children from year to year. It can be seen from the data of the last five years, which are, of course, referred to as predictive data for the following years. The increasing number of street children is mostly caused by economic problems. Demands to fulfill life's needs make children late to experience the education that is appropriate to their age.

Street children who want to live better certainly become the main trigger in taking risky actions, namely taking to the streets. The desire to improve economic life and get money to meet daily food needs is the main thing they do. Apart from that, parents who feel less prosperous in the economy also prefer to employ their children to sing or just ask for mercy from others. On average, the street children in Yogyakarta themselves come from the Yogyakarta environment, which belongs to the marginal area.

Apart from these increases, the program and also assistance to tackle street children are not small, many communities of observers of street children in Yogyakarta who volunteered to help them reach their dreams, of course, did not leave the realm of education. The program carried out by the Social Service itself is in the form of a program that provides services and protection for children with social problems. However, the prediction of the increasing number of street children in Yogyakarta cannot be avoided. Therefore, it is necessary to conduct an in-depth evaluation so that it is known how the Education Center program is implemented at the Rumah Impian Foundation to have an impact on street children.

Rational Stage

At this stage, the implementation of education is viewed from a philosophical aspect, namely the purpose of education for street children. The background of education for street children at the Rumah Impian Foundation itself is first seeing the enthusiasm of street children who want to achieve their dreams, but many of the street children are hit with facilities and other things that make them not have the confidence to achieve their dreams. Starting from this point, education is directed at how they achieve their dreams (Interview data, December 12, 2018). Another thing behind the implementation of education for street children is opening their dreams that were once on the road to what they dreamed of. Through the desire to make appropriate changes for street children the founders of the foundation make a breakthrough, namely to open the horizons of street children, especially in the field of education.

Then, from 2003 to 2009, the number of street children in Yogyakarta seen at intersections or major roads continued to grow from 10 to 15 children by developing the initial meeting as a transit house (Interview data, December 12, 2018). The implementation of the background is in accordance with what is in the field; the foundation has an education center which is cur-

rently engaged in education for street children. Not only learning, playing, and honing skills, but more than that the foundation also facilitates and fosters them if they want to continue to a higher level of education (Interview data, 12 December 2018). From document studies, there are education center activities that are still running today and are developing by recruiting volunteers who care about street children.

Antecedent Stage

The antecedent stage is what the input is, which is the condition of the initial implementation of education for street children. This first step is to recruit volunteers who must comply with the rules or policies issued by the Rumah Impian Foundation. Recruitment of volunteers is conducted every three months, taking into account several criteria, namely: (1) must comply with child protection, because the foundation has a policy on child protection, starting from taking photos to joking with street children also regulated in this policy; (2) cooperating with several internship institutions: the Netherlands and United States of America.

Some of these volunteers were deployed directly to foster and assist street children who were also assigned to the office to create programs for street children. Then, for the financial input, the foundation is carried out using a donation system, because the foundation is engaged with the nonprofit sector so that the financial system is also limited.

From the results of the field studies and the input implementation documents, they have proceeded. Accordingly, they have recruited volunteers with due regard to the existing criteria, and then for the apprentice volunteers, they are also given the same policy that must comply with child protection. Then, the approach to street children is done by home visits and also a family approach to assisting street children to pay attention to their dreams and want to develop themselves. In addition, to ensure the safeguard rights and protections related to the personal data of street children, the foun-

dation also applies strict things, that is, by not publishing the names of street children they are taking care of, related to their personal rights.

Assessment for children assisted by the Rumah Impian Foundation was carried out by looking at the characteristics of street children. When the research took place, the mentors who were also carrying out formal education are 120 people, and they still lived with their families. There are two places of shelter that accommodate street children for 24 hours, and three districts are available for the assisted activities, namely Yogyakarta Municipality, Bantul Regency, and Sleman Regency. This activity was also supported by street contact or outreach children directly on the road.

Transaction Stage

The transaction stage is related to the process of implementing street children education at Rumah Impian Foundation. The process for carrying out street children's education is to direct them by developing the potential that exists within them. Provide direction that education is not only obtained through formal education, but education can be obtained in many ways. Indeed, street children need to receive a formal education. This process is carried out by cooperating with various education providers, both through education, chasing packages and entering children into schools or early education institutions that work with the foundation.

For now, the facilities provided by the Rumah Impian Foundation are helping them to enter formal schools, because the school is a stepping stone to find their identity, a place in which they are trained to be able to adapt to the environment outside the street (in the sense of having friends who are not from the street environment and channeling children to outside school activities such as soccer schools and other courses). According to the foundation, to achieve the need to pass dreams, namely through bridges and bridges, the development of soft skills and formal schools is also developed.

The Rumah Impian Foundation also collaborates with the school founded by Romo Mangun. The collaboration provided includes a discount for school fees. The foundation considers that formal education is necessary for street children as a stepping stone to achieve their dreams. Then the involvement of volunteers in conducting training was also seen by the development of activities in the education center. Not only learning, but street children are also given the knowledge of other self-development such as self-actualization.

The process activities are certainly in accordance with what was done by the foundation. It was evidenced by the collaboration document between the school and the foundation, then also the schedule of learning activities at the Foundation. Also, the process of involvement of street children is starting to be active in achieving their dreams by formally attending school.

Outcome Stage

The stages of the outcome are related to the results to be achieved by the Rumah Impian Foundation, that are, the education of street children, in which street children must be able to actualize themselves, make them better and more aware to the environment. Not a few street children who continue their education to a higher level, such as entering a university with a direction they want include sports teachers, midwives, and becoming nurses, in addition to street children who attend vocational schools they choose to work directly in accordance with the majors they take.

Also, there are street children who have succeeded in becoming student council leaders in their schools, of course, this is proof that the desired expectations have been consistent with the achievement of street children's education. The number of street children carried out with assistance in formal schools amounts to 120 people who are spread throughout Yogyakarta.

The foundation issues a policy if the child can stand on their dreams; each of them will be expelled or handed over to their

parents. Then, the foundation will issue a support system that can channel the dreams of street children into reality.

The results in the field show the compatibility between what is supposed to be and the reality. In this case, the foundation continues to develop education centers through fun learning to prevent street children from going back to the streets and also giving them opportunities to work and make a career. When the street children have succeeded in achieving what they want, the foundation does not break the relationship but builds good relations with the family; this happens in all street children assisted in the assisted locations in each district. This activity is also carried out by using an activity matrix that can be carried out and can be controlled by each activity.

Conclusion

Street children's education carried out by Rumah Impian Foundation is an education that leads to self-actualization of street children. Education is directed towards achieving the dreams they wish for.

The evaluation produces conclusions at several stages. The first is the rational stage, in which the educational background for street children emerges, namely with concern to make the founders of the foundation establish a foundation that focuses on achieving the dreams of street children and establishing an education center involving street children accompanied by volunteers. The second is antecedent stages, in which the education implementation input education input should involve education, in this case, the foundation has also involved volunteers to conduct fun-based learning activities, with the hope that street children are not quickly bored with the learning done. It is in accordance with the conditions of street children in the Rumah Impian Foundation, where, the input is made in accordance with the volunteer recruitment and also provides rules for the interests of the rights of street children fostered by the foundation. The third stage is the transaction, in which the process of education for street children

should be carried out in accordance with their conditions.

The findings indicate that there is a match between what the foundation does and what it should do. The foundation facilitates street children who need education by helping them complete their vocational education courses, including those in early childhood education, and also in collaboration with schools that can facilitate formal education for street children. Then, in terms of the accuracy of the teaching process in the education center, training is also given to volunteers to foster street children according to their talents and abilities, so that it is appropriate.

The fourth stage is the outcome, in which the Rumah Impian Foundation requires street children not to go back to the street anymore, and children can succeed in actualizing themselves. In other words, it is called the termination process, carried out to foster the independence of street children, and also provide a broader life experience by providing connections to the world of work through the schools they have chosen. Data in the field shows that the foundation has carried out this termination process well, proven by the street children who had passed through the intervention process to get a better life. It is certainly in accordance with the goals of the Rumah Impian Foundation.

Acknowledgment

The authors deliver the gratitude to the Rumah Impian Foundation and the volunteers involved in it.

References

- Aptekar, L., & Stoecklin, D. (2014). *Street children and homeless youth: A cross-cultural perspective*. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-94-007-7356-1>
- Article 31 of the 1945 Constitution of the Republic of Indonesia (1945). Republic of Indonesia.

- Goenawan, R., & Harnoko, D. (1993). *Sejarah sosial Daerah Istimewa Yogyakarta: Mobilitas sosial DI. Yogyakarta periode awal abad duapuluh*. Jakarta: Direktorat Jenderal Kebudayaan.
- Law No. 23 of 2002 on Child Protection (2002). Republic of Indonesia.
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).
- Lusk, M. (1992). Street children of Rio de Janeiro. *International Social Work*, 35(3), 293–305.
- Malia, I. (2016, November 3). 16 Provinsi rawan anak jalanan. *Harian Nasional*. Retrieved from <http://www.harnas.co/2016/11/03/16-provinsi-rawan-anak-jalanan>
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: SAGE Publication.
- Neolaka, A., & Neolaka, G. A. (2017). *Landasan pendidikan: Dasar pengenalan diri sendiri menuju perubahan hidup*. Jakarta: Kencana.
- Patel, A. (1983). *An overview of street children in India*. New York, NY: Covenant House.
- Ray, S. (2017). A street child's perspective: A grounded theory study of how street children experience and cope with grief. *The Qualitative Report*, 22(1), 291–308.
- Regional Development Planning Agency. (2018). Penyandang masalah kesejahteraan sosial dan sarana kesejahteraan sosial. Retrieved from http://bappeda.jogjaprovo.go.id/dataku/data_dasar/cetak/105-penyandang-masalah-kesejahteraan-sosial-dan-sarana-kesejahteraan-sosial
- Rochatun, I. (2015). Eksploitasi anak jalanan sebagai pengemis di kawasan Simpang Lima Semarang. *Journal of Unnes Civic Education*, 1(1), 22–29.
- Sajiwo, R. G. (2018). *Model evaluasi pembelajaran sejarah: Studi kasus pada Yayasan Rumah Impian di Kalasan, Sleman Yogyakarta*. Fakultas Dakwah dan Komunikasi Universitas Islam Negeri Sunan Kalijaga, Yogyakarta.
- Stake, R. E. (2005). Qualitative case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 443–466). Thousand Oaks, CA: SAGE Publications.
- Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2000). *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.). Boston, MA: Kluwer Academic Publishers.
- Worthen, B. R., & Sanders, J. R. (1973). *Educational evaluation: Theory and practice*. Worthington, OH: Longman.

THE IMPLEMENTATION OF ATTITUDE ASSESSMENT IN CURRICULUM 2013 AT ELEMENTARY SCHOOLS

Ikhya Ulumudin

The Center for Research on Education Policy and Culture, Ministry of Education and Culture

Sisca Fujanita

The Center for Research on Education Policy and Culture, Ministry of Education and Culture


Abstract


This study is aimed at describing the implementation of attitude assessment conducted by teachers starting from the assessment planning, implementation, results processing, its utilization, and the follow-up. This study employed a qualitative and quantitative method. The subject was elementary school teachers in Bekasi city who had taught in classes applying Curriculum 2013. The data were collected using questionnaires and focus group discussions. The result shows that the implementation of attitude assessment had not been done optimally because of teachers' lack of understanding. Besides, too many techniques of attitude assessment cause teachers to need a lot of time to learn and conduct the assessment. In Curriculum 2013, students' involvement is emphasized, yet most students were quite passive in class, and there was no dynamic and active discussion in the learning process. It means teachers need to apply different methods to encourage students to be more active. Lastly, the utilization and follow-up of the assessment results were not optimal. Some recommendations that can be proposed are: (1) schools are hoped to be able to provide independent training for the implementation of assessment in Curriculum 2013; (2) mandatory attitude assessment required to be done by each subject teacher is observation, while other required assessments are done by the classroom teacher; (3) variables in attitude assessment needs to be completed and clearly stated especially in social aspects (KI-2), which are accepting and responding; (4) there should be a detailed description of attitude assessment in students' reports.

Keywords: *attitude assessment; assessment planning; assessment implementation; assessment result processing; assessment result utilization; assessment result follow-up*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.23391>

Contact *Ikhya Ulumudin*

 *ikhya.puslitjak@gmail.com*

 *The Center for Research on Education Policy and Culture, Research and Development Agency, Ministry of Education and Culture of Republic of Indonesia
Jl. Jenderal Sudirman, Komplek Kemendikbud gedung E lantai 19, Senayan, Jakarta
10270, Indonesia*

Introduction

The government, through the Ministry of Education and Culture (MoEC), continues making innovations in education, one of them is by renewing the curriculum of primary and secondary education. Gradually, starting in 2013, the government has implemented Curriculum 2013 at primary and secondary education. It is done as a way to control the education quality so that Indonesia can have generations which will be able to compete worldwide.

The changes in the curriculum affect the assessment system. Conducting the assessment of learning outcomes is a way done to control education quality. According to Clements and Cord, assessment is an essential component in the learning process and its' environment, and it also has a role in knowing the learning outcomes. Assessment is hoped to be the instrument of the quality insurance, quality control, and quality improvement in the education system as in the scale of class or school (Sutama, Sandy, & Fuadi, 2017). Assessment is a series of activities to obtain qualitative and quantitative information in the beginning, ongoing, or at the end of the learning process, aimed to evaluate and diagnose things need to be fixed so that teachers and students can review, plan, and apply the action to be done next to achieve the learning goals (Purnomo, 2013).

Assessment of learning outcomes in primary and secondary education level consists of the assessment conducted by teachers, schools, and government. It covers the assessment of attitude, knowledge, and skill. This study will further discuss the elementary school attitude assessment that is conducted by teachers. Assessment of learning outcomes is conducted by the teacher to monitor and evaluate the learning process, learning progress, and also to improve student's learning outcomes continuously. Moreover, the attitude assessment is specifically conducted by the teacher to obtain the descriptive information of student's attitude (Regulation of the Minister of Education and Culture of Republic of Indonesia No. 23 of 2016 on Educational Assessment Standard).

The authentic assessment is used as an assessment system in Curriculum 2013. Authentic assessment is a significantly meaningful measurement of a student's learning outcomes in the attitude, knowledge, and skill aspect. It can describe the student's competencies more comprehensively and objectively, even when the student has completed his education (Wuryani & Irham, 2014). However, the changes in the assessment system caused problems for the teachers. Utsman (2014) in his study mentioned that the changes of the assessment system in Curriculum 2013 confused the teachers because the traditional assessment system which was already established and easy to conduct had to be changed with the alternative assessment in the form of authentic assessment. His study described that many teachers have difficulty in understanding the Curriculum 2013 assessment technique, especially on how to do the authentic assessment. Most teachers know the assessment but do not know how to apply and to adjust it as demanded in Curriculum 2013. Another study stated that some teachers still have difficulty in applying the Curriculum 2013; the greatest one faced by teachers is in applying the authentic assessment (Haryana & Gimin, 2015).

In Curriculum 2013, in addition to knowledge and skill assessment, attitude assessment explicitly has to be done as well by the teacher. It adds to the difficulty faced by the teacher, because in the previous curriculum, teachers were not asked to conduct the attitude assessment. Yulia, Bakhtiar, and Fauzi (2017) mentioned that teachers' problems in applying the assessment are its' complexity and the limited amount of time. Besides, Mahmud (2014) states in his study, "all teachers in Delima Cluster have difficulty in conducting an assessment, especially the attitude assessment, which is considered as a complex assessment."

Attitude assessment is needed to face global competition so that the graduates can be successful. Someone's success is influenced by some factors, one of which is his attitude. It can be seen from a survey con-

ducted by Thomas J. Stanley, Ph.D. that is written in his book, *The Millionaire Mind*. The survey was conducted in the United States of America with a total of 1001 respondents, of which 733 respondents were billionaires. Based on the research, there are 30 determinants of success. The top 10 determinants are attitude competencies such as: being honest, discipline, socializing skill, having a supportive partner, work harder than others, love what is done, great and strong leadership, passionate and competitive personality, good life management, and the ability to present ideas and products (Hatmoko, 2016).

Attitude assessment is intended as an assessment of student's attitude in the learning process, consisting of spiritual attitude (Core Competency-1/KI-1) and social attitude (Core Competency-2/KI-2). Attitude assessment has different characteristics from knowledge and skill assessment, so the techniques are also different. In this case, the attitude assessment focuses more on fostering students' behavior to build their character. Attitude assessment consists of the main assessment and supporting assessment. Main assessment is obtained by daily observation, which is written in the daily journal. Supporting assessment is obtained by doing self and peer assessment, which its result can be used as a confirmation of the assessment done by the teacher. The assessment technique used is observation through an interview, anecdotal record, and incidental record as the main assessment element. The spiritual competencies (KI-1) which are being observed are accepting, doing, and respecting the religion they adhere to. Meanwhile, the social (KI-2) ones include several attitudes, among others: honest, disciplined, responsible, polite, caring, and confident in interacting with family, friends, neighbor, and the country (Directorate for the Development of Elementary School, 2016). Preparing a generation with excellent characters should be started since their early age, which means elementary schools. A study conducted by Riscaputantri and Wening (2018) shows the affective descriptions or student's attitudes

are (1) based on the affective levels proposed by Krathwohl, Bloom and Masia: characterizing level 42.9%, organizing 50.9%, judging 6.2%, and accepting and responding 0%; (2) based on Curriculum 2013: on excellent category 62%, good 38%, adequate 1% and poor 0%.

Process of the attitude assessment starts from planning, implementation, processing, utilization, and the follow-up. The planning is done based on KI-1 and KI-2. The teacher plans and decides the attitude being assessed in the learning process. In addition, for the assessment outside the learning process, the teacher observes the natural behavior. The implementation of attitude assessment is adjusted to the learning approach taken during and outside the learning process. The result of attitude assessment will be discussed and reported in the form of a descriptive score. The result is utilized to get the information of student's attitude, identify the progress, and do the follow-up.

Based on the description above, the researchers are prompted to conduct a study of the implementation of the attitude assessment in Curriculum 2013 at elementary schools. This study aimed to describe the implementation of attitude assessment conducted by teachers starting from the assessment planning, the implementation, the assessment results' processing, its' utilization, and the follow-up.

Research Method

The study was conducted on August 8-10, 2018, in Bekasi, West Java. This study used a mixed-method (qualitative and quantitative method), and the data were collected using questionnaires and focus group discussions. The questionnaires were online and could be accessed using a smartphone application with google drive as the basis. Respondents could fill in the questions in the instruments through <https://goo.gl/forms/AEwhFql30oS1acn23>. The regional education office distributed the instruments page (URL link) through the elementary school principals group chat on What's App (WA).

Then, the principals distributed it to the teachers in their own school. The FGDs were conducted in the discussion room of Bekasi City Education Regional Office, inviting qualified-teachers who were selected by the education regional office staff.

The subject were elementary school teachers, both public and private school in Bekasi, who had taught in class applying the Curriculum 2013. The subjects are the informants in the Focussed Group Discussion (FGD) and questionnaire respondents. A total of eight persons, consisting of six elementary school teachers, one elementary school supervisor, and one staff of the regional education office managing the elementary level, were involved as the informants in FGD. The staff of regional education office selected the informants with some provisions, such as, the selected teacher is an instructor of Curriculum 2013, or he/she is the one who made the questions for the National-Standard School Examination questions or the one who had an experience of teaching the class applying Curriculum 2013. There are a total of 1,528 questionnaire respondents, who came from the teachers of Bekasi public and private elementary school which have been implementing Curriculum 2013. The regional education office coordinated the selection of the questionnaire respondents.

The questionnaire used closed questions with three choices of answers: understood, partly understood, not at all understood. The questionnaire-filling is aimed at finding out teachers' understanding (1) in making an instrument format of attitude assessment, (2) on different techniques of attitude assessment, and (3) in processing and reporting the attitude assessment. Whereas, the FGD guidance consisted of three opened questions: (1) confirmed and explored teachers' questionnaire answer; (2) discussed teachers' tendency in answering particular choices; and (3) identified the difficulties and problems faced by teachers in conducting the attitude assessment. Thus, the FGD aimed to confirm and explore teachers' questionnaire answer and to identify the difficul-

ties and problems faced by teachers in planning, implementing, processing the results, utilizing, and doing the follow-up of the attitude assessment.

The field research procedure consisted of three steps. First, the researchers came to the regional education office to ask for permission to conduct a study and ask the regional education office to distribute the questionnaire to elementary school teachers, with the letter of introduction about the appeal of filling out the questionnaire from the head of regional education office being inserted. Second, the result of the processed-questionnaire can be seen directly on Google Drive. Third, the distribution of the questionnaire was confirmed, and the answer was explored through FGD. Last, in FGD, the researchers have also identified the difficulties and problems faced by teachers in planning, implementing, processing the results, utilizing and doing the follow-up of the attitude assessment.

The data were analyzed using a quantitative descriptive technique in the form of a percentage of each question's answer. It was done in several steps: (1) analyzing the questionnaire answers; (2) conducting triangulation of the questionnaire results by confirming to FGD informants; (3) identifying and exploring the cause of the tendency of respondents' answers; (4) identifying the difficulties and problems faced by teachers in planning, implementing, processing the results, utilizing, and doing the follow-up of the attitude assessment while conducting triangulation of the three informants (teachers, supervisors, and the staff of regional education office); (5) drawing conclusions of the research.

Findings and Discussion

Findings

Findings are elaborated based on the purpose of the study, which is to describe the implementation of attitude assessment conducted by the teacher, starts from planning, implementing, processing the results, utiliz-

ing and doing the follow-up. Each step is explained as follows.

Assessment Planning

The planning of attitude assessment is a teacher's preparation in making instruments or assessment format of each attitude assessment technique. The techniques are observation, self-assessment, and peer assessment. The instruments of observation are unstructured observation form, structured observation form, social and spiritual journal. While the instruments of self-assessment are self-assessment form and for the peer assessment is peer assessment form. The attitude assessment instruments cover the spiritual aspect (KI-1) and social aspect (KI-2), which appear naturally during the learning process inside or outside the classroom. KI-1 which were being observed was accepting, doing, and respecting the religion students adhere to. Meanwhile, KI-2 which were being observed included attitude among others: honest, disciplined, responsible, polite, caring, and confident in interacting with family, friends, neighbor, and the country.

Planning an attitude assessment requires teachers' understanding of the KI-1 and KI-2 indicators which are being observed, so teachers can easily make the attitude assessment instruments. However, many teachers did not understand it. Based on the questionnaire results, a total of 45.22% elementary school teachers in Bekasi City understood how to elaborate the indicators of KI-1 and KI-2, and turn it into the form of an assessment instrument. Meanwhile, 52.03% of teachers partially understood, and 2.75% did not understand. In FGD, it was revealed that teachers' lack of understanding was specifically in two things: the elaboration of attitude aspect indicator into the instrument and the making of scoring rubrics.

First, the teacher had not been able to elaborate the attitude aspect indicator into the instrument. Teachers found it difficult to elaborate the indicators of spiritual and social aspect into questions/statement or assessment form. For example, in turning the

indicators of spiritual aspect (accepting, doing, and respecting the religion students adhere to) into questions/statement or assessment form. As well as the social aspect, it was complicated for teachers to turn the indicators (honest, disciplined, responsible, polite, caring, and confident in interacting with family, friends, neighbor, and the country) into questions/statement or assessment form.

Teachers thought of some solutions to overcome those problems, such as used assessment instruments with the same format as the example given by the Directorate for Development of Elementary School in *Panduan Penilaian Sekolah Dasar*. The majority of teachers used precisely the same self and peer assessment instruments as the sample in the guidance book without any development.

Second, teachers found difficulties in making the scoring rubric of attitude assessment. For example, in KI-1 attitude assessment, there was a student who prays regularly but also often disturbs other friends who are praying. On the other hand, he always thanks others when receiving help or gifts. The teacher was confused about what score (excellent, very good, adequate, or poor) to give because, in one particular variable, the student had strengths and weaknesses. It created different perceptions among teachers in deciding the attitude score.

According to teachers, the social aspect indicators which include honest, disciplined, responsible, polite, caring, and confident in interacting with family, friends, neighbor, and the country, need to add some more variables to fit the Curriculum 2013 learning system. Curriculum 2013 learning emphasizes on students' involvement. However, in fact, most students were passive during the learning process; only a few of them were active. It made the learning atmosphere becoming less active. The research conducted by Retnawati (2015) concluded that the main difficulty in the implementation of scientific learning was to make the student become more active and applying 5M.

In the FGD, it was revealed that in guidance book, there had not been explicitly contained information on how to determine the attitude score. In the guidance book, it was only stated about the scoring predicate, namely A (excellent), B (good), C (adequate), and D (poor) but there was no explanation on how a student could get a particular score. Teachers hope that the information on how to make the scoring rubrics could be put in the guidance book, to minimize the different perceptions among teachers.

Teachers' competencies to comprehend the various attitude assessment technique could affect their competency in doing the assessment. The same thing was mentioned in the research done by Retnawati (2015) that "teachers' understanding of social attitude assessment significantly affected their skill in implementing social skill assessment." For this reason, teachers should improve their understanding of attitude assessment continuously, and school should provide training of Curriculum 2013 assessment for teachers independently.

Assessment Implementation

Teacher implements attitude assessment during and out of the learning process. The assessment technique consists of the main assessment and supporting assessment. Main assessment is obtained by daily observation, which is written in the daily journal. The assessment technique used is observation through an interview, anecdotal record, and incidental record. The supporting assessment is obtained by doing self and peer assessment, which its result can be used as a confirmation of the assessment done by the teacher. The technique used is distributing questionnaires, both in self and peer assessment.

Teachers are required to master various assessment techniques, both main and supporting assessment. However, not all elementary school teachers mastered those techniques. From the questionnaires, it is known that 54% of elementary school teachers had a good understanding of the assessment techniques, 42.54% with partial under-

standing, and 2.88% did not understand the techniques. In FGD, it was mentioned that observation by making an incidental record was the most used assessment technique. While interview, self-assessment, and peer assessment were rarely used because the teacher did not really understand the techniques.

Observation techniques by making incidental record were used by teachers to record students' "extreme" attitude (good and bad attitude). Based on FGD, steps done by teacher in making incidental record observation were: (1) students' attitude which was recorded in journal were the very good and the very bad ones and related to the attitude items of spiritual and social aspect; (2) if student has had unfavorable records, and if on another occasion the student has shown good attitudes in the same indicator, then in the journal it was written that the student has been good or even very well; (3) records in the journal were not limited to the bad and good attitude, but also the development towards the expected attitude; (4) based on the journal, teacher made a description of students' attitude assessment in a semester.

Two supporting attitude assessments function as data confirmations, namely self-assessment, and peer assessment. Self-assessment is an assessment technique where student assesses themselves by identifying their strengths and weaknesses in behaving. Meanwhile, peer assessment is an assessment technique where the student assesses another students' attitude. Not all teachers in Bekasi City conducted self and peer assessment.utama et al. (2017) mentioned the same thing on his study, "in conducting the attitude assessment, teachers only used observation technique and wrote it in the daily journal."

Self and peer assessment were usually conducted at the end of the semester because the assessment results were used as data confirmation of student attitude development during one semester. Steps done by teachers in conducting self and peer assessment were: (1) conducted self-assessment first and then peer assessment in the same

period of time; (2) in peer assessment each student usually assessed one to two other students; (3) the results of self and peer assessment were used as confirmation data by matching it with the result of observation, and then used to determine the attitude score for student's report.

Based on these reasons, teachers want the assessment that has to be done by all teachers is the assessment using the observation technique only. If a classroom teacher does self and peer assessment, this will save time, because, in the authentic assessment, there are a lot of techniques and students to be assessed. It is in line with the study conducted by Merta, Suarjana, and Mahadewi (2015) which found that "teachers' difficulties in conducting authentic assessment were a large number of students, many assessment techniques to be done, and the availability of time to conduct the assessment."

The Processing of Assessment's Results

Results of attitude assessment were recapitulated every semester. The result would be discussed and reported in the form of description. Steps in making a description score were: (1) classroom teacher and subject teachers classified and marked the records of students' attitude written in the journal, both spiritual and social aspects; (2) classroom teachers recapitulated the students' attitude within one semester (the period can be adjusted according to the school consideration); (3) classroom teachers collected the short description of students' attitude from religious teacher and PE teacher and also from the school residents (extracurricular teachers, librarian, janitors, and also school guards); (4) classroom teacher inferred and made the final description of students spiritual and social achievement.

Elementary school teachers were still lack of understanding the processing and reporting of attitude assessment. The result of the questionnaire showed that 41.49% of elementary school teachers in Bekasi City understood the processing and reporting of attitude assessment, 52.57% with partial understanding, and 5.76% did not under-

stand. It occurred in every teacher or classroom teacher.

Majority of elementary school teachers in Bekasi City did an attitude assessment by doing observation and making incidental records of students' good and bad attitude. The student with a good attitude was given rewards such as getting praised by the teacher. Meanwhile, students with bad records were given guidance until their attitude change. The development of students' bad attitude was recorded in the journal. Teachers had difficulties in collecting and inferring the short description of assessment from various indicators and techniques.

Classroom teachers had difficulties in recapitulating and inferring the short description. Inferring the attitude assessment was confusing, for example, when there is one student who gets different attitude score from different teachers and the scores are even opposite each other. The attitude score had not yet represented students' attitude, it consisted of only several indicators, and it was still general. This problem arose because the formulation of determining the predicate and description of attitude assessment had not yet explained explicitly in the guide book. Teachers hope the process of attitude assessment can be simplified so the results can be utilized appropriately. In the guidance book, it was explained that school had the authority to determine the criteria of attitude description, in fact, the school did not make the criteria for the reason that the direction on making the criteria had not yet included in the guide book. On his study, Setiadi (2016) recommended the government to simplify the Curriculum 2013 guidance book, to conduct socialization and training of attitude assessment, to conduct training of thematic learning assessment technique for elementary school teachers, and to guide teachers in analyzing instrument and revising test items.

Another problem in processing the attitude assessment was that there was an intervention from the stakeholders. In FGD, some teachers mentioned that there were principals who intervened the assessment by requiring the teachers to give a minimum

score B (good) even when the students had a bad attitude, even though it could backfire the student itself.

The Utilization and Follow-Up

The results of attitude assessment were not optimally used. Teachers considered that the results were still general, so it could not be utilized optimally. The results should be described in details so students or parents could improve the attitude score that is still bad. Here is an example of the attitude description by a teacher: "Spiritual aspect score: Fauzan prayed regularly, be grateful, and was able to develop his religious tolerance. Social aspect score: Fauzan was very honest, confident, and needed guidance in discipline". Students or parents did not understand what it was meant by "very honest and needed guidance."

Description in the attitude assessment was considered meaningless by teachers. As in the students' report, it was only written "student was very obedient, very honest, needed guidance and counseling," this description was meaningless because student or parents were not specifically informed of students' attitude. It would be more meaningful if it were stated technically. For example, student was in need of guidance because he was often late, and influenced his friends for not attending school; students did not do Jumat prayer. By describing it specifically, both teachers, school, and parents could utilize and follow-up the report.

The teacher said that it would be meaningful if students' behavior were stated explicitly. For example, in the spiritual aspect score: Fauzan prayed regularly (praying and doing congregational prayer in an orderly manner), was grateful (for his meal even though it was not enough), and was able to develop his religious tolerance (befriended with students from different religions). Social aspect score: Fauzan was very honest (he found money at school and returned it to the teacher, he never cheated), confident (always answered teacher's questions loudly), and discipline (came to school on time, being patient while queuing).

The follow-up of attitude assessment is very important, so students can immediately change their bad attitude. In FGD, it was mentioned that the majority of teachers had done the follow-up. The steps in doing the follow-up of the attitude assessment results were: (1) the spiritual and social behavior being observed were recorded in teachers' journal, it was used as the follow-up by school; (2) teachers immediately did the follow-up of students' bad attitude by giving guidance; (3) classroom teachers could develop counseling and mentoring services for students had bad attitude; (4) if the students had bad attitude records and had not shown any positive change, the description of their attitude would be discussed in the teacher council meeting at the end of semester. The meeting decided on the predicate and description of students' attitude that should be written in the report and decided the follow-up for the students.

Discussion

The planning of attitude assessment is necessary. However, only some elementary school teachers in Bekasi City understood how to do the planning, especially on how to turn the attitude indicators into an instrument of the assessment and how to make a scoring rubric. It was because attitude assessment was a new addition in Curriculum 2013. Meanwhile, in the previous curriculum, teachers were not asked to conduct the attitude assessment (only observing and scoring). Besides, there are too many instruments to be mastered by teachers, namely unstructured observation form, structured observation form, spiritual and social journal, self-assessment form, and peer-assessment form. Teachers ability to understand the attitude assessment technique could affect the quality of the assessment. Based on these reasons, teachers suggested that the assessment done by all teachers is the observation technique only, while the classroom teacher does other techniques.

Curriculum 2013 emphasizes on students' involvement in the learning process. However, most students were passive during

the learning process and made the learning atmosphere becoming less active. Therefore, some variables need to be added in the attitude assessment, namely the variable of accepting and responding, which are included in Bloom's Taxonomy affective assessment. Accepting refers to the ability to show attention and give a response to the appropriate stimulation and the ability to show attention and appreciation for others, for example: listen to and appreciate other's opinions. Besides, responding refers to students' affective involvement in the learning process, become participants and interested in the learning materials, such as asking questions, actively participating in the in-class discussion, etc.

There were several techniques in implementing attitude assessment during or outside the learning process, but there were only some elementary school teachers in Bekasi who understood the techniques. Most teachers conducted the attitude assessment through observation technique by making an incidental record. While interview, the self-assessment, and peer assessment were rarely used because the teacher did not really understand the techniques. Some reasons of the incomprehension mentioned by teachers were (1) teachers did not fully understand in how to make the self, and peer assessment instruments; (2) teachers were confused on deciding the final score (the final score was a combination of observation score, self, and peer assessment score); and (3) teachers felt that there was not enough time to conduct self and peer assessment.

Both teachers' and classroom teachers' understanding of processing the attitude assessment were still lack. Teachers had difficulties in collecting the records and inferring the short description of the assessment result from various indicators and assessment techniques. Meanwhile, the classroom teachers had difficulties in recapitulating and inferring the short description from several teachers and school residents. In FGD, some teachers mentioned that some principals required their teachers to give a minimum score of B (good) even when the students had a bad attitude.

The utilization of the attitude assessment results was not optimal. In the students' report, it was only written "student was very obedient, very honest, needed guidance and counseling," which was considered too general by teachers, so it could not be utilized by students or parents to conduct evaluation. Majority of teachers did an immediate follow-up of students' bad attitude by giving them guidance and counseling.

Conclusion

The implementation of attitude assessment which is conducted by teachers starting from the assessment planning, the implementation, the assessment results' processing, its' utilization, and the follow-up is not optimal. It is because of teachers' lack of understanding. Generally, the lack of understanding is because the attitude assessment is a new addition in Curriculum 2013, and in the previous curriculum, teachers were not explicitly asked to conduct the attitude assessment. Moreover, there were too many techniques of attitude assessment. It needed much time to learn, implement, and process the assessment. Curriculum 2013 emphasizes on students' involvement in the learning process. However, most students were passive during the learning process and made the learning atmosphere becoming less active, so the teachers need to think of a way to increase students' participation. The utilization of the attitude assessment was not optimal.

To optimize the implementation of attitude assessment in Curriculum 2013 for primary education level, researchers had summarized some recommendations obtained from the study, including: (1) schools are expected to provide independent training for the implementation of assessment in Curriculum 2013, because in the previous curriculum, teachers were not explicitly asked to conduct the attitude assessment; (2) MoEC is expected to change the regulation of the use of attitude assessment techniques, namely: the attitude assessment required to be done by each subject teacher is observation, while other required assessments are to be

done by the classroom teacher, because teachers had difficulties in processing and formulating the attitude assessment combined from several techniques; (3) MoEC is expected to add more variables in attitude assessment especially in social aspects (KI-2), accepting and responding. Accepting refers to the ability to show attention and give a response to the appropriate stimulation and the ability to show attention and appreciation for others. Meanwhile responding refers to students' affective involvement in the learning process, become participants and interested in the learning materials. It is in line with Curriculum 2013 that emphasizes on students' involvement in the learning process; (4) there should be a detailed and technical description of attitude assessment scoring in students' reports, for example student X are given C for the attitude score because he was late 30 times, 20 times did not come to school, etc. This description is more meaningful and can be utilized by students and parents.

Acknowledgment

This work is supported by the Center for Policy Research in Education and Culture, Research and Development Agency, Ministry of Education and Culture under the budget of Center for Policy Research, in 2018.

References

- Directorate for the Development of Elementary School. (2016). *Panduan penilaian untuk sekolah dasar*. Jakarta: Kementerian Pendidikan dan Kebudayaan. Retrieved from <http://ditpsd.kemdikbud.go.id/wp-content/uploads/2017/06/Panduan-Penilaian-untuk-Sekolah-Dasar.pdf>
- Haryana, G., & Gimin, G. (2015). Hambatan yang dihadapi guru ekonomi SMA dalam implementasi Kurikulum 2013 di Kota Pekanbaru. *PEKBIS (Jurnal Pendidikan Ekonomi Dan Bisnis)*, 7(2), 146–151.
- Hatmoko, W. (2016). 10 Faktor kesuksesan seseorang. Retrieved January 7, 2019, from <https://ciptacendekia.com/forums/topic/10-faktor-kesuksesan-seseorang/>
- Mahmud, M. (2014). Kendala guru dalam melakukan penilaian pada proses pembelajaran kurikulum 2013 di Sekolah Dasar Gugus Delima Banda Aceh. *Pesona Dasar (Jurnal Pendidikan Dasar Dan Humaniora)*, 2(3), 33–44. Retrieved from <http://www.jurnal.unsyiah.ac.id/PEAR/article/view/7497>
- Merta, I. M. E. D., Suarjana, I. M., & Mahadewi, L. P. P. (2015). Analisis penilaian autentik menurut pembelajaran kurikulum 2013 pada kelas IV SD No. 4 Banyuasri. *Journal PGSD Universitas Pendidikan Ganesha Jurusan PGSD*, 3(1). Retrieved from <https://ejournal.undiksha.ac.id/index.php/JJPGSD/article/viewFile/5818/4207>
- Purnomo, Y. W. (2013). Keefektifan penilaian formatif terhadap hasil belajar matematika mahasiswa ditinjau dari motivasi belajar. In *Seminar Nasional Matematika dan Pendidikan Matematika, in the theme of "Penguatan Peran Matematika dan Pendidikan Matematika untuk Indonesia yang Lebih Baik."* Yogyakarta: Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta.
- Regulation of the Minister of Education and Culture of Republic of Indonesia No. 23 of 2016 on Educational Assessment Standard (2016).
- Retnawati, H. (2015). Hambatan guru matematika sekolah menengah pertama dalam menerapkan kurikulum baru. *Jurnal Cakrawala Pendidikan*, XXXIV(3), 390–403. <https://doi.org/10.21831/cp.v3i3.7694>
- Riscaputantri, A., & Wening, S. (2018). Pengembangan instrumen penilaian

- afektif siswa kelas IV sekolah dasar di Kabupaten Klaten. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 22(2), 231–242. <https://doi.org/10.21831/pep.v22i2.16885>
- Setiadi, H. (2016). Pelaksanaan penilaian pada kurikulum 2013. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 20(2), 166–178. <https://doi.org/10.21831/pep.v20i2.7173>
- Sutama, S., Sandy, G. A., & Fuadi, D. (2017). Pengelolaan penilaian autentik kurikulum 2013 mata pelajaran matematika di SMA. *Jurnal Manajemen Pendidikan*, 12(1), 105–114. <https://doi.org/10.23917/jmp.v12i1.2967>
- Utsman. (2014). Penilaian otentik berbasis kurikulum 2013. In *Proceeding of Seminar Nasional Evaluasi Pendidikan, in the theme of "Pengembangan Pendidik: Implementasi Asesmen Otentik Pendidikan dalam Rangka Meningkatkan Kompetensi dan Kinerja Profesional Berkelanjutan"*. Semarang: Universitas Negeri Semarang.
- Wuryani, W., & Irham, M. (2014). Penilaian dalam perspektif Kurikulum 2013. *Insania: Jurnal Pemikiran Alternatif Kependidikan*, 19(1), 181–199. <https://doi.org/10.24090/INSANIA.V19I1.470>
- Yulia, L., Bakhtiar, B., & Fauzi, F. (2017). Kendala guru dalam mengimplementasikan buku paket kurikulum 2013 di SD Negeri 50 Banda Aceh. *Jurnal Ilmiah Mahasiswa Pendidikan Guru Sekolah Dasar*, 2(1), 204–211. Retrieved from <http://www.jim.unsyiah.ac.id/pgsd/article/view/2549>

THE APPLICATION OF THE GENERALIZED LORD'S CHI-SQUARE METHOD IN IDENTIFYING BIASED ITEMS

Herwin

Faculty of Education, Universitas Negeri Yogyakarta

Sophak Phonn

Ministry of Education, Youth, and Sport of Cambodia

Abstract


This study aims to find out the results of the analysis of item bias by using Generalized Lord's Chi-square test method on the test instrument of elementary school examination in a sub-district of Gowa Regency, Indonesia. This research is explorative research using quantitative approach. This research was conducted in the second semester of the academic year of 2017/2018 at the Bontomarannu District Elementary School in Gowa Regency. Data collection technique used is documentation. The data in this study were analyzed using the DIF method of Generalized Lord Chi-Square test. The results show that in using the Generalized Lord Chi-Square method, from 20 items of mathematics test of school exam in Bontomarannu District Elementary School in Gowa Regency of the academic year 2017/2018, there are two items which are detected containing the bias (DIF), i.e., item 5 and item 11, while the rest are not.

Keywords: *DIF, generalized Lord's chi-square, school examination*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.20665>

Contact *Herwin*

 herwin89@uny.ac.id

 Department of Elementary-School-Teacher Training Education, Faculty of Education,
Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia

Introduction

Assessment is an important component of education. One effort that can be done to improve the quality of education is to improve the quality of the assessment system. A good assessment system that will encourage teachers to determine good teaching strategies in motivating students to learn better. Therefore, to improve the quality of education, it is necessary to improve the assessment system that is applied. From this assessment, the teacher will obtain the ability of the portrait or profile of students to achieve the basic competencies set out in the curriculum (Herwin, 2016, p. 1).

A good rating system must be supported by various components. One component that supports the implementation of a good assessment is a quality assessment instrument or measuring instrument. In creating a measurement, a valid and consistent (reliable) measurement tool is needed. By using a measuring instrument that meets both of these criteria, the results of repeated measurements will be obtained in accordance with what is to be measured without being affected by other factors. A condition on a test instrument that is influenced by other factors, other than what is being measured, is called a bias on a test (Retnawati, 2014, p. 125).

One indication of a good instrument is an instrument that is free from the threat of bias. Bias is defined as a systematic error in the measurement process (Osterlind, 1983, p. 10). Bias is a negative condition that can threaten the quality of the test. Item bias can harm a group of students based on gender, religion, ethnicity, social, and economic status of different groups, when, in fact, the two groups have the same knowledge and also abilities.

The Differential Item Function (DIF) is a procedure used to identify bias test items. DIF is a psychometric method which is generally used to overcome justice in achieving standards, intelligence, certification, and testing of licenses (Jodoin & Gierl, 2001, p. 329). DIF occurs when participants from various groups show different probability of success

in the item after matching the ability underlying the item intended to be measured (Zumbo, 1999, p. 12). The DIF analysis aims to detect differences in item responses in questionnaires, rating scales, or tests in different subgroups (e.g., sex) while controlling ability levels (Zhang, 2015, p. 1). Bias is a threat to the test. Bias can appear as an invalid problem, for example, in terms of construction problems or content that is not suitable for a particular group of students. The bias of items in an instrument can affect test results that are not good. That is, the decisions obtained from the test will be less objective if the instrument contains an item bias. The bias presence of items must be detected so that the test can be of higher quality.

Since the function and purpose of the test are very important, the measurement instrument used must be of high quality (free of bias). Thus, the detection of bias items in a measuring instrument is very important. The bias item procedure is used to determine whether individual items in the examination function in the same way for two groups of examinees are usually defined by racial and ethnic background, gender, age, and experience, or disability conditions (Scheuneman & Bleistein, 1999, p. 220).

Statistically, bias items occur when there is a difference in support (alignments) of an item against a particular group. To understand the source of the bias that occurs in an item, it is necessary to check whether the response to the item is systematically related to certain characteristics of the respondents such as gender, age, ethnicity and other differences (Bares, Andrade, Delva, Grogan-Kaylor, & Kamata, 2012, p. 387).

Many different ways can be used in detecting bias items, but in this study, the method used is the Generalized Lord's chi-square test method. The Generalized Lord's chi-square test method (Kim, Cohen, & Park, 1995), is a DIF identification method which is often referred to as Qj statistics, making it possible to detect uniform or non-uniform differential function among many groups by setting the appropriate item re-

sponse model. Therefore, based on some of these descriptions, this research was conducted. If the DIF indication level is practically significant, it can be tested by using a certain statistical test or just by looking at the index, then the item in question is said to be exposed to DIF, load DIF, or detected as DIF item (Budiyono, 2009, p. 3).

Based on the description of the background of the problem that has been explained, the research question is how the results of the bias item analysis using the Generalized Lord's Chi-square test method on the elementary school exam test instrument in Bontomarannu District, Gowa Regency in 2018? Based on the research question, the implementation of this study is aimed at determining the results of the bias item analysis using the Generalized Lord's Chi-square test method on the elementary school exam test instrument in the district of Bontomarannu, Gowa Regency, in the academic year of 2017/2018.

Research Method

This research is exploratory research using a quantitative approach, aimed at detecting the bias items or differential item functioning (DIF) on the class II elementary school exam test instrument in the district of Bontomarannu, Gowa Regency, in 2018. This research was conducted in the elementary school of Bontomarannu District, Gowa Regency, in the even semester of the academic year of 2017/2018.

Data collection techniques are carried out by documentation. The data in this study are class II elementary mathematics questions on school exams in Bontomarannu District, Gowa Regency in 2018, consisting of 20 items (15 multiple-choice items and five filling items). In addition to the question set, this research data are the results of the response or answers of the students as many as 400 answer sheets to be continued on the analysis of item bias detection. Data from the responses or answers of participants in the form of 1-0 dichotomous data are derived from the multiple choice objective test answers and objective filling tests.

In this study, the grouping is divided into four groups: rural women students, urban women groups, rural male groups, and urban male groups. In addition, in this study, the reference groups were labeled in rural women's groups. Meanwhile, focal groups were labeled in urban women's groups, rural men's groups, and urban male groups.

The data of this study were analyzed using the Generalized Lord Chi-Square test DIF detection method. This method is used to detect bias items in cases of more than one focal group (Magis, Béland, Tuerlinckx, & De Boeck, 2010, p. 852). Kim et al. (1995) expand the Lord Chi-square test method to be more than one focal group in a procedure called the Generalized Lord Chi-square test. Lord Chi-square statistics are then generalized to Equation 1:

$$Q_j = (Cv_j)'(C\Sigma_j C')^{-1}(Cv_j) \quad (1)$$

Referring to Equation 1, v_j is obtained by combining the vector parameter items estimated in the reference group and focal group. Σ_j is a diagonal block matrix where each diagonal block is a matrix of item variance-covariance parameters in each group. C matrix is a design matrix that shows the parameters of the items you want to compare between groups (for more details, see Kim et al., 1995). The threshold (or cut-score) for classifying items as DIF is calculated as the quantile of the chi-square distribution with a lower-tail probability of one minus alpha and p degrees of freedom.

Findings and Discussion

Findings

This research is focused on the application of the Generalized Lord Chi-square method to detect biased items in the school exams test instrument in the Elementary School of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018. This Generalized Lord Chi-square method is applied because in this case, there is more than one focal group. This research is based on problem analysis using data from the test

participant's response results or the results of student answers to mathematical questions that have been done.

In the DIF study, there were at least two groups, namely, focus and reference groups. Focus groups are basically a minority group, for example, are potentially disadvantaged groups. Groups that are considered potentially benefited by this test are called reference groups. However, it should be stressed that naming groups is not always clear. Naming groups in such cases is often done randomly (Karami, 2012, p. 60).

There are two types of DIF, namely DIF that is uniform and non-uniform. Uniform DIF occurs when groups perform better than other groups at all levels of ability. That is, almost all group members outperform almost all other group members who are at the same level of ability. In the case of non-uniform DIF, members of one group are favored to the extent of the ability scale and from that point on the inverse relationship. Thus, there is an interaction between grouping and level of ability (Karami, 2012, p. 60).

Based on the results of the research obtained from the question instrument, documentation, and responses, it is found that in the 2018 elementary school examination in Bontomarannu Subdistrict, specifically for Class II and Mathematics subject, the results showed that the instruments used were a multiple-choice objective test consisting of 15 items and five filling items. In addition, the results of the students' answers were also obtained in the form of response answers on each question that was done.

Analysis of item bias or detection of DIF in this study was viewed from two aspects, namely gender aspects (male and female) and location aspects (urban and rural). Data distribution from the subject of this study is described in Table 1.

Table 1. Distribution of Research Data Sources

Gender	Location	
	Rural	Urban
Female	127	130
Male	73	70

Table 1 shows that the overall data of respondents amounted to 400 response answers or 400 answer sheets. The data in the form of answers to the test participants are included in the application, namely the R Program, which is then estimated using the Generalized Lord Chi-Square test method. The results of the bias item analysis obtained from the R Program are presented in Table 2.

Table 2. Results of Item Bias Analysis using the GLC Method

Item	GLC	Cut-score	p	α	Bias
1	0.31	7.8147	0.94	0.05	no
2	0.63	7.8147	0.88	0.05	no
3	1.09	7.8147	0.77	0.05	no
4	0.39	7.8147	0.94	0.05	no
5	12.2	7.8147	0.00	0.05	yes
6	5.68	7.8147	0.12	0.05	no
7	0.94	7.8147	0.81	0.05	no
8	1.25	7.8147	0.74	0.05	no
9	1.11	7.8147	0.77	0.05	no
10	0.73	7.8147	0.86	0.05	no
11	32.8	7.8147	0.00	0.05	yes
12	2.58	7.8147	0.45	0.05	no
13	0.79	7.8147	0.85	0.05	no
14	1.92	7.8147	0.58	0.05	no
15	2.24	7.8147	0.52	0.05	no
16	0.43	7.8147	0.93	0.05	no
17	0.53	7.8147	0.91	0.05	no
18	7.03	7.8147	0.07	0.05	no
19	1.17	7.8147	0.75	0.05	no
20	2.77	7.8147	0.42	0.05	no

Table 2 shows information regarding the Generalized Lord's Chi-square statistical coefficient which is the analysis coefficient of DIF, the threshold/cut-score coefficient which is the comparative criterion, p-value which is an opportunity to reject the null hypothesis, α which is the significance level used, and the conclusion item analysis. Based on the results of the analysis presented in Table 2, it can be concluded that of the 20 items of mathematics questions in the school examination in the state elementary school

of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, there are 18 items that are categorized well (free from gender and location bias), while two items others were detected significantly containing bias items, namely: Item 5 and Item 11. Both items were in the form of multiple choices. In addition to the generalized Lord's Chi-square statistics and p-value, the results of the analysis are also presented in the form of item plots as shown in Figure 1.

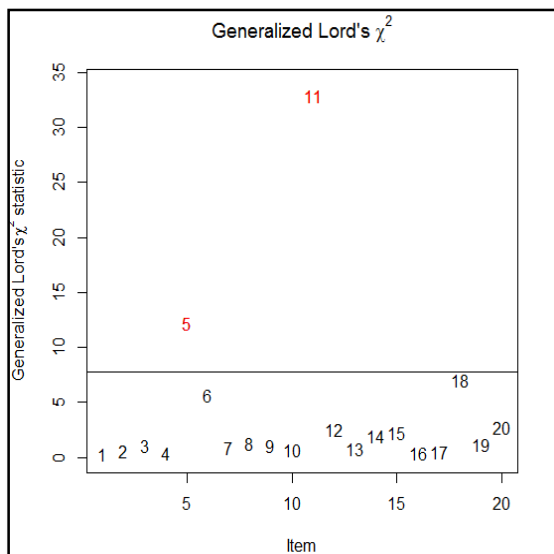


Figure 1. Generalized Lord's Chi Square Item Plot

The plot in Figure 1 provides information regarding item distribution based on bias analysis. Basically, the information in the picture is the same as the results in Table 2 in which two items are significantly detected containing DIF, namely Item 5 and Item 11. The straight line that intersects the plot is the cut-score area of 7.8147. This area serves to limit item bias. It is used to group items that contain significant bias or DIF and items that are free of DIF.

The Generalized Lord's Chi-square statistical method is an item bias detection method or DIF-based item response theory. For this reason, we can describe the item response theory curve to see the difference between items containing DIF and items that are free of DIF. The curve referred to is presented in Figure 2 and Figure 3.

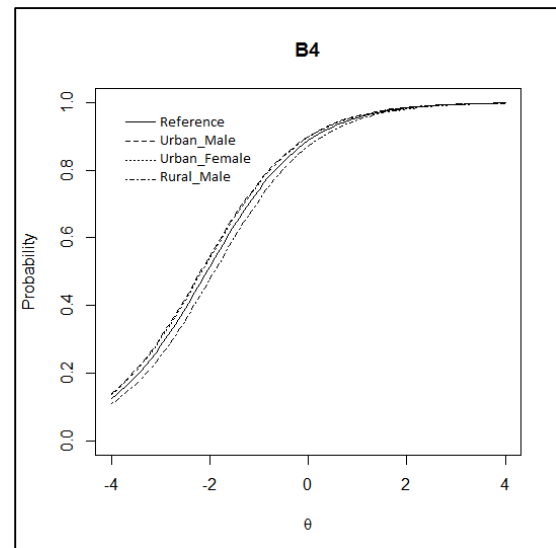


Figure 2. Plot Item 4

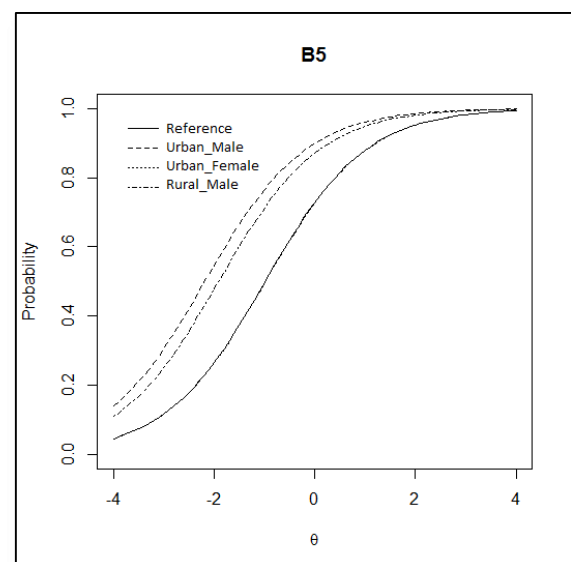


Figure 3. Plot Item 5

Figure 2 and Figure 3 show two different item characteristics: Item 4 and Item 5. Based on the previous analysis on the Generalized Lord's Chi-square coefficient and p-value, Item 4 includes items free of bias or temporary DIF, while Item 5 includes items that significantly load DIF. Figure 2 shows that in Item 4, the curve between the four groups looks like each other. It shows that the item does not contain DIF because the chances of test-takers with the same ability to answer Item 4 are also relatively the same.

Another thing is found in Item 5. From Figure 3, it can be seen that the curve

between groups is separate. It shows that Item 5 contains DIF because the probability of examinee with the same ability to answer Item 5 correctly seems different/not the same. It becomes the basis for concluding that Item 5 on the math questions of the school examination in the State Elementary School of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, if viewed from the aspect of the item, may be in unfavorable category, and need further revision and evaluation for reuse in future.

In addition to Item 5, the results of this study also found that one more item detected significantly contained DIF, namely Item 11. Based on the previous analysis on Generalized Lord's Chi-square, statistical coefficient of 32.89 with a threshold/cut-score of 7.8147 and p-value of 0.00 with α 0.05 indicates that Item 11 includes items that significantly contain bias or DIF. For more details, Item 11 is presented in the form of characteristic items, as shown in Figure 4.

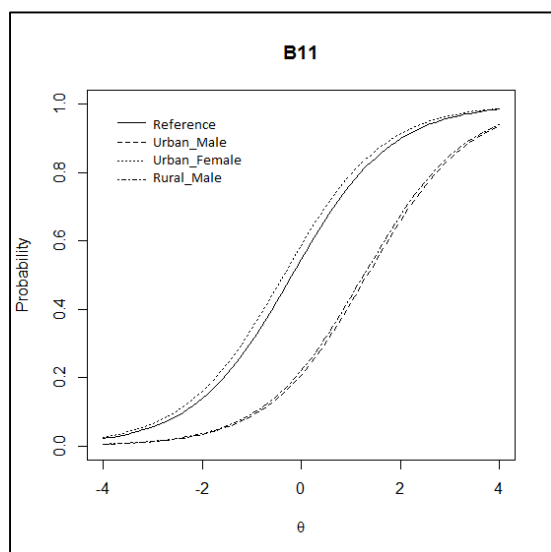


Figure 4. Plot Item 11

Figure 4 shows that in Item 11, there is a curve between separate groups. It is the basis for concluding that Item 11 on the math questions of the school exam in the elementary school of Bontomarannu District, Gowa, in the academic year of 2017/2018 if viewed from the aspect of the item, maybe in unfavorable category and need further revision and evaluation for reuse in future.

The results of this study found that from 20 math questions in the school examination in the elementary school of the District of Bontomarannu, Gowa Regency, in 2017/2018, two multiple-choice items were detected containing bias (DIF) (item 5 and item 11), while the remainder were not detected as bias items. Thus, the two biased items need further evaluation while the other 18 items are good and can be maintained to be used in the test in the future.

Discussion

The results of the study obtained empirical findings; namely, two questions contain bias or contain the DIF (Item 5 and Item 11). It is undoubtedly a material for future evaluations for developers of math questions in school exams in Elementary Schools of Bontomarannu District, Gowa. According to Retnawati (2013, p. 276), ideally, the implementation of the test is based on objectivity, transparency, accountability, and non-discrimination. If a test contains items that are in favor of a particular group, then the test is said to contain bias or contain DIF.

Item 5 is one of the items that is significantly detected to contain bias. Figure 4 shows that the item contains gender bias. The Plot Item 5 shows that the item favored the group of male students. It shows that the chances of male students answering Item 5 correctly are greater than female students, even though their abilities are the same. Further analysis can be done by observing the context of the Question Item 5. For more details, Item 5 (in Indonesian) is presented in Figure 5.

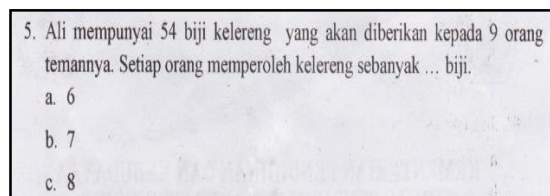


Figure 5. Question Item 5

Note: Translation for Question Item 5
 Ali has 54 marbles which will be given to 9 persons.
 Thus, each person will have ... marbles.

Item 5 contains the distribution operation material. Item 5 questions are made in the form of story questions with an alternative choice of 3 answers. Based on the story contest discussed in Item 5, it can be seen that the developer questions raised the story of the marbles to be given. Based on the results of this study, the question turned out to favor male students (see Figure 3) in the Bontomarannu District Elementary School in Gowa, in the academic year of 2017/2018. It shows that the context of the story about the game of marbles seems to be closer to male students. This is attributed to the charge bias in Item 5. According to Kim et al. (1995, p. 1), item bias is basically a condition in which there are differences in the chances of answering correctly on a question item in several groups of respondents or test participants even though the ability between several groups is the same.

In addition, Lautamo, Laakso, Aro, Ahonen, and Törmäkangas (2011, p. 223) explain that bias or often known as DIF occurs when people from different groups have different probabilities of getting a specific score on a test item. DIF analysis provides an indication of unexpected behavior based on items on the test. DIF is also basically a process of validating people's responses (test-takers). Bias can be examined by DIF analysis using a sample consisting of subgroups that differ in essential characteristics that will affect the measured phenomenon. Thus, when one or more item parameters differ between groups, in that condition the item contains DIF (bias).

Further, Karami (2012, p. 59) asserts that Differential Item Functioning (DIF) has been increasingly applied in the study of justice in psychometric circles. DIF occurs when two groups with the same level of ability, not equally capable (different opportunities) answer an item correctly. In other words, one group does not have the same opportunity to get the right item even though its members have a level of ability compared to other groups. If the factor leading to DIF is not part of the construct that is being tested, then the test is biased.

Based on this view, Item 5, which has been detected by DIF, has a different function in the sub-group analyzed, namely gender. Based on the results of empirical analysis of male groups and groups of women have different opportunities to answer Item 5 correctly, even though compared to the same ability parameters.

Beside Item 5, Item 11 is also one of the items that are significantly detected to contain bias. Item 11 in Figure 4 shows that the item contains gender bias. The Item 11 plot shows that the item favored the female student group. It shows that the chances of female students answering Item 11 correctly are higher than male students, even though their abilities are the same. Further analysis can be done by observing the context of the Item 11 question. For more details, Item 11 (in Indonesian) questions are presented in Figure 6.

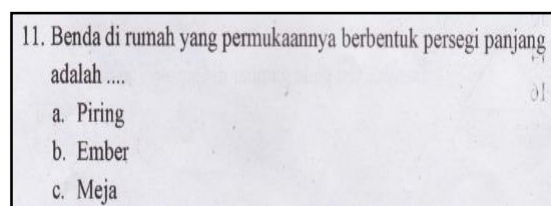


Figure 6. Question Item 11

Note: Translation for Question Item 11
An object in the house which has a rectangle-shaped surface is
A. Plate
B. Bucket
C. Table

Item 11 basically contains flat material. Based on the question theme discussed in Item 11, it can be seen that the developer of the question raised about examples of objects in the house, such as plates, buckets, and tables. Empirically, based on the results of this study, the question turned out to favor female students in elementary school of Bontomarannu District, Gowa, Academic Year 2017/2018. It shows that the context of objects in the house, such as plates, buckets, and tables seems to be closer to female students. It is attributed to the content of bias on Item 11. Bias items are important be-

cause the purpose of the DIF analysis is to detect differences in item responses in questionnaires, rating scales, or tests in different subgroups (e.g., sex) while controlling level of ability (Zhang, 2015, p. 1).

De Leo, Van Dam, Hobkirk, and Earleywine (2011, p. 570) state that the DIF charge on an item is often caused by inaccurate questions. These inaccurate questions present one group higher than the other group on certain traits because group membership is not the difference in the characteristics of the real abilities. The results of this study relate to this study that the bias found in the two items is caused by gender differences alone, not differences in the characteristics of the actual ability of the test takers. It is what needs attention in the future related to the development of better question items.

Based on the IRT model, an item displays DIF if the probability of responding in different categories varies across the groups studied, given an equivalent level of the underlying attributes. If the DIF is contained in an instrument, there are several options available. One extreme option is to remove the items from bank items, but the risk is that if the number of items omitted is large, then the condition has a disadvantage, namely, the measurement of precision and flexibility of administration of items which is blocked. Another alternative is available when the number of items with DIF is relatively small, namely making repairs or revisions to items detected loading DIF. In this case, in-depth analysis is needed regarding the causes of the existence of the items in a particular group. If it can be shown that an instrument is free from bias items, then all items in the instrument can be used to correctly estimate the value of the ability parameter (Weisscher, Glas, Vermeulen, & De Haan, 2010, p. 545).

Item 5 and Item 11 are the items that experience DIF. Basically, the two items are unwanted items on the school exam in the Elementary School of Bontomarannu District of Gowa in the academic year of 2017/2018 because they will tend to show differences in the number of subject attributes that actually do not exist. Items like this are items

that can discriminate against one group compared to another, so it needs to be repaired or excluded from the test on the next examination.

The findings of this study indicate that there is a bias in gender aspects only, while in the location (urban and rural) aspects, no significant items were found to contain DIF. It means that the location of the school (both rural and urban) does not affect the alignment of the test items. It means that each item on the exam of the Elementary School of Bontomarannu District of Gowa Regency in the academic year of 2017/2018 does not favor the sub-groups of student locations (rural and urban).

Another thing happens in the aspect of gender. The existence of items that have different treatment in terms of gender aspects causes the item detected to contain bias. It is in line with the results of research conducted by Chiesi, Ciancaleoni, Galli, Morsanyi, and Primi (2012, p. 391) that gender is one of the potential aspects that might lead to bias in measuring ability in general. Although some studies (Colom & García-López, 2002, p. 445) also state that there is no significant difference in terms of gender, but empirically this study found a difference in the probability of successfully answering items correctly even in the case of similar abilities.

The bias that occurs in the case of this study is internal bias, which is also commonly referred to as item bias. This item's bias is an aspect of the bias in the test relating to the psychometric properties of a test item and the overall test. This internal bias is focused on investigations about whether each item has similar behavior, namely the similarity in measuring psychometric properties (Adams, 1992, p. 178). If this view is related to the results of this study, it turns out that two items have behavioral inequality or tend to favor certain groups. It is what underlies that there is an interaction between group members on the performance of the items in the examination of Elementary School in the District of Bontomarannu, Gowa Regency, in the academic year of 2017/2018.

Refractive analysis of items is done to check whether each item is fair for each student (examinee) without being caused by inherent differences from the student, such as differences in sex, language, ethnicity, parental education, and others. Detection of bias items in a measuring instrument is very important, considering that the community has time to be more critical or very concerned with the results of measurements such as school exams, especially if these measurements can have a direct impact on students, such as whether a student can be categorized as failed or successful on a test.

The general approach to managing item bias is to issue any item that shows or loads the DIF of an instrument. However, maintaining all items is much better because the development of expensive and time-consuming instruments and testing is quite long. Therefore, one way to deal with items that show DIF is to correct bias by maintaining matching items with the opposite bias on an item. To match the appropriate items and address DIF items at the scale level, the direction, and type of DIF must be recognized correctly (Cho, Martin, Conger, & Widaman, 2010, p. 176). This view implies that, basically, detected items loaded with DIF are generally acted upon by removing or removing the item from the instrument. On the other hand, removing items in the instrument will interfere with the validity of the instrument's contents so that the item is fixed, and is a good alternative for energy efficiency and time for test developers. However, to correct and maintain the item, the developer must recognize the direction and source of bias correctly.

The bias content of the appraisal item is a bad thing and decreases the level of credibility of the test (assessment). If an assessment instrument is dominated by many bias items, then, the results of decisions or conclusions issued from the implementation of the assessment will also be biased, because one of the indicators of the quality of the assessment is the quality of the instrument or appraisal itself, although there is no doubt other factors besides instrument factors.

In the case under study, namely the elementary school exam instrument in the Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, it is found that in general, the instruments used are basically good. Only a small number of items detected contain bias, while most others have good quality in terms of bias. It is the basis of evaluation for items that are still biased and retain items that have been assessed as good.

Conclusion and Suggestions

Based on the results of the research conducted, it can be concluded that by using the Generalized Lord Chi-square method showing 20 items of math questions in the school examination in the elementary school of Bontomarannu District, Gowa Regency, in the academic year of 2017/2018, two items contain detected bias (DIF), which are item 5 and item 11. Meanwhile, 18 other items such as Items 1, 2, 3, 4, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20 are not detected as bias items.

Based on the conclusions obtained in this study, some suggestions are proposed. (1) To the developer of the School Examination in the State Elementary School of Bontomarannu District, Gowa Regency: to always evaluate the test instruments used every year by biased detection of items so that the quality of the instruments used annually can be guaranteed. It is considered crucial because basically, the right measurement results will produce objective decisions. (2) To analyze bias items, it is recommended to use the Generalized Lord Chi-square method, because by using this method, we can detect DIF to more focal groups so that the sub-groups analyzed can be more complex. (3) The application of the Generalized Lord Chi-square method in this study only reaches the 1-parameter model so that it is suggested to be further analyzed in the next parameter (2-parameters and 3-parameters). (4) In conducting the next research, it is recommended to use more than one method to compare accuracy.

References

- Adams, R. J. (1992). Item bias. In J. P. Keeves (Ed.), *The IEA technical handbook*. The Hague: The International Association for the Evaluation of Educational Achievement (IEA).
- Bares, C., Andrade, F., Delva, J., Grogan-Kaylor, A., & Kamata, A. (2012). Differential item functioning due to gender between depression and anxiety items among Chilean adolescents. *International Journal of Social Psychiatry, 58*(4), 386–392. <https://doi.org/10.1177/0020764011400999>
- Budiyono, B. (2009). Ketepatan metode Mantel-Haenszal, Sibtest, dan Regresi Logistik untuk mendeteksi differential item function. *Jurnal Penelitian Dan Evaluasi Pendidikan, 13*(1), 1–20. <https://doi.org/10.21831/pep.v13i1.1398>
- Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences, 22*(3), 390–396. <https://doi.org/10.1016/j.lindif.2011.12.007>
- Cho, Y. I., Martin, M. J., Conger, R. D., & Widaman, K. F. (2010). Differential item functioning on antisocial behavior scale items for adolescents and young adults from single-parent and two-parent families. *Journal of Psychopathology and Behavioral Assessment, 32*(2), 157–168. <https://doi.org/10.1007/s10862-009-9145-1>
- Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences, 32*(3), 445–451. [https://doi.org/10.1016/S0191-8869\(01\)00040-X](https://doi.org/10.1016/S0191-8869(01)00040-X)
- De Leo, J. A., Van Dam, N. T., Hobkirk, A. L., & Earleywine, M. (2011). Examining bias in the impulsive sensation seeking (ImpSS) Scale using Differential Item Functioning (DIF) – An item response analysis. *Personality and Individual Differences, 50*(5), 570–576. <https://doi.org/10.1016/j.paid.2010.11.030>
- Herwin, H. (2016). An application of the generalized logistic regression method in identifying DIF. In *Proceeding of the International Conference on Educational Research and Evaluation (ICERE)*. Yogyakarta: Universitas Negeri Yogyakarta.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2
- Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*(2), 59–76.
- Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261–276. <https://doi.org/10.2307/1435297>
- Lautamo, T., Laakso, M.-L., Aro, T., Ahonen, T., & Törmäkangas, K. (2011). Validity of the play assessment for group settings: An evaluation of differential item functioning between children with specific language impairment and typically developing peers. *Australian Occupational Therapy Journal, 58*(4), 222–230. <https://doi.org/10.1111/j.1440-1630.2011.00941.x>
- Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(2), 226–238. <https://doi.org/10.6027/0893-4152.a0000202>

- 42(3), 847–862. <https://doi.org/10.3758/BRM.42.3.847>
- Osterlind, S. J. (1983). *Test item bias. Series: Quantitative applications in the social sciences*. Beverly Hills, CA: Sage.
- Retnawati, H. (2013). Pendeteksian keberfungsian butir pembeda dengan indeks volume sederhana berdasarkan teori respons butir multidimensi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 17(2), 275–286. <https://doi.org/10.21831/pep.v17i2.1700>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.
- Scheuneman, J. D., & Bleistein, C. A. (1999). Item bias. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 220–234). New York, NY: Elsevier.
- Weisscher, N., Glas, C. A., Vermeulen, M., & De Haan, R. J. (2010). The use of an item response theory-based disability item bank across diseases: Accounting for differential item functioning. *Journal of Clinical Epidemiology*, 63(5), 543–549. <https://doi.org/10.1016/j.jclinepi.2009.07.016>
- Zhang, Y. (2015). Multiple ways to detect differential item functioning in SAS. In *Proceedings of SAS Global Forum 2015 Conference* (pp. 1–9). Dallas, TX: SAS Global Forum. Retrieved from <https://pdfs.semanticscholar.org/01f8/0a01287893f8f3f1029aa817b9cce3983901.pdf>
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation.

ANALYSIS OF THE QUALITY OF TEST INSTRUMENT AND STUDENTS' ACCOUNTING LEARNING COMPETENCIES AT VOCATIONAL SCHOOL

Nur Ichsanuddin Achmad Kurniawan
Universitas Negeri Yogyakarta

Sudji Munadi
Universitas Negeri Yogyakarta

Abstract


The study is aimed at describing: (1) characteristics of the items of the national examination try-out test of the accounting subject matter in the 2015/2016 academic year on classical test theory and modern test theory; and (2) classification of students' masteries in the learning of accounting. The study is explorative research. Analyses are conducted using the classical and modern test theories for item characteristics and descriptive quantitative for students' masteries in accounting using the test set for the national examination try-out in the 2015/2016 academic year. A total of 414 students do the Package A test. Results show that (1) based on the classical test analyses, a number of 11 items (27.5%) belong to the "easy" category, 22 items (55%) "medium" category, and 7 items (17.5%) "difficult" category allowing a total of 19 (47.5%) to be categorized as good items; meanwhile, on the modern-theory analyses, a total of 34 items (85%) belong to the "good" category. (2) Around 38% of the students have competencies of the medium and low categories. Most students have difficulty in answering questions of the higher-order thinking levels.

Keywords: *test item characteristics, accounting, learning competencies, Rasch Model*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.22484>

Contact *Nur Ichsanuddin Achmad Kurniawan*

 *nur.ichsanuddin@gmail.com*

 *Department of Educational Research and Evaluation, Graduate School of Universitas Negeri Yogyakarta*

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

Introduction

Education takes an important role in the development of human resources of a country and nation. In the Law of Republic of Indonesia No. 20 of 2003, it is mentioned that education is a conscious and planned effort to provide learning conditions and processes so that learners actively develop their potentials to acquire spiritual, religious strengths, personalities, intellects, decent traits, and skills needed by themselves, the society, the nation, and the country. National education is to function in developing the learners' awareness of their potentials for the sake of the good of the nation in the frame of educating the life of the nation.

In the frame of educating the nation, the government issued the Regulation of the Minister of National Education No. 19 of 2005 about the national standard of education (NSE). The NSE is a minimal criterion of education systems in the legal area under the state of the Republic of Indonesia. The NSE becomes the basis for the planning, implementation, and control of education to realize qualified national education. The NSE also functions to ensure the quality of national education in intellectualizing the nation and building a civilization of the nation. With NSE, it is expected that the quality of education improves.

The NSE consists of eight standards that must be achieved by all education units. These are graduate's competency, content competency, process competency, teacher's competency, facility, management, funding, and evaluation. These standards must be obeyed by teachers and school personnel in running educational programs to develop students' competencies and forming the characters and civilization of the nation. The graduate's competency standard (GCS) becomes the main reference in developing the other standards. This way, evaluation of the instructional processes must be oriented to the GCS.

Outcomes of instructional processes can be seen from the results of the students' scores in examinations. Learning outcomes are interpreted through standardized evalua-

tion processes. Many education systems still use the results of exams as an indicator of students' progress and mastery of knowledge. As a consequence, society tends to look at students achievements, mainly from final scores of the instructional activities. This view has caused students to have a burden to acquire the highest possible scores (Manoppo & Mardapi, 2014). The magnitude of students' learning outcomes obtained through evaluation processes is then regarded as a judgment for the instructional processes. Such evaluation processes cannot be separated from the assessment processes that are done using particular measuring instruments.

Evaluation is an important component in the running of education programs. Education evaluation is the quantification of phenomena or objects involved in the education process. It is expected that, through a good evaluation system, teachers can devise appropriate learning strategies, and that will motivate students to learn better. Evaluation is a tool that can be used to obtain information on the students' learning achievement.

In the Regulation of the Minister of National Education No. 19 of 2005 Chapter 63 Item (1), it is mentioned that evaluation of learning outcomes at the primary and secondary school levels consists of (a) evaluation of learning outcomes by teachers, (b) evaluation of learning outcomes by the school, and (c) evaluation of learning outcomes by the Government. In line with the advancement in the world of education, the evaluation system that is presently used is the criterion-referenced evaluation. Criterion-referenced evaluation is aimed at knowing a person's competencies on a certain criterion (Mardapi, 2012, p. 186). The criterion-referenced evaluation compares examinee's test scores with an absolute criterion determined by the teacher. So far, results of the criterion-referenced examination are pass or fail. An examinee is regarded as passing if his score is the same with or higher than the given minimal limit and failing if it is lower.

The minimal limit, more familiarly referred to as the minimal passing criterion

(MPC), is the minimum level of competency that a student has in order to be able to be declared as passing a particular education level. MPC is used to know the level of competency a student achieves. The passing label means that a student has achieved the required level of competency and failing means that a student has not.

In Sleman regency, Yogyakarta Special Region, the Business and Management Study Program of the vocational schools have had good graduates. It is shown by the fact that, between 2013 and 2015, the passing percentage of the graduates is 100%. However, the criterion is mainly passing, without information of the extent to which the graduates have the competencies of the subject matters. During the pre-survey with teachers, members of the subject matter professional group, it is found that no empirical review has been done on the levels of graduates' competencies. It is important to know the classification of students' competencies to be used as consideration in developing learning outcome evaluation. The present study is an effort to do just that.

In order to know the nature of the competency of students who take the examination, an initial effort must be made to look at the test instrument. It is a fact that, up to the present time, the test instrument that is used for the national examination in accounting is not well reviewed. A teacher in the interview stated that the test items that had just been developed for trial examination were administered right away, before being tried out first. A good test item must first go into analyses of differentiating power, difficulty level, and distractor function. This way, a student's competency can be classified into very low, low, medium, high, or very high.

Based on the preceding background, the researchers are interested in empirically attempting to look at the quality of test items of the exam and classification of the competencies of students of the accounting study program of all the vocational schools in Sleman regency. Empirical evidence is obtained by collecting responses of students

taking the try-out of the national examination of the three subject matters of the accounting subjects of the 2015/2016 academic year developed by members of the Accounting Teachers' Association of Sleman Regency.

Research Method

The study employs quantitative research approach of the descriptive explorative method. Results of the study are expected to be able to describe the quality of the test items and students' competencies in the accounting subject matter. Data were taken from students' responses in the regional test trial of the national examination developed by accounting teachers, members of the accounting teachers' professional association of Sleman Regency.

Findings and Discussion

The classical-based item analyses conducted in this study produce levels of item difficulty, discriminating powers, test reliability, and standard errors of measurement. There are 19 items accepted as good items.

Characteristics of the Test on the Classical Theory

Level of Item Difficulty

Results of the item analyses show that the levels of difficulty of the test items are found to range between 0.075 and 0.971 with a mean score of 0.556. Referring on the criterion by Crocker & Algina (1986, p. 313) and Wright & Masters (2008, p. 227), 27.5% or 11 items are of the easy category, 55% or 22 items are of the medium category, and 17.5% or 7 items are of the difficult category.

Discriminating Power

All items have positive discriminating powers, although of various degrees. It means that all the correct answer has functioned well. Most of the items can differentiate high-achieving students from low. Most of the high-achieving students choose the correct answers, while the low-achieving

students choose the distractors. Scores of the discriminating powers range from 0.032 to 0.698 with a mean score of 0.412.

The point-biserial correlation of the item analyses shows that ten items (25 %) are weakly discriminating. These ten items have a discriminating power of lower than 0.3 (Kartowagiran, 2012; Reynolds, Livingston, & Willson, 2009).

Reliability and Standard Error

The reliability index (alpha) is 0.880 with a standard error of measurement (SEM) of 2.582. It means that the test can be categorized as reliable since the alpha index satisfies the minimum line of 0.7 (Linn, 1989, p. 106; Mardapi, 2014). It is in agreement with Safrudin Amin’s study that finds a reliability index of 0.874. Meanwhile, the SEM score of 2.582 means that, on the confidence level of 95%, a student with a raw score of X will have his real score on the interval of $X \pm 2 \text{ SEM} = X \pm 5.164$.

Characteristics of the Test on the Modern-Theory Approach

IRT Pre-requisite Test

Uni-dimensionality

For the requirement of the factor analysis, the analysis sample of the study can be categorized as “good” since it is higher than 300. Williams, Onsmann, & Brown, 2003 suggest that an analysis sample should minimally be 100 or over. More specifically, it is stated that a sample of 100 is poor, 200 is fair, 300 is good, 500 is very good, and 1000 or more is excellent. Feasibility of test samples can be determined by KMO-MSA and Barlett’s test of sphericity (see Table 1).

Table 1. KMO-MSA and Barlett’s Test of Sphericity

KMO-MSA and Barlett’s Test			
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.878	
Barlett’s Test of Sphericity	Approx. Chi-Square Df. Sig.	4075.252 780 0.000	

A KMO-MSA value is regarded adequate if it passes 0.5 (Field, 2009, p. 660). Results of the study show that the KMO-MSA is 0.878. Calculations show a Barlett’s Test of the Sphericity significance level of 0.000. It means that the requirement is fulfilled since the significance level obtained is lower than 0.05.

According to Reckase (1979), Smits, Cuijpers, & van Straten (2011), and Wu et al. (2013), the uni-dimensionality assumption is fulfilled if “the first factor should account for at least 20 percent of the test variance”. The variance that can be explained amount to 57.351% and the contribution of the first factor is 19.860% (see Table 2). Since the first factor accounts for almost 1/5 of the variance test, it can be concluded that the unidimensionality assumption is satisfied.

Table 2. Eigenvalue and Component of Variance (13 Components)

Component Number	Eigen Value	Proportion	Cumulative
1	7.944	19.860	19.860
2	1.788	4.471	24.331
3	1.506	3.764	28.095
4	1.445	3.613	31.708
5	1.312	3.279	34.987
6	1.257	3.141	38.129
7	1.180	2.949	41.078
8	1.146	2.866	43.944
9	1.139	2.846	46.790
10	1.083	2.707	49.497
11	1.074	2.685	52.182
12	1.062	2.655	54.837
13	1.006	2.514	57.351

Egan, Sireci, & Swaminathan (1998) add that “if a data set is unidimensional, then the first eigenvalue should explain a relatively large proportion of the variance”. Results of the factor analysis in Table 2 show that more than 13 eigenvalues have a score higher than 1, where the first factor is the most dominant, 7.944; almost five times higher than those of the following factors which are almost equal. Since the variance scores have a linear comparison with the eigenvalue (Field, 2009, p. 652; Johnson & Wichern, 2002, p. 441) and the first factor accounts for a bigger

contribution than the other factors, then the assumption of uni-dimensionality is fulfilled.

The results of the uni-dimensionality test presented graphically in a scatter plot can be seen in Figure 1. According to Hambleton & Rovinelli (1986), as cited by Stage (2003), the number of significant factors is usually shown by the appearance of an “angle” in the plot. The scree plot in Figure 1 signifies that an angle has been formed at a point on the left side. It means that the uni-dimensionality assumption is fulfilled.

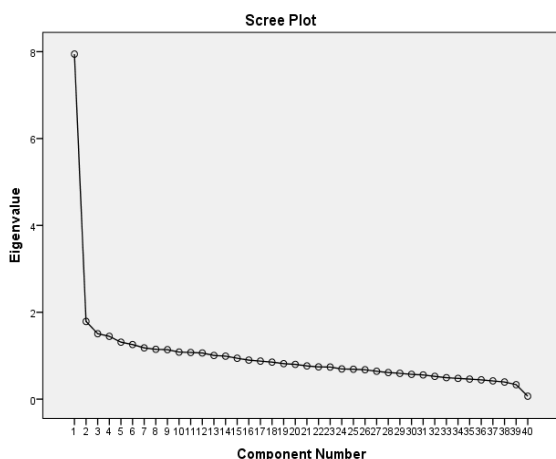


Figure 1. Eigenvalue Scree Plot

Local Independence

In general, all the elements outside the main diagonal matrix are too small, closing to zero. It can show that the local independence assumption has been fulfilled.

Parameter Invariance

Results of the item parameter estimation (in this case, levels of difficulty since the analysis uses the Rasch model) of each sample are presented in a scatter plot and correlated. Positive high correlation shows that parameter invariance is satisfied (Retnawati, 2014, p. 8). Figure 2 presents an estimation plot for item parameter invariance. From Figure 2, it can be seen that the estimate values are located relatively close to the straight line with a high correlation score (0.9881). It can be concluded then that the assumption for the item parameter invariance is fulfilled.

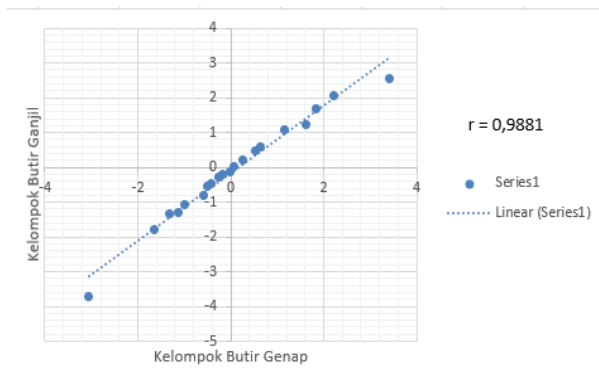


Figure 2. Parameter Invariance Plot of Item Difficulty Levels

To test the competence parameter invariance (θ), the forty test items are divided into two groups of subtests according to the results of the item parameter estimation of each sample are presented in a scatter plot and correlated. Positive high correlation shows that parameter invariance is satisfied. to the item numbers, subtest I consisting of odd numbers and subtest II even numbers (Retnawati, 2014, p. 9). Figure 3 presents the scatter plot of the competence parameter in accordance with the item groups done by the students. In Figure 3, the estimate values are located close to the straight line with a high (substantial) correlation score of 0.9989.

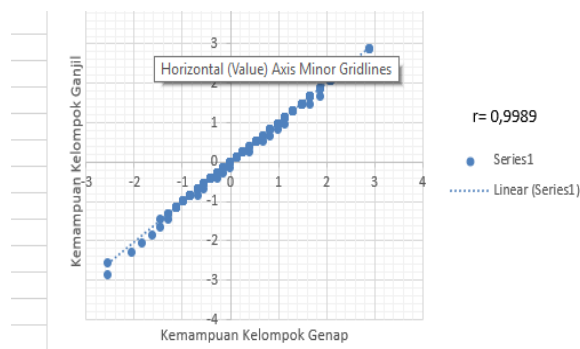


Figure 3. Parameter Invariance Plot of Students' Competencies

Instrument Characteristics

Analyses on the characteristics of the test under study include model fits, item parameter, and testees' characteristics, TIF, and SEM. Each characteristic is elaborated as follows.

Model Fit

The analysis carried out in the study makes use of the WINSTEPS program of IRT of the Rasch model. A test is regarded as fit with the item difficulty and the testees if the outfit MNSQ value is in the range of 0.5-1.5 (Linacre, 2002). Results of the study show that five items are found to be not fit with the model. These are 6, 11, 18, 31, and 40. In term of the testees, 59 students are found to be not fit with the Rasch model since they are outside the MNSQ outfit range.

Item Parameter and Testees' Characteristics

A total of 40 items and 414 students are subjected to the analyses. An item is categorized as "good" if it fulfills two requirements, namely: it has a good difficulty level (-2 logit ≤ bi ≤ +2 logit) (Hambleton & Swaminathan, 1985) and it has a model fit. In the study, six items (15%) of the total 40 are not in the "good" category. They are items 1, 2, 18, 31, 37, and 40. Item 18 has the highest difficulty level (+3.4 logit), and item 2 has the lowest difficulty level (-3.72 logit). Meanwhile, testee number 146 has the highest competency (+2.89 logit) and the testee number 89 the lowest (-2.87 logit). Figure 4 presents the distribution of the item difficulty levels. From Figure 4, it can be seen that 34 items can be accepted.

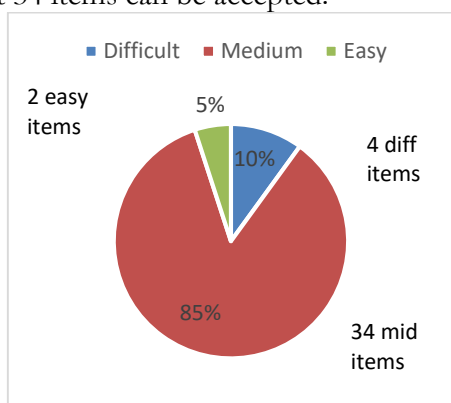


Figure 4. Distribution of Item Difficulty Levels

Test Information Function (Tif) and Sem

Results of the analyses using the Rasch model show that the test set has a maximum

information function (TIF) amounting to 16.737 on competencies around -0.2 logit. According to Hambleton (in Wiberg, 2004), a reliable test has a TIF value of ≥ 10. In the study, the test instrument can be regarded as reliable in measuring the testees' competencies in accounting. Meanwhile, SEM values have a reverse criterion from TIF. It means that the test will have a good TIF if it has the lowest SEM value (0.2444) and answered by testees with a competence level of around -0.2 logit (of the mid-high category). Using the SEM value and what has been calculated, the interval of testees' competencies can be obtained using this equation (Hambleton & Swaminathan, 1985, p. 90):

$$\theta - z \frac{\alpha}{2} [I(\theta)]^{-\frac{1}{2}} \leq \theta \leq \theta + z \frac{\alpha}{2} [I(\theta)]^{-\frac{1}{2}}$$

Since $[I(\theta)]^{-\frac{1}{2}} = SEM$ at the confidence level 95%, the formula becomes:

$$\theta - 1.96 SEM \leq \theta \leq \theta + 1.96 SEM$$

Based on this equation, it can be stated that the test will give good information (TIF) if taken by testees of the interval range of -0.678 logit ≤ θ ≤ 0.278 logits. Visualization of the test TIF and SEM is presented in Figure 5.

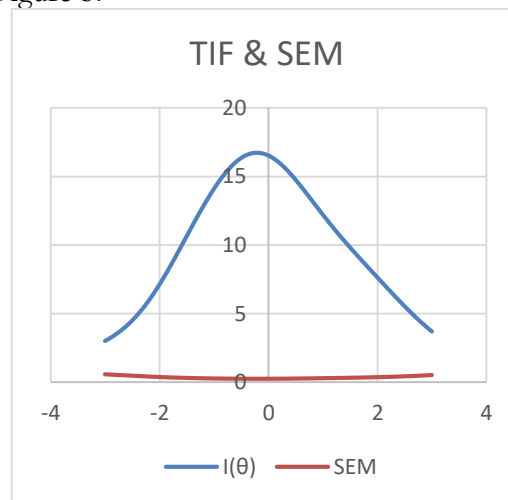


Figure 5. Relation between TIF and SEM of the Test

Classification of Students' Competencies in Accounting

Students' competencies can be graded into five categories: (1) very high, (2) high, (3) medium, (4) low, and (5) very low. Students' competencies can be seen from the

tetha measure in the analysis using Winstep software program. Prior to this, to use the Winstep, the test items that are not fit for the Winstep program are not included. Items that are not used in the analysis are numbers 6, 11, 18, 31, and 40. Results of the data analysis of the 414 students can be seen in Table 3. From Table 3, it can be seen that the very high category of students' competencies is occupied by 39% of the students, the high category 3%, the medium category 6%, the low category 7%, and the very low category 45%.

Table 3. Learning Competency Categories of the Accounting Students

Category	Number of Students	Percentage (%)
Very high	159	38
High	14	3
Medium	24	6
Low	30	7
Very low	187	45
Total	414	100

Conclusion

By the classical theory approach, it is found that the average measure of the item difficulty level is in the “medium” category, the test items have a good measure of distractor functions, and the test is reliable. Concerning the difficulty level, discriminating power of, and distractor functioning, a total of 19 items (47.5%) of the test are in the “good” description. By modern-approach analyses, it is found that the average of the difficulty level is in the “medium” category. Given the difficulty level and model fit, a total of 34 items (85%) are in the “good” category. Based on the measures of the test information function (TIF) and SEM, the test allows the best for students with a competency range of $-0.678 \text{ logit} \leq \theta \leq 0.278 \text{ logit}$. In view of the item response theory (IRT), students' competencies can be grouped into five categories; namely very high with 159 students (38%), high with 14 students (3%), medium with 24 students (6%), low

with 30 students (7%), and very low with 187 students (45%).

One implication that can be given is for the results of the study to be an input to teachers of the Accounting Teachers Professional Association in Sleman Regency in developing test items. Since students' accounting competencies in the 2015/ 2016 academic year cannot be measured maximally because of the low quality of the test, training is needed for teachers to develop and analyze test items. Use of the IRT and classical-theory analysis gives different results; caution is therefore needed in reviewing the existing tests.

References

- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.
- Egan, K. L., Sireci, S. G., & Swaminathan, H. (1998). Effect of item bundling on the assessment of test dimensionality. In *the paper presented at the annual meeting of the National Council on Measurement in Education*. San Diego, CA.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd 3d.). London: Sage Publications.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Kartowagiran, B. (2012). Penulisan butir soal. In *the paper presented in Training on Writing and Analysis of Items for the Civil Servant-Rekinpeg Resource*. Hotel Kawanua Aerotel, Jakarta.
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).

- Linn, R. L. (1989). *Educational measurement*. New York, NY: Macmillan.
- Manoppo, Y., & Mardapi, D. (2014). Analisis metode cheating pada tes berskala besar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 18(1), 115–128. Retrieved from <https://journal.uny.ac.id/index.php/jpep/article/view/2128/1773>
- Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.
- Mardapi, D. (2014). Authentic assessment. In *the paper presented at HEPI Conference*. Denpasar, Bali.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207–230. <https://doi.org/10.3102/10769986004003207>
- Regulation of the Minister of National Education No. 19 of 2005, on National Standard of Education (2005). Republic of Indonesia.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.
- Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.
- Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1), 147–155. <https://doi.org/10.1016/j.psychres.2010.12.001>
- Stage, C. (2003). *Classical test theory or item response theory: The Swedish experience*. Santiago, Chile: Centro de Estudios Públicos.
- Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test*. Stockholm: Umea Universitet.
- Williams, B., Onsmann, A., & Brown, T. (2003). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 8(3), 1–13. Retrieved from <https://ajp.paramedics.org/index.php/ajp/article/view/93/90>
- Wright, B. D., & Masters, G. N. (2008). *Rating scale analysis: Rasch measurement*. Chicago, IL: Mesa Press.
- Wu, Q., Zhang, Z., Song, Y., Zhang, Y., Zhang, Y., Zhang, F., ... Miao, D. (2013). The development of mathematical test based on item response theory. *International Journal of Advancements in Computing Technology*, 5(10), 209–216.

CHINA'S K-12 TEACHER QUALIFICATION SYSTEM

Dani Surya Lee

Southwest University, Chongqing China


Abstract

A strict teacher qualification system is vital to ensure a steady stream flow of high-quality human resources. In 1993, the People's Republic of China established the Teacher Law demanding that teachers must have legal status and qualifications. It has put into effect a system of qualification that could determine the readiness of K-12 teacher candidates to be practitioners of education. This country's teacher qualification system has experienced changes and adaptations until it has reached the current form in 2016's Teacher Qualification Examination, as more than a tool merely for formality or evaluation of rote learning, but an actual assessment of pre-service teachers' knowledge and abilities to conduct classroom teaching in authentic settings. This paper introduces this qualification examination system, while also mentioning some important and distinguishable aspects of teachers' situations in China, in the hope that it will shed some light on how this Asian giant is currently conducting its education.

Keywords: *teacher qualification; K-12 teachers; education in China*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.24065>

Contact *Dani Surya Lee*

 danilee009@hotmail.com

 Faculty of Education, Southwest University, Chongqing
No. 2 Tiansheng Road Beibei District, Chongqing 400715, P.R. China

Introduction

Teacher quality plays more roles in affecting student achievement than any other school factors. Effective teachers may enable students to overtake their peers who were learning from less effective teachers. On the other hand, learning with poor-quality teachers has a long-lasting residual effect, which may be problematic to compensate in later grades even when studying under effective teachers (Looney, 2011).

Teachers must possess the required knowledge and skills. It is a general requirement that is probably the same in every part of the world, albeit different specifics. As Looney (2011) states, high-quality teachers are intellectually and verbally able; have good knowledge on the subjects they are teaching; are skilled in many diverse teaching methods; and are skilled assessors of their students. Chinese scholars and researchers have also stated that teachers should have the corresponding vocational knowledge, as well as the education and teaching skills concerning teaching design, teaching implementation, and teaching evaluation, and also able to correctly use them in conducting teaching practice (Cheng, 2016; Fu, 2003; Han, Zhong, Liang, Peng, & Chen, 2017). These skills must be continuously enhanced with considerations towards teaching practice aspects such as subject management, lesson preparation, class management, feedback reflection, etc. (Ministry of Education of the People's Republic of China, 1993).

On the other hand, teachers must be mentally healthy. Vesely, Saklofske, and Leschied (2013) insist that Emotional Intelligence (EI) is critical in contributing to the prevention of occupational stress and burn-out while improving a teacher's classroom management, therefore, helpful in mitigating the effect of stress. Cheng (2016) also argued that a teacher must own a sense of professional ethics and psychological literacy which must come from a healthy mind.

The People's Republic of China is a nation growing rapidly. Its economic development is undeniable and is the subject of many discussions and research. However, its

development cannot be separated from how it conducts education, which, unfortunately, received very little attention, especially in how it selects its primary school teachers.

Like most countries, China has adopted a nine-year compulsory education system, including infant school education, primary education, secondary education, and higher education (Ministry of Education of the People's Republic of China, 1986, 1995). Through this research, we hope to introduce China's K-12 teacher qualification system used to meet its demand for effective teachers, with a focus on the new teacher qualification examination. It is hoped that this paper may serve as a base of reference for future research, and for the practical development of teacher qualification as well.

Research Method

The instruments used are documentary analysis, expert discussions, and interviews. The documentary analysis comprises of a comprehensive literature study of English journals, Chinese journals, and relevant research reports. The interviews include discussions with two experts in the field of Chinese education – teachers of the Faculty of Education in a well-known university in China, as well as text-interviews with three Chinese teachers and teachers-in-training that have undergone the old or new teacher qualification exam. This paper serves as a qualitative research report where the phenomenon is first generally discussed with an expert, followed by collecting individual stories and extensive documentary analysis. Another discussion with an expert followed, and finally, the result, i.e., the qualification system, is reported chronologically in order to create a comprehensive report on how the teacher qualification method works in China.

Findings and Discussion

Chinese Teacher Education

At the beginning of the last century, China has already had specialized schools to train teachers, and teachers were treated as

public officials, paid by the state. As China's population grew out of control, however, the state was unable to bear the huge cost of education. Minimizing educational costs, the extensive use of non-teacher graduates or non-academically qualified teachers occurs. After the adoption of a socialist market system in the 1990s, educational development was strongly influenced by market forces and had two striking features: quantitative growth and qualitative enhancement (Li, 2010).

In 1986, the state's teacher service system already started to provide different grades of professional teachers, along with the technical titles and requirements. In 1993, the Teacher Law established a statutory teacher qualification system, demanding that teachers must have legal status and qualifications. This system became the national implementation of a statutory career permit system. According to this law, teachers who had not received professional education to teach must enroll in a professional academic training institution and obtain the certificate of accreditation. The Chinese Ministry of Education (MoE) (2009) published its English version of educational laws, including Teacher Law.

The categories and proportion of curriculum structure of most normal schools in China (Chen, 2014) are: (1) Public courses, e.g., politic, physical education, English, etc. This category accounted for approximately 15%. (2) Pedagogy courses, e.g., theories of education, theories of psychology, the practice of education, etc. accounted for approximately 6 to 10%. (3) Professional (or discipline) courses, according to each student's preference. This category accounted for approximately 70%. (4) Fieldwork, or education practice for six to eight weeks.

In 2001, the state launched a fully implemented teacher qualification system through a provincial qualification examination. In 2012, the Ministry of Education gradually abolished the provincial Teacher Qualification Examination and began to implement the national Teacher Qualification Examination system with a distinguishable

content and procedure, from mere rote learning to a more authentic setting of teaching practices (Liu & Bai, 2016; Shi, 2017). The newest change, however, was in 2015 when the exam was made a mandatory requirement even for graduates of normal schools, and in 2016 when the content and the process were changed significantly again to emphasize actual teaching ability in classrooms (Cheng, 2016; Zeng, 2016).

There is a very distinct requirement in China for teacher qualification. After their graduation from normal schools, teacher candidates must apply and pass the teacher qualification examination given by or acknowledged by the state, and obtain the accreditation in the corresponding level and subject from the state's accreditation system for teachers (Fu, 2003; Gao, 2005).

To apply for the qualification examination, applicants must first already have obtained the necessary academic qualifications. The following is an example from the website of Chongqing Shi Jiaoyu Kaoshi Yuan [Chongqing Educational Examination Institute] (2017): Graduate of an infant normal school, college graduate or higher for kindergarten; college graduate or higher for primary school teachers; and, bachelor's degree or above for junior and senior high school.

However, in fact, the demand is much higher than this. Most of today's teachers in China are at least 4-year graduates of normal schools or are universities graduates, and lots of them have master degrees. Part of this unwritten requirement was established due to the tight competition of workforce in China. The higher the degree, the higher the possibility of securing a good job or position. Another thing to consider is that it is not actually mandated for people who want to work as a teacher to have a normal school graduation certificate, but, the ownership of a certificate would give him/her a better chance in procuring a job in better schools.

The Current Condition of Chinese Teachers

In recent years, there is a significant improvement in the status of general K-12 teachers. China's MoE has formally written

the education law that the whole society shall respect teachers. This regulation, however, has no practical use, as Chinese culture encourages the people to base respect on tangible aspects such as position, wealth, achievements, and seniority. Instead, the improvements in teachers' condition come from the increase in salary and other benefits.

The Teacher Law in China stated that the average salaries of teachers should not be lower than the average salaries of the local civil servants. Currently, K-12 teachers' average base salary is approximately between 36,000 to 45,000 RMB net annually (Mei, 2017). This number is a fixed amount within the total salary. In addition to the base salary, teachers' total salary generally include 3 elements as follow: Allowance for length of service, given as a part of the monthly total salary, amount to 60 RMB per year of service (e.g., if a teacher has taught in a school for three years, then in the fourth year his monthly total salary will have an additional 180 RMB); in-class hourly salary, given as a part of the monthly total salary for hours they teach students in classrooms; and performance pay, given per semester and also annually, can amount near to 10% of the total salary. The tax rate for teachers does not differ from other professions and will be deducted from the monthly salary before given to the teachers.

Other than the four components of the salary mentioned above, teachers are also entitled to other benefits. For example, the State's required schools to give teachers pocket money called hongbao (lit. translated into "red envelope") on their birthdays with the sum of money as much as 100 RMB as a symbol of consideration and good intention. There are also welfare benefits prescribed by the State; housing supports given by some schools, as well as the approximately three months of leave with pay in winter and summer vacations.

From our discussion with experts, we have found that an essential reform of teachers' welfare is the fact that now it is more dependent on the individual schools' policy instead of the State's in providing additional

teacher benefits. There is, however, a decrease in teachers' welfare for some good schools (i.e., schools that has obtained higher accreditation). For example, before, good schools may give salary up to 10,000 RMB for K-12 teachers, not including housing supports and other benefits, now there is no such privilege. This decrease is balanced by the State's law of a higher general salary for all teachers. An expert argued that even though the welfare seemed to decline at some schools, due to the broader spread of equality among teachers' salary, there is an increase in overall welfare.

However, the most attractive thing about being a teacher in China is job stability. There is basically no teacher layoffs, provided that they perform their duties as prescribed by the law. Among other things, political awareness and ideological level are crucial aspects of teacher evaluation in China. We have confirmed through our discussion with experts that fulfilling these two aspects would secure the position while failing would mean termination despite any achievements or success.

We can see from the discussion above that due to the improvement of the overall status of teachers, improved treatment, occupational stability, basically no laid-off, and other advantages, the attraction of being a professional educator in China has become more popular compared to the period before the qualification system based on the Teacher Law was established.

Challenges in the Teacher Aspect of Education

Even after the method of qualification was established, there were still some challenges to be faced in the teacher aspect of education. Fu (2003) listed some of these challenges. However, following the rapid development of China's education system, some of them have been currently well-addressed. Through our discussion with experts in China's education, we have triangulated whether currently these challenges have been addressed or not and whether they can be a cause for concern.

Fu (2003) stated that many primary and secondary school teachers had not received vocational training. It is not true today. As the qualification exam has evolved to be closer to an authentic assessment that actually focuses on teacher's practical ability, it is almost impossible to pass the exam without proper vocational training. Fu says many teachers choose to learn some easy-to-obtain qualifications to improve academic level and meet academic requirements. The disciplines of what they have learned, however, is inconsistent with the subject they are or will be teaching. For example, Chinese literature graduates teaching physics, Politics graduates teaching English, Math graduates teaching computers. This problem still exists, especially in the rural area, but rare for higher education institutions (there are still some in universities, but they happen only between sub-divisions courses, not interdisciplines). Experts argued that it should not be a problem. In China, this issue is seen mainly in the primary and secondary level of education.

As the current Chinese teachers are well equipped to teach these basic subjects or disciplines, regardless of what their advance academic specialties are, the quality of the teaching and learning process should not be much influenced by it. In other words, if the teacher can master the new subjects and teach them effectively to the students, then he/she can still be considered an effective teacher despite his/her original subject credentials. Besides, Fu argued that there is an uneven number of teachers in rural areas compared to urban areas. It also currently still happens. However, through the discussions with experts, it is found that, currently, there is a trend that many teachers prefer to work in the rural area instead of the big cities, especially for older teachers. In the same province, teacher salary is more-or-less the same. The school location, whether it is in the urban or rural area, does not have any influence. The stress level and demand of the job, however, is obviously different. This equal level of salary plus a lower level of demand has increased the attraction of being a teacher in the rural area.

The Law Aspect of Teacher Requirement

In China, the law aspect of an educator's job is a truly vital one. Failure in obeying the law shall result in a restriction to obtain qualifications for teachers as well as the forfeit of already obtained qualifications.

Written in the law, teachers must abide by the constitution, laws and professional ethics; raise their ideological level and political consciousness, and carry out schools' teaching plans, fulfill teaching contracts and accomplish educational and teaching tasks (Ministry of Education of the People's Republic of China, 1995).

According to China's Ministry of Education in Teachers Law of the People's Republic of China, teachers can conduct educational and teaching activities and experiments, as well as engage in scientific research and academic exchanges, join professional academic societies and fully express their views in academic activities. They are also entitled to put forward opinions and suggestions regarding education, teaching, management of schools and the work of the administrative departments of education, and to participate in the democratic management of schools through congresses of teachers, staff and workers, or through other forms (Ministry of Education of the People's Republic of China, 1993).

It is important to remember that the expressions of their opinions and suggestions must always be parallel to the law of China and the view of the Chinese political party that they align themselves to. There are currently eight institutional minor-parties in China, but only one major party, which is the Communist Party of Chinese with the President of China as its General Secretary. Practitioners of the field stated that a teacher might choose not to align him/herself with any political party, but joining a party may give some advantages such as access to some jobs or positions that may otherwise be unavailable. Whether they are a member of a political party or not, however, expressing, supporting, or teaching the opposing views could still result in termination.

In dealing with the students, according to the Teacher Law, teachers can guide students in their studies and development and evaluate their works and academic achievement. It is also emphasized in the law that the teachers are to be paragons of virtue and learning in terms of ideology i.e., (the State's) morality, culture, science, and technology, at the same time, pay attention to their individual differences, teach students on the basis of their aptitude and promote full development of students. Thus, from the discussion above, it is indicated that Chinese teachers must be able to guide students to balance their individualism's creativity and nationalism at the same time, to make them grow but keep them in line, to inspire students to be brilliant, not for themselves, but for the good of the country as a whole.

Teacher Qualification Examination

Mandated tests measure the competence of teacher candidates by assessing their knowledge of best teaching practices and their knowledge of the skills which are needed for being an effective teacher (Goodman, Arbona, & de Rameriz, 2008). Gao (2005) stated that China's teacher qualification system serves to improve the quality of teachers, and functions as an essential element to facilitate teacher standardization, as well in establishing teacher training and education.

China's MoE indicates clearly that all Chinese citizens within legal working age, who abide by the constitution and laws, take a keen interest in education, have sound ideological and moral character, possess a record of formal schooling as stipulated in the Teacher's Law and have passed the national teachers' qualification examinations may obtain qualifications as teachers.

National teacher qualification examination has become the only standard for obtaining K-12 teacher qualification (Cheng, 2016). Since 2015, all of those who apply for kindergarten, primary school, junior high school, senior high school, secondary vocational school teacher qualification, and secondary vocational school practice instructor qualification, are required to take the exam-

ination. Even students of normal schools are no longer directly qualified as teachers after graduation and must take the exam as well. This requirement, however, does not influence graduates of normal schools before the rules were applied.

The professional titles of K-12 teachers engaging in compulsory education that can obtain through the qualification examination are classified into pre-school, primary, intermediate and senior (Zhongguo Jiaoshi Zige Wang [China's Teachers Qualification Network], 2008; Zhongguo Jiaoyu Kaoshi Wang [China Educational Examination Network], 2017). Each candidate can choose the exam classification according to the educational level where he/she is planning on teaching. By obtaining the pre-school qualification, teachers are qualified to teach in kindergartens; primary qualification is for primary schools; intermediate qualifications, junior high schools; and senior qualifications, senior high schools. Candidates need to pass the exam only once to work as an educator in the corresponding level (periodical evaluation or assessment process will take over afterward). A teacher that obtains a higher level of classification can also work in the level lower than his/her qualification, but not the other way around.

All Chinese citizens who abide by the constitution and laws, whether they are normal school or non-normal school graduates, may take the National Teacher Qualification Examination as a condition to procure the teaching positions. Zeng (2016) stated that the exams take place twice a year.

Gui (2014) argued that in the past, China's teacher qualification exam had only focused on "rote" ability of the test. It only examined the pedagogy and psychology aspect of teacher candidates without testing their subject knowledge and teaching abilities, which makes any professional person can apply for teacher qualification. However, now, the exam assesses applicants' practical knowledge and analytical thinking, instead of merely a simple memory ability. It gives the graduates of normal schools a huge advantage.

Since 2016, there is a huge change in the way the examination is being conducted and the content of the examination (Zeng, 2016). The current examination is divided into two parts: written examination and interview examination. The participants have to grasp basic teaching knowledge and skills, as well as the corresponding knowledge of the discipline they want to teach (Chen, 2014; Han et al., 2017).

To prepare for the exam, candidates must familiarize themselves with two basic books: *Theory of Education Examination* and *Principles of Educational Psychology Examination*. These books are developed by the Personnel Division and Examination Center of the Ministry of Education, printed by East China Normal University Press. The books are publicly marketed, easy to obtain, affordable, and are distinguished into three levels: for primary school, middle school (junior and high school), and higher education teacher candidates. Many training centers offer courses and supplement books to prepare candidates for the exam, which also included the contents of the above mentioned two mandatory books.

In the new system of 2016, however, there is additional knowledge that candidates must master to pass the exam: the comprehensive quality of a teacher, teaching knowledge and ability, education regulations and policies, and teacher professional ethics. For primary and higher levels of qualification,

candidates must also partake in one among a few subject-related exams according to their teaching focus, such as Chinese language, and literature, mathematics, physics, or other subjects-related knowledge. All of these will be included in the written exam according to the level of qualification (Zeng, 2016).

There is also a significant reform in the examination nature. Before, the emphasis is on rote memorization of *Educational Theories and Principles of Educational Psychology*, the two books mentioned above. Now, although rote memorization of those principles still has its place, the national examination pays more attention to the comprehensive quality and practical ability, focusing on the candidates' ability to solve the problems of teaching practices. In the written exam, other than basic principles, the qualification process also focuses on specific case studies and the application of basic educational principles, so that the candidates in their daily work life are proven able to identify the problem, ask questions, and solve the problem. In the interview examination, the focus is on the candidates' classroom teaching designs, teaching practice, as well as incident handlings, contingencies and other aspects of a teacher's comprehensive ability (Cheng, 2016; Shi, 2017; Zeng, 2016).

Figure 1 explains the overall workflow of the National Teacher Qualification Exam. The following are the details concerning the individual parts, as described in Figure 1.

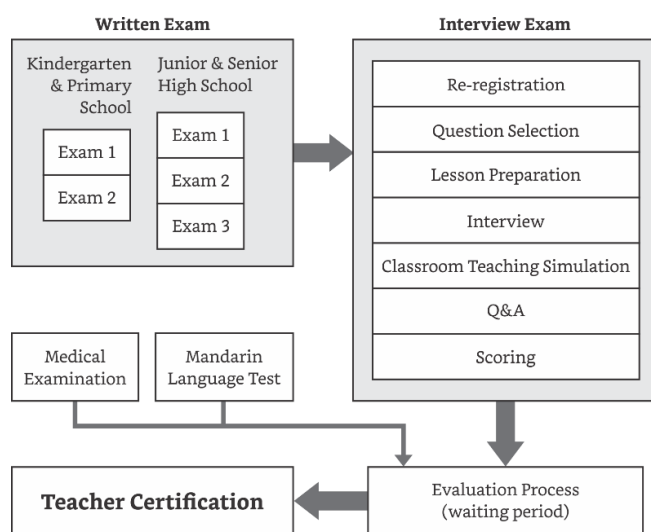


Figure 1. China's National Teacher Qualification Exam

Written Examination

In the former version of the teacher qualification examination, Theory of Education and Principles of Educational Psychology have separate examinations, with separate outlines and knowledge points. The new exam does not separate pedagogy and educational psychology but integrates them into parts of the exam in the assessment process, which better reflects the real ability of the teacher candidates.

Table 1 describes the difference in the content of the written exam of different K-12 examination levels (Zhongguo Jiaoyu Kaoshi Wang [China Educational Examination Network], 2017). Pen and paper are used in the written examination. The allotted time is 120 minutes. The tests include multiple-choice questions, short answer questions, and essays. The content, other than theories, are questions concerning case analysis, the practice of principles, teaching and activity designs, case studies, etc.

Passing all exams within the written examination segment of the qualification process is the prerequisite to take the interview exam. There are 2 to 3 tests within this segment, depending on the level of qualification. If an applicant fails a test (or tests) in the written exam, then he/she can choose to make up for the failed test(s) in the next term. However, applicants must now pay attention that after the exam's reform in 2016, the result of the individual tests will be valid only for two years. The failure to pass the

whole exam and obtain the certification within these two years will result in a total repetition of the whole process (Shi, 2017; Zeng, 2016).

Interview Examination

Structured interviews and multiple authentic assessment methods, such as lesson preparation and classroom teaching simulation, were employed in this segment. The main assessment covers the applicant's professional ethics, psychological quality, mental health, manners, speech and expressions, basic literacy of teaching and teaching design ability, implementation of the teaching plan, teaching evaluation ability, and other basic teaching skills. Further, Chongqing Shi Jiaoyu Kaoshi Yuan [Chongqing Educational Examination Institute] (2017) describes the workflow of the interview exam process of K-12 teacher qualification on their website. Each stage is described as follows.

Re-registration

Candidates arrive on the site on time with the interview ticket and ID, then proceed into the waiting room. The ticket can be printed from the MoE Examination Center's website a week before the examination date.

Question Selection

Candidates log in the software. Then, they were given a randomly-selected topic by the computer, give confirmation, and then receive a printed copy of the topic.

Table 1. Content of the Written Exam of Different K-12 Levels of the Teacher Qualification Exam

Level of Examination	Written Exam			Interview Exam
	Exam 1	Exam 2	Exam 3	
Kindergarten	Comprehensive Quality	Teaching Knowledge and Ability	—	Education and Teaching Practice Ability
Primary School	Comprehensive Quality	Educational Teaching Knowledge and Ability	—	Education and Teaching Practice Ability
Junior High School	Comprehensive Quality	Educational Knowledge and Ability	Subject Knowledge and Teaching Ability	Education and Teaching Practice Ability
Senior High School			Subject Knowledge and Teaching Ability	Education and Teaching Practice Ability

Lesson Preparation

The candidates go into the preparation room and then, using pen and paper, prepare a lesson plan. The duration of this period of preparation is 20 minutes.

Interview

A panel of examiners randomly selects two questions from the question bank. Candidates answer both questions. The duration is five minutes.

Classroom Simulation

Candidates demonstrate their teaching ability in practice according to the prepared lesson plan. The presentation should be done within 10 minutes.

Q&A

Examiners ask questions concerning the teaching demo. The duration is five minutes.

Scoring

Examiners give the candidates a comprehensive score for the whole process – success or failure. Then, the examiners submit the score into the interview evaluation system.

The board of examiners consists of three people, and are composed of experts from colleges and universities, as well as excellent teachers in primary and secondary schools and kindergartens. They are required to take provincial or higher educational examination institutions' training.

Mandarin Language Test and Medical Examination

Mandarin Language Test acts to confirm the candidate's qualification, as there might be some teacher candidates whose daily language are still heavily influenced by their local dialects. The law of China stated that all teachers must teach using standardized Mandarin or Putonghua. For Mandarin Language teacher candidates, there will be a higher standard for this test. The Mandarin

test can be taken in a different period from the qualification examination, even years before, but must be obtained before the qualification certificate can be given.

Physical test also acts to confirm the qualification. It is especially crucial for kindergarten teacher candidates. However, different from the language test, the physical test must be taken within six months apart from the qualification exam. Some municipalities even required a period no longer than three months.

Evaluation of the Result and Certification

The MoE will examine the result, and after a waiting period, candidates can know their scores and qualification through the internet. Scores for the individual tests of written examination will be kept for two years. The MoE will also provide the Teacher Qualification Examination Certificate.

Important to note that, according to the new regulation of 2016, this certification is valid for three years (Shi, 2017), means that if a candidate does not engage in the field of education within three years after he/she received the certificate, then the certificate will be invalid or expired.

Conclusion

In recent years, we can see a significant improvement in the status, job treatment, occupational stability, and other advantages for teachers in China. The country's qualification system has demanded more from teacher candidates than it used to be, but at the same time promises fair compensations for the fulfilling of those demands through a general increase in teachers' welfare.

Chinese National Teacher Qualification Examination has become the only standard for obtaining teacher qualification. It is divided into two parts: written examination and interview examination, and have different contents according to the level of qualifications. The focus of the written exam is comprehensive educational knowledge and case studies to test how the candidates would identify the problem, do the thought process, and solve the problem. The focus of the

interview examination is on candidates' mental health, classroom teaching design, and preparation, as well as other teaching practice aspects that are related to the profession.

By improving the overall quality of normal school students, or K-12 teacher candidates, we are ensuring the quality of future school teachers. The implementation of the teacher qualification system in China provides an effective personnel system guarantee to direct effective and qualified future teachers into the proper places and positions in society. With the addition of the demand for Chinese teachers to teach their students to think not only of him/herself but of the whole country, this system would ensure the continuous stream of quality human resource to support the growth of the nation as the whole.

References

- Chen, Q. H. (2014). Jiaoshi zige kaoshi zhidu gaige gei xiaoxue jiaoshi zhiqian peiyang dailai de tiaozhan yu jiyu [Challenges and opportunities brought by the teacher qualification examination system reform for primary school pre-service teacher training]. *Jiaoshi Jiaoyu Luntan*, (27), 54–56.
- Cheng, Y. (2016). Jiaoshi zigezheng guokao dui xueqian jiaoyu zhuanke rencai peiyang de tiaozhan yu jiyu [The challenges and opportunities of the national teacher qualification exam for talent cultivation in pre-school education major]. *Jiangxi Guangbo Dianshi Daxue Xuebao*, (4), 89–91.
- Chinese Ministry of Education. Teachers law of the People's Republic of China (2009). People's Republic of China. Retrieved from http://en.moe.gov.cn/Resources/Laws_and_Policies/201506/t20150626_191385.html
- Chongqing Shi Jiaoyu Kaoshi Yuan [Chongqing Educational Examination Institute]. (2017). Chongqing shi 2017 nian shang bannian zhong xiaoxue jiaoshi zige kaoshi mianshi gonggao [Primary and secondary school teacher qualification examination interview notice of Chongqing city for the first half of 2017]. Retrieved from <http://www.ntce.cn/html1/report/1704/932-1.htm>
- Fu, Y. (2003). Jiaoshi zhuanke hua jincheng de lishi huigu yu jiaoshi zige zhidu de zhongyao zuoyong [The historical review of teacher specialization and the important role of teacher qualification system]. *Chengdu Shibfan Gaodeng Zhuanke Xuexiao Xuebao*, (22), 41–43.
- Gao, K. (2005). Jiaoshi zige renzheng zhidu de zuoyong ji qi wanshan [The function and other excellence of teacher qualification system]. *Zhiye Shikong*, (4), 22–23.
- Goodman, G., Arbona, C., & de Rameriz, R. D. (2008). High-stakes, minimum-competency exams: How competent are they for evaluating teacher competence? *Journal of Teacher Education*, 59(1), 24–39. <https://doi.org/10.1177/0022487107309972>
- Gui, W. L. (2014). Woguo jiaoshi zige kaoshi xinjiu bishi dagang duibi yu fenxi [Comparison and analysis of China's current and former teacher qualification written examination]. *Kaoshi Yanjiu*, (3), 86–91.
- Han, C. H., Zhong, W. M., Liang, F., Peng, R., & Chen, Z. L. (2017). Jiaoshi zige kaoshi tizhi gaige xia shifansheng mianlin de tiaozhan he yingdui cuoshi [Challenges for teacher-training school students and the countermeasures under the reform of teacher qualification examination system]. *Jiaoyu Guancha*, (6), 133–137.
- Li, Y. (2010). Quality assurance in Chinese higher education. *Research in Comparative and International Education*, 5(1), 58–76. <https://doi.org/10.2304/rcie.2010.5.1.58>

- Liu, D. M., & Bai, Z. F. (2016). Jiaoshi zigezheng gaige beijing xie shifan jiaoyu zhuan ye jiaoxue gaige de sikao [Reflections on the teaching reform of the education in teacher-training schools under the background of teacher qualification reform - study case of Yulin College]. *Yulin Xueyuan*, (26), 91–93.
- Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement. *European Journal of Education*, 46(4), 440–455. <https://doi.org/10.1111/j.1465-3435.2011.01492.x>
- Mei, F. (2017). Xiaoxue laoshi gongzi yiban duoshao [The general wages for primary school teachers]. Retrieved from <http://www.yjbys.com/wage/251172.html>
- Ministry of Education of the People's Republic of China. Compulsory Education Law of the People's Republic of China (1986). Retrieved from http://en.moe.gov.cn/Resources/Laws_and_Policies/201506/t20150626_191391.html
- Ministry of Education of the People's Republic of China. Teachers Law of the People's Republic of China (1993). People's Republic of China. Retrieved from http://en.moe.gov.cn/Resources/Laws_and_Policies/201506/t20150626_191385.html
- Ministry of Education of the People's Republic of China. Education Law of the People's Republic of China (1995). People's Republic of China. Retrieved from http://en.moe.gov.cn/Resources/Laws_and_Policies/201506/t20150626_191385.html
- Shi, C. Y. (2017). Jiaoshi zige 'guokao' beijingxia youhua tijiao zhuan ye rencai peiyang moshi yanjiu [A study on the model of optimizing talent training in physical education under the background of 'the national examination']. *Jilin Tiyu Xueyuan Xuebao*, (33), 90–93.
- Vesely, A. K., Saklofske, D. H., & Leschied, A. D. W. (2013). Teachers—The vital resource: The contribution of emotional intelligence to teacher efficacy and well-being. *Canadian Journal of School Psychology*, 28(1), 71–89. <https://doi.org/10.1177/0829573512468855>
- Zeng, H. (2016). Xin jiaoshi zigezheng 'guokao' de tedian ji zhongyao yingxiang fenxi [An analysis of the characteristics and important influence of the new teacher qualification certification]. *Dangdai Jiaoyu Shijian Yu Jiaoxue Yanjiu*, (4), 230–231.
- Zhongguo Jiaoshi Zige Wang [China's Teachers Qualification Network]. (2008). Jiaoshi zige tiaoli [Teacher certification regulations]. Retrieved from http://www.jszg.edu.cn/portal/policy_regulation/whole_policy?id=232
- Zhongguo Jiaoyu Kaoshi Wang [China Educational Examination Network]. (2017). Kaoshi jieshao [Introduction to exams]. Retrieved from <http://www.ntce.cn/html1/folder/1507/1181-1.htm>

EVALUATING THE IMPLEMENTATION OF ENGLISH COMMUNICATION THERAPY (ECT): AN OBJECTIVE STRUCTURED CLINICAL ASSESSMENT (OSCA) APPROACH

Nursanti Dwi Yogawati
Universitas Negeri Yogyakarta

Widibastuti
Universitas Negeri Yogyakarta

Abstract


This study is aimed at evaluating the implementation of the English Communication Therapy (ECT) instructional program oriented to the Objective Structured Clinical Assessment (OSCA) for the students of the D-3 Nursing Study Program at STIKES Al-Irsyad Al-Islamiyyah, Cilacap. Three research problems are put forward concerning (1) the fit of the program in terms of the variables context, input, process, and product; (2) obstacles faced by students; and (3) attainment of the ECT learning objectives. The study is evaluation research using the qualitative and quantitative CIPP model. The research subjects are 168 students of the Nursing Study Program of STIKES Al-Irsyad Al-Islamiyyah. Data were collected using interviews, questionnaires, and documentation. The findings show that (1) evaluation of the four variables of context, input, process, and product shows that ECT can be stated as in the good or fit category; (2) main obstacles faced by students are lack of vocabulary and low level of self-confidence; and (3) students are able to achieve the passing grade of 74 at each station on the competency of English speaking about nursing matters.

Keywords: *CIPP research model, English communication therapy, OSCA*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.22449>

Contact *Nursanti Dwi Yogawati*

 *shantyyogawati@gmail.com*

 *Department of Educational Research and Evaluation, Graduate School of Universitas Negeri Yogyakarta*

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

Introduction

The Health Institute of Al-Irsyad Al-Islamiyyah is a health-based educational institute in the city of Cilacap, Indonesia, that has seven study programs. The institute offers two education levels of Diploma 3 (D-3) and Bachelor level (also known as Strata-1 [S-1]). There are four study programs in Diploma 3 level, there are Nursing, Midwifery, Pharmacy, and Physiotherapy study programs. In the Bachelor level, there are Medical Aid, Pharmacy, and Nursing study programs.

The vision and mission of the Institute are becoming Islamic, and superior and advanced in the global era with goals of producing graduates who are qualified and competitive. The institute expects the students to have hard and soft skills to support their competencies in facing the world of work. The D-3 program in Nursing is the most preferred study program that has most work in producing graduates that are competitive not only in the country but also overseas.

Competency assessment is an evaluation of nurses' competencies, which is, nowadays, is conducted by both the Computer-Based Testing (CBT) and Paper-Based Testing (PBT), using the method of Objective Structured Clinical Assessment (OSCA). Development of the preparation of nurses to take the competency assessment is done by the OSCA stream since this method has the best way to measure nurses' competencies up to the phase when examinees perform their competencies (the show how phase) (Masfuri et al., 2016). The OSCA method includes the evaluation of knowledge, communicating skills, physical examination skills, diagnostic analysis, and interpretation skills, knowledge, diagnostic skills, and interpersonal relation skills (Oermann & Gaberson, 2009).

The OSCA method of assessment is given in the curriculum with the expectation that the students are prepared for the assessment. The syllabus includes cognitive, affective, and also psycho-motoric skills of nurses to be able to carry out therapeutic

communication with the patients (Yanti & Pertiwi, 2008).

The OSCA is a fair assessment model since every examinee receives the same exam materials. It is structured seeing that specific clinical skills are tested using specific evaluation sheets. Each station is constructed in a way that resembles the real clinical conditions. For each station, specific time allocations are determined. The examinees go through the stations consecutively. In each station, the examinees are given questions or tasks to answer or demonstrate, graded by the assessors. Subjects that are included in the OSCA examination are Medical Surgery Nursing, Child Nursing, Psychic Nursing, Maternity Nursing, Community Nursing, Emergency Unit, Islamic Nursing, Nursing Basics, and Physiology Anatomy.

In 2013, the D-3 program of Nursing, STIKES Al-Irsyad Al-Islamiyyah Cilacap, added one subject matter, namely English. According to Sahraini and Madya (2015), development of English education in many forms and places has motivated the D-3 Nursing study program to require students to be able to speak not only in the daily language but also in a foreign language, in this case, English. Competency in English speaking will help graduates develop their career when working abroad. In the English syllabus, the students receive English for Specific Purposes (ESP), where students are trained to communicate in English about nursing. The study program expects that graduates will be able to work not only in the domestic regions but also in international places. The ESP subject in the D3 Nursing study program is called English Communication Therapy (EST). It is aimed at helping students and graduates conduct two-way communication events with foreign patients in all situations, especially during clinical checks.

Education is a conscious and planned effort to provide learning situations and processes so that learners will be able to develop their potentials actively to acquire spiritual, religious competencies, self-management, personalities, intellectuality, noble

behaviours, and practical skills needed by themselves, the society, the nation, and the country (Law of Republic of Indonesia No. 20 of 2003). One of the behavioral changes is that in language proficiency. Acquisition of language, however, is attained by steps (Batang, 2014), influenced by the interaction between the learners and their external environment. In language learning, the learners will be expected to master aspects of the language (pronunciation, grammar, and vocabulary) and the four language skills reading, writing, listening, and speaking (Cambridge Assessment English, 2014).

According to Sahraini and Madya (2015), the learning of teaching of English in the field of language learning has passed through vast development and changes. Brown (2007, p. 7) states that learning can be defined as “showing or helping someone to learn how to do something, giving instruction, guiding in the study of something, providing with knowledge, causing to know or understand”. Concerning program evaluation, Irambona and Kumaidi (2015) stress that “it is understood that program evaluation is the use of different social research methods to check the effectiveness of social programs, find out if the programs have been implemented as planned by the society, organization, or by the government”. In this relation, the learning and teaching processes include showing and helping the learner how to do something, give instructions, make an analysis, prepare something, and understand the effectiveness of learning management and evaluation.

According to Yanti and Pertiwi (2008), the components needed to prepare for conducting OSCA are (a) station, post or place for the testing of examinees’ knowledge and skills; (b) assessor team, examinees who work in the station to observe and evaluate examinees’ demonstrations using check lists, (c) simulation patient, somebody who is trained well to play a role of patients on designed scenarios, specifically and realistically, usually in the form of cases; (d) fasilitator, person or unit who prepares materials and tools for demonstrations and provides things

needed by the examinees; (e) timer, person who is in charge of moving the examinees from station to station using a bell making sure that examinees movement is on time; (f) time allocation, amount of time given to examinees to do the demonstrations of each case, considering the examiner’s judgement and level of difficulty of the task; (g) cleaning service, person similar to facilitator, who is in charge of keeping the place clean and conducive during the examination; and (h) grader, person who calculate the scores according to the standard guide provided by the assessing body to produce the decision whether or not an examinee passes the exam in each station. Passing is decided on not in the form of a total score, but on the score of each station.

The OSCA method is regarded to be more valid, reliable, and objective, compared to the conventional aural case examination in assessing clinical competencies, communication skills, and personalities and behaviors. The OSCA reliability level is dependent upon the number of stations. With six stations, in 90 minutes each, the reliability coefficient ranges only between 0.5 and 0.6. On the other hand, with 40 stations, in a total of 4 hours, the coefficient ranges up to 0.8 (Yanti & Pertiwi, 2008).

OSCA could be used as a valid and reliable instrument for assessing the nurses’ skills if it used for selection or training of nurses with standard patients (McWilliam & Botwinski, 2012). A number of weaknesses of the OSCA method, however, are the small number of the stations in order to obtain adequate, reliable information about performances, the limited time allocation, the medium-level checklists that are not too easy nor too difficult, patients that are not standard, examiners who give scores not using the provided standard guide, and administering problems such as the noisy rooms and unorganized staff aids. The OSCA scenario may not be able to simulate an actual and ideal clinical situation because OSCA is quite expensive, needs a great number of logistics, and takes much time.

Therefore, this study is aimed at describing the English Communication Therapy (ECT) instructional processes using the Objective Structured Clinical Assessment (OSCA) method for students of the D-3 Nursing program, Al-Irsyad Al-Islamiyyah Cilacap. The study is an evaluation research using the CIPP (Context, Input, Process, Product) model. The study is expected to be able to give an evaluation to the ECT in terms of: (1) degree of fit between the instruction program and the context, input, process, and product (CIPP) system; (2) obstacles faced by students in attending the program; and (3) attainment of the objectives of the program.

Research Method

The study is evaluation research using the descriptive, qualitative, and quantitative approach. The research design refers to the CIPP evaluation model which was developed by Stufflebeam. This model was chosen for the reason that it is more comprehensive than other evaluation models. The four components to be used to evaluate the ECT program include context, background, and objectives of the ECT program; input, students’ preparation and materials of the ECT program; proses, implementation of the ECT program; and product, results or achievement of the instructional process.

The subjects of the study were all of the 168 students of the D-3 Nursing Study Program. Data were collected by interviews, questionnaires, and documentation. The re-

search instruments were developed based on operational definitions of the indicators.

Content validity was computed on the Aiken’s V formula (Azwar, 2012, p. 113) with a criterion for a valid item of ≥ 0.70 (Sireci & Geisinger, 1995, pp. 246–247). The results of the Aiken computation with the Explanatory Factor Analysis (EFA) on SPSS software showed a figure of 0.869. with a Determinant of the correlation matrix of 0.003. The following analysis using the KMO-MSA gave a score of 0.748 and Barlett’s Test of Sphericity of 926.668 at the degree of significance 0.000. Based on these statistics, it was decided that data analyses could be done. Further, it was found that, from the communality measures, all the variable had a score of $> 50\%$. It was concluded that all variable explained the factors and, from the total variance, a total of 8 constructed gave a contribution to the validity at the figure of 67.083%.

Reliability was estimated by Intraclass Correlation Coefficients (ICC) with a bottom line of 0.4 (Fleiss, 1999, p. 7). Results of the ICC analysis showed a figure of a single measure of 0.607, meaning that the reliability level was in the “good enough” category.

For the questionnaire, data were analyzed using the qualitative analysis technique. The first step was categorizing the data into the level of tendencies. Reference for score interpretation can be seen as in Table 1 (Mardapi, 2012, p. 162).

Table 1. Reference for Score Interpretation

No	Category Norm	Interpretation
1	$X > \bar{X}_i + \frac{1}{2} S_{bi}$	Highly fit/very good
2	$\bar{X}_i \leq X < \bar{X}_i + \frac{1}{2} S_{bi}$	Fit /good
3	$\bar{X}_i - \frac{1}{2} S_{bi} \leq X < \bar{X}_i$	Medially fit/good enough
4	$X < \bar{X}_i - \frac{1}{2} S_{bi}$	Not fit /poor

Notes:

X = scores obtained

\bar{X} = mean $\frac{1}{2}$ (highest ideal score + lowest ideal score)

S_{Bi} = standard deviation $\frac{1}{6}$ (highest ideal score - lowest ideal score)

Findings and Discussion

Findings

Results of the analyses on the students' achievement of the ECT can be seen in the following presentations.

Context Evaluation

The highest score obtained by the 168 students is 715, and the lowest score is 678. The ideal mean can then be computed $\frac{1}{2}(715 + 678) = 696.5$ with a standard deviation of $\frac{1}{6}(715 - 678) = 6.2$. (see Table 2).

Meanwhile, frequency distribution of the context component variable can be seen in Table 3. These score results show that a

total of 80 student respondents (47,6%) state "highly fit", 64 respondents (38,1%) state "fit", and 24 students (14,3%) state "not fit".

Input Evaluation

The highest score of the 168 students is 709, and the lowest is 646. The ideal mean is then $\frac{1}{2}(709 + 646) = 677.5$, and the ideal standard deviation is $\frac{1}{6}(709 - 646) = 10.5$ (see Table 4).

In addition, the frequency distribution of the scores of the input evaluation can be seen in Table 5. It can be seen that 59 student respondents (35.1%) state "highly fit", 72 students (42.9%) state "fit", 26 students (15.5%) state "fit enough", 10 students (6%) state "not fit", and 1 student (0.6%) state "highly not fit".

Table 2. Ideal Mean and Standard Deviation of the Scores of the Context Evaluation

Mean	Ideal Mean	Ideal St. Deviation	Max Score	Min Score
698.5	696.5	6.16	715	678

Table 3. Frequency Distribution of the Scores of the Context Component Variable

	Context			
	Freq.	%	Valid %	Cum. %
Not fit	24	14.3	14.3	14.3
Fit	64	38.1	38.1	52.4
Highly fit	80	47.6	47.6	100.0
Total	168	100.0	100.0	

Table 4. Ideal Mean and Standard Deviation of the Scores of the Input Evaluation

Mean	Ideal Mean	Ideal St. Deviation	Max Score	Min Score
677.7	677.5	10.5	709	646

Table 5. Frequency Distribution of the Scores of the Input Component Variable

	Input			
	Freq.	%	Valid %	Cum. %
Highly not fit	1	.6	.6	.6
Not fit	10	6.0	6.0	6.5
Fit enough	26	15.5	15.5	22.0
Fit	72	42.9	42.9	64.9
Highly fit	59	35.1	35.1	100.0
Total	168	100.0	100.0	

Process Evaluation

Of the 168 student respondents, the highest score is 717 and the lowest score of 629. The ideal mean $\frac{1}{2}(717 + 629) = 673$, and the ideal standard deviation is $\frac{1}{6}(717 - 629) = 14.6$ (see Table 6).

Frequency distribution of the scores of the input component variable can be seen in Table 7. For this component variable, it is clearly presented that a total of 66 student respondents (39.3%) state “highly fit”, 71 students (42.3%) state “fit”, 25 students (14.9%) state “fit enough”, and 6 students (3.6%) state “not fit”.

Product Evaluation

The results of the product evaluation show that, of the 168 student respondents, the highest score is 609, and the lowest score is 502. The ideal mean is then $\frac{1}{2}(609 + 502) = 555.5$, and the ideal standard deviation is $\frac{1}{6}(609 - 502) = 17.8$ (see Table 8).

Frequency distribution of the scores of the product component variable can be seen in Table 9. It can be seen that 24 student respondents (14.3%) state “highly fit”, 82 students (48.8%) state “fit”, 37 students (22%) state “fit enough”, 16 students (9.5%) state “not fit”, and 9 students (5.4%) state “highly not fit”.

Table 6. Ideal Mean and Standard Deviation of the Scores of the Process Evaluation

Mean	Ideal Mean	Ideal St. Deviation	Max Score	Min Score
678.6	673	14.6	717	629

Table 7. Frequency Distribution of the Scores of the Process Component Variable

	Process			
	Freq.	%	Valid %	Cum. %
Not fit	6	3.6	3.6	3.6
Fit enough	25	14.9	14.9	18.5
Fit	71	42.3	42.3	60.7
Highly fit	66	39.3	39.3	100.0
Total	168	100.0	100.0	

Table 8. Ideal Mean and Standard Deviation of the Scores of the Product Evaluation

Mean	Ideal Mean	Ideal St. Deviation	Max Score	Min Score
562.1	555.5	17.8	609	502

Table 9. Frequency Distribution of the Scores of the Product Component Variable

	Product			
	Freq.	%	Valid %	Cum. %
Highly not fit	9	5.4	5.4	5.4
Not fit	16	9.5	9.5	14.9
Fit enough	37	22.0	22.0	36.9
Fit	82	48.8	48.8	85.7
Highly fit	24	14.3	14.3	100.0
Total	168	100.0	100.0	

Discussion

Based on the preceding presentation of the results of the study, a discussion can be made as follows. (1) The context evaluation gives a situation that can be regarded as good or fit. In the interviews, students state that there is relevance between the vision and missions of the Institute and the learning-teaching process of the ECT. It is seen that one objective of the D-3 program is creating a campus atmosphere that supports learning processes that global-oriented. (2) The input evaluation is also categorized as being good or fit. In this component evaluation, the objects of evaluation include students' preparedness, plan for learning material/cases and learning support services. Students are active and aware of the existence of ECT in the OSCA as a support to the improvement of students' competencies in aural communication. The instructional materials and cases in the ECT are developed on the basis of the curriculum and designed in accordance with the students' needs to improve their competencies. The availability of learning aids and facilities has supported the instructional processes of the ECT. (3) The process evaluation also shows results that are good or fit. It can be seen from the fact that students are prepared, instructional materials are developed and presented clearly, time management is in accordance with the amount of the materials/cases, and the learning and teaching processes follow the guides given in OSCA. It can be stated that the implementation of ECT runs in a conducive situation. Students' understanding of the roles helps in supporting the learning processes such that obstacles can be minimized and supports can be maximized. (4) The product evaluation is in the good or fit category. It is seen from the score mean of 562.1, which is higher than the success category of 555.5. For graduation, students are declared as graduating if they obtain the passing grade of 74 at every exam station. It also applies to ECT. Students are regarded as competent if they can communicate in English as a nurse.

The obstacles in the implementation of ECT are uncovered from the interviews with students. Among others, obstacles that are often mentioned by students are: (1) over-nervousness experienced by students, (2) low motivation caused by inability to converse in English, (3) pessimism in their thought that ECT is not relevant to nursing, (4) students' low level of vocabulary mastery, and (5) low level of students' self-confidence to communicate in English.

From the questionnaires and students' evaluation scores, it is shown that students acquire scores that are the same or higher than the passing grade. It can be concluded that students do well in the ECT program. Students' competencies to communicate in English can be regarded as quite good as they can communicate in English with other people in clinical matters that they master. From the point of view of the OSCA, all students can fulfill the demand of the passing grade determined by the study program. Students' scores are consistent with their performances in the evaluation sheets which carry weighted points. It can be concluded that students can fulfill the requirements of the ECT program.

Conclusion

Based on the research problem, theoretical reviews, data analyses, and research findings, four items of conclusion can be drawn as follows: (1) implementation of the ECT for the students of the D-3 Nursing Study Program at STIKES Al-Irsyad Al-Islamiyyah Cilacap can be categorized as good or fit for the four components of context, input, process, and product; (2) the most prominent obstacles is students' lack of vocabulary inventory so that they are not confident enough to speak English; (3) attainment of the ECT learning objectives can be seen from students' abilities to achieve the passing grade of 74. Students' competencies are in the form of their ability to communicate in English about nursing matters; (4) implementation of ECT with the foci of context, input, process, and product carries in an implication for students in the form of

support to students to be able to communicate not only in their mother tongue but also in a foreign language.

References

- Azwar, S. (2012). *Reliabilitas dan validitas*. Yogyakarta: Pustaka Pelajar.
- Batang, B. L. (2014). Communicative competence and language learning styles of prospective teachers of English. *Jurnal of Arts, Science & Commerce*, 5(4), 182–187.
- Brown, H. D. (2007). *Principle of language learning and teaching* (5th ed.). New York, NY: Pearson Education.
- Cambridge Assessment English. (2014). Framework competency statements. In *Cambridge English Teaching* (pp. 1–11). Cambridge: Cambridge Assessment English. Retrieved from <http://www.cambridgeenglish.org/images/172992-full-level-descriptors-cambridge-english-teaching-framework.pdf>
- Fleiss, J. L. (1999). *The design and analysis of clinical experiments*. New York, NY: John Wiley & Sons.
- Irambona, A., & Kumaidi, K. (2015). The effectiveness of English teaching program in senior high school: A case study. *REiD (Research and Evaluation in Education)*, 1(2), 114–128. <https://doi.org/10.21831/reid.v1i2.6666>
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).
- Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.
- Masfuri, Sriyono, Maria, R., Budiati, T., Irawati, D., Nursanti, I., ... Hamzah, A. (2016). *Panduan penyelenggaraan OSCE keperawatan*. Jakarta: PPNI, AIPNI, AIPViKI, LPUK-Nakes.
- McWilliam, P. L., & Botwinski, C. A. (2012). Identifying strengths and weaknesses in the utilization of Objective Structured Clinical Examination (OSCE) in a nursing program. *Nursing Education Perspectives*, 33(1), 35–39.
- Oermann, M. H., & Gaberson, K. B. (2009). *Evaluation and testing in nursing education*. New York, NY: Springer.
- Sahraini, S., & Madya, S. (2015). Model evaluasi internal kompetensi guru bahasa Inggris (Model_EIKGBI) SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 19(2), 156–167. <https://doi.org/10.21831/pep.v19i2.5576>
- Sireci, S. G., & Geisinger, K. F. (1995). Using subject-matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement*, 19(3), 241–255. <https://doi.org/10.1177/014662169501900303>
- Yanti, & Pertiwi, H. W. (2008). *OSCA: Panduan praktis menghadapi UAP DIII kebidanan*. Yogyakarta: Mitra Cendikia.

LEARNING OUTCOME OF MATHEMATICS AND SCIENCE: FEATURES OF INDONESIAN MADRASAH STUDENTS

Kusaeri

Universitas Islam Negeri Sunan Ampel Surabaya

Ali Ridho

Universitas Islam Negeri Maulana Malik Ibrahim Malang

Abstract


This study aims to analyze the learning outcome of mathematics and sciences in the Indonesian National Examination from 2015 to 2018 of MTs (Islamic secondary school). The results of the analysis are used as the basis for making alternative policy as a possible way to improve the quality of mathematics and science learning. With the sample size of 360, 371 MTs students in East Java, the achievement was measured by using 40 multiple choice questions with each having four answer options. Split-plot and latent profile analysis of the data show that there was a consistent decrease in the achievement of mathematics and sciences of the moderate group of MTs from 2015 to 2018 with a dramatic drop in mathematics and tendency to drop for science. The fundamental implication of these findings is that there is a need for learning innovations to improve mathematics and science learning outcomes for 21st century learning. The findings can also provide data support for decision making for the revision of the mathematics and science curriculum and can be further used as empirical evidence for the developed countries in finding ways to improve the quality of mathematics and science learning outcomes for 21st century learning, in particular, to assist the developing countries such as Indonesia.

Keywords: *madrasah, mathematics, science, national examination*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.24881>

Contact *Kusaeri*

 *kusaeri@uinshy.ac.id*

 *Department of Mathematics Education, Faculty of Tarbiyah and Teacher Training,
Universitas Islam Negeri Sunan Ampel Surabaya
Jl. Ahmad Yani No. 117, Surabaya, 60237, Jawa Timur, Indonesia*

Introduction

The 4.0 industrial revolution has become an important topic in education in Indonesia. This is evident from the refinement of educational curriculum policies at the level of primary, secondary and higher education to accommodate important ideas from the industrial revolution 4.0 (Ibda, 2018; Subekti, Taufiq, Susilo, Ibrahim, & Suwono, 2018). Therefore, students must be equipped with 21st century skills to prepare them for the competition in the era of the industrial revolution 4.0. The 21st century skills that students must possess include scientific creativity and innovation, critical thinking, problem solving, literacy, and also metacognitive and collaborative skills (Care, Griffin, & Wilson, 2018; Isnawati, Indana, & Susantini, 2018; Jatmiko et al., 2018; Kusaeri & Aditomo, 2019; Pandiangan, Sanjaya, & Jatmiko, 2017; Siswanto, Susantini, & Jatmiko, 2018; Suyidno, Nur, Yuanita, Prahani, & Jatmiko, 2018; Wahab, Mahmud, & Tiro, 2018).

Madrasah students in Indonesia have not acquired such 21st-century skills well (Kusaeri, 2018). One of the most visible indicators is student's weak scientific ability and creativity in solving mathematics and science problems. For example, in the last four years of the national examination (NE), madrasah students were worse off when they did the mathematics and science problems that required them to think critically, creatively and to use high-level thinking (Ministry of Education and Culture, 2018). This problem certainly calls for special attention to find a solution so that madrasah students in Indonesia can compete in the era of industrial revolution 4.0.

The possible cause for such low achievement in the two subjects assessed in the NE is the heavy study load of madrasah students. They are required to accomplish not only general subjects (as learned by public school students) but also Islamic lessons (Qur'an, *hadith*, jurisprudence, *aqidah akhlaq* (Islamic creed and ethics), *tasawuf* or *tariqah* (path of spiritual development), history of Islamic culture, *nahwu*, *sorof* (Arabic terms)

and *balagha* (rhetoric) (Lukens-Bull, 2010). The learning process at the madrasahs also contributes to such low achievement. Generally, madrasah students tend to be encouraged to remember or memorize the contents of textbooks, especially in madrasahs, managed under the auspices of or affiliated with *pesantren* (Barkey, 2007). This learning method seems to be incompatible with the principles of mathematics and science learning that mainly require the inquiry process of learning and finding concepts (McClure, 2009).

Madrasahs in Indonesia are always interesting to study because Indonesia is one of the countries with the largest Muslim population in the world (Woodward, 2015). Madrasahs serve as a place of learning and instillation of Islamic values to millions of young Indonesian Muslims. Among the 49,402,000 students in Indonesia, 6,422,260 students study at madrasahs. It indicates that madrasahs have a strategic role in building and coloring the mindset of Indonesian's young generation.

Madrasah has also attracted researchers across the globe. Broadly speaking, studies on madrasah can be categorized into three groups. First, studies have focused on historical reviews of madrasahs and have been intensively carried out since the attack on the New York World Trade Center (WTC) tower, September 11, 2001. Since then, researchers such as Al-Hasani, Ismail, Kazeemkayode, and Elegu (2017), Asadullah and Chaudhury (2016), Lukens-Bull (2010), McClure (2009), Nizah (2016), Rao and Hossain (2011) have paid attention to the madrasah education system. They explored the madrasah curriculum because madrasah as an education institution is assumed to be a place to instill radical Islam for Muslim younger generation. Second, studies are concerned about madrasah management problems such as the study by Ahid (2010). The studies signified that madrasahs with all their limitations leave several problems starting from a limited budget, low student input, underqualified teachers, and limited learning facilities. The studies also spotted how each madrasah transforms to minimize their limi-

tations. Third, studies have concerned about the achievements of madrasahs. A study by Asadullah and Chaudhury (2016) compares the female and male students ability in mathematics between madrasah students and public schools in Bangladesh. The study of Ali and Furqon (2016) portrays the abilities of Indonesian madrasah students in the NE. Other researchers have concerned about the dynamics of achievement of madrasah students in the context of the NE in the last three years (Kusaeri, 2016, 2018).

The aforementioned studies, however, have mainly discussed the achievement of Indonesian madrasah student in general context and have not specifically focused on student achievement in mathematics and science. Also, previous studies, such as those by Ali et al. (2011) and Ali and Furqon (2016), were carried out within a limited year or a small sample. This calls for a more thorough study on MTs student achievement on specific subject matters tested in the NE in the multi-year period. Therefore, more detail and robust empirical data can be collected to understand madrasah student achievement in Indonesia better. This research, hence, portrays the trends in mathematics and science achievement of MTs students (MTs in Indonesia is equivalent to grade 7-9 of junior high school) and describes the profile of NE participants in the multi-year period for the last four years based on MTs status (public-private), number of students, mathematics and science scores.

The focus on MTs is particularly important as the results of the NE at this level will in part become the basis for students' enrolment to the higher level of MA (Islamic senior high school) or SMA (general high school) (Kusaeri, Aditomo, Ridho, & Fuad, 2018). Low scores in the NE will make it difficult for the students to be enrolled in the school or madrasah they expect. Such situation has encouraged MTs students to be more seriously working on the NE. Hence, the results of the NE in both subjects can highly reflect the student's real abilities. The study of the achievements of MTs students also becomes important in the context of

evaluating the quality of learning in MTs (Ali & Furqon, 2016). Ultimately, which MTs needs more policy support than others can also be more accurately identified.

This research evaluates the MTs student score in mathematics and science in Indonesian NE in the years of 2015-2018. The evaluation is focused on finding alternative solutions and policy to improve the quality of education in Indonesia, especially in the fields of mathematics and science education for MTs students. This evaluation research is expected to draw empirical evidence that provides scientific contributions to developed countries to improve the quality of mathematics and science learning outcomes in the 21st century, particularly to assist the developing countries to improve their education.

The research questions (RQ) of this study are: (1) what is the MTs (state vs. private) student achievement on mathematics and natural sciences across the examination years? (2) what is the profile of the madrasahs participating in the NE based on the MTs status (state-private), number of students, examination year (2015, 2016, 2017, 2018), mathematics scores, and science scores? (3) what is the alternative solution to the problem related to the mathematics and science learning outcomes found in points 1 and 2?

Research Method

Type and Sample of Research

This study is a policy research. The sample of this study was MTs in East Java province. In 2018, there were 998,072 students from all over Indonesia taking part in the NE, spread over 17,009 institutions. Most MTs (20.49%) were in East Java Province (Ministry of Education and Culture, 2018). Considering the big number of MTs in the province, East Java was selected as the sample province. Based on the comprehensiveness of the data obtained continuously from 2015 to 2018, the number of MTs students in East Java involved in this study is 360,371 comprising of 39,834 students (in

2015), 40,380 (in 2016), 39,740 (in 2017), and 40,417 (in 2018). The students are spread in 630 MTs (46 public, 584 private), 2015 (M = 173.04, SD = 83.23), 2016 (M = 172.89, SD = 82.40), 2017 (M = 171.20, SD = 87.88), and 2018 (M = 181.76, SD = 84.22).

Instrument and Procedures

The mathematics and science scores are the results of MTs students in the NE from 2015 to 2018. The mathematics and science questions tested in the NE from 2015 to 2018 consist of 40 multiple-choice items with four answer choices, parallel from year to year. Thus, the results can be used to compare the quality of the schools, students across different schools, or students across different districts/cities or provinces.

Data Analysis

In answering RQ-1, split-plot analysis techniques were used (Pituch & Stevens, 2016). In this analysis, the status of private-state madrasahs is between variables, while achievement of the scores in both subjects from year to year becomes within variables. RQ-2 was answered through Latent Profile Analysis (LPA) (Gibson, 1959). The model

was estimated by MPlus Version 7 (Muthén & Muthén, 1998-2015). The score for each variable was converted into a standardized score, *z*. A series of procedures were used to determine the number of profiles, following Meyer, Stanley, and Parfyonova (2012) using Akaike Information Criteria (AIC) (Akaike, 1987), Bayesian Information Criteria (BIC) (Schwarz, 1978), Sample-Adjusted Bayesian Information Criterion (SABIC) (Sclove, 1987), Bootstrapped Likelihood Ratio Test (BLRT) (McLachlan & Peel, 2000), entropy (Celeux & Soromenho, 1996; Flaherty & Kiff, 2012), and likelihood ratio test p-value LMR adjustment (aLMR) (Lo, Mendell, & Rubin, 2001). The smaller the value of AIC and BIC means the more suitable the model (Raftery, 1995). Further, high entropy (close to 1) shows a more fit model (Ramaswamy, Desarbo, Reibstein, & Robinson, 1993).

Findings and Discussion

Findings

The study used 630 madrasahs as the sample. Based on this sample, Table 1 presents exposure to mean, standard deviation, and the correlation of all variables.

Table 1. Mean, Standard Deviation, and Pearson Correlation for All Measures

No	Variable	M	SD	1	2	3	4	5	6	7	8	9	10	11
1	Science 2015	63.23	64.24	-	-	-	-	-	-	-	-	-	-	-
2	Science 2016	64.10	66.42	.964*	-	-	-	-	-	-	-	-	-	-
3	Science 2017	63.08	67.54	.952*	.970*	-	-	-	-	-	-	-	-	-
4	Science 2018	64.15	70.55	.927*	.943*	.958*	-	-	-	-	-	-	-	-
5	Math 2015	44.80	17.82	.0452	.0333	.0416	.0378	-	-	-	-	-	-	-
6	Math 2016	43.14	11.16	.0024	-.0121	-.0073	-.0118	-	-	-	-	-	-	-
7	Math 2017	39.76	7.62	.114*	.106*	.103*	.104*	.383*	.475*	-	-	-	-	-
8	Math 2018	37.48	5.51	.171*	.176*	.178*	.176*	.134*	.187*	.468*	-	-	-	-
9	Science 2015	45.84	8.82	.198*	.197*	.195*	.187*	.149*	.097*	.163*	.202*	-	-	-
10	Science 2016	43.88	7.57	.232*	.226*	.232*	.224*	.0741	.083*	.196*	.231*	.816*	-	-
11	Science 2017	43.15	5.22	.217*	.208*	.224*	.228*	.0338	.019	.244*	.240*	.710*	.751*	-
12	Science 2018	43.23	4.46	.284*	.285*	.285*	.277*	-.0642	-.034	.110*	.275*	.229*	.363*	.437*

*Statistically significant at $p < .05$; $n = 630$

The Achievement in Mathematics and Science of State and Private MTs in the NE Across the Years

In general, the trend of average NE scores tends to decline (Figure 1). The decline from year to year was proven to be multivariate, $F(6, 623) = 22.49$, partial $\eta^2 = .178$, $p < .05$. Analyzed from the status of the madrasah, there were also differences in scores between state and private madrasahs, $F(2, 627) = 27.09$, partial $\eta^2 = .080$, $p < .05$. In addition, the interaction between madrasah status (public-private) and the NE scores across the year (2015, 2016, 2017, 2018) was also proven to be multivariate, $F(6, 623) = 9.62$, partial $\eta^2 = .085$, $p < .05$. These results suggest that there is a significant influence of the status of the madrasah (state-private) on the NE scores and the year of examination on the NE scores.

The differences in NE achievement from year to year are explained as follows. The effect of the examination year variable can be considered based on the decrease or increase in NE achievement. In the mathematics test, the mean of 52.53 (2015) becomes 50.33 (2016), $F(1, 628) = 14.07$, partial $\eta^2 = .022$, $p < .05$. In the following year (2017), it fell to 46.93, $F(1, 628) = 5.30$, partial $\eta^2 = .002$, $p < .05$. The 2018 NE results rose slightly to 47.24, $F(1, 628) = 26.71$, partial $\eta^2 = .041$, $p < .05$.

In the science test, the decline of the mean was 45.24 (2015) to 43.37 (2016), $F(1, 628) = 41.78$, partial $\eta^2 = .062$, $p < .05$. In 2017, it fell to 42.85, $F(1, 628) = 70.58$, partial $\eta^2 = .101$, $p < .05$. The 2018 NE results rose to 42.91, $F(1, 628) = 15.42$, partial $\eta^2 = .024$, $p < .05$.

In the science test, the decline of the mean was 45.24 (2015) to 43.37 (2016), $F(1, 628) = 41.78$, partial $\eta^2 = .062$, $p < .05$. In 2017, it fell to 42.85, $F(1, 628) = 70.58$, partial $\eta^2 = .101$, $p < .05$. The 2018 NE results rose to 42.91, $F(1, 628) = 15.42$, partial $\eta^2 = .024$, $p < .05$.

The Profile of MTs Student Taking the NE

In this study, four profiles were explored (2 to 5 profiles; Table 2). Between the fitting models, the entropy value is the most conclusive criteria among the other criteria. The highest value is obtained in the three-profiles model. AIC, BIC, and SABIC show reasonable support to these three profiles. Although BLRT shows that the four-profiles model is more suitable than the three-profiles model, aLMR concludes, otherwise, the three-profiles model is better than the four-profiles model. Based on these compatibility criteria, the most reasonable three-profiles model is selected (Figure 2).

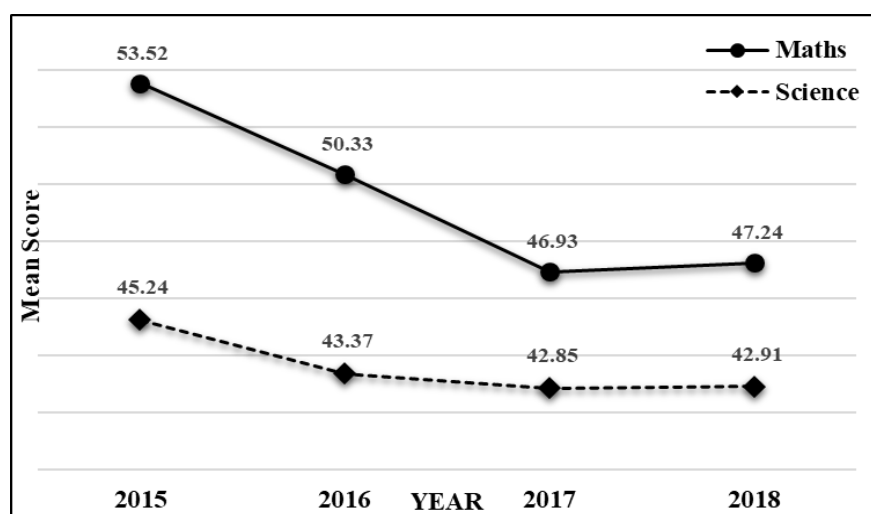


Figure 1. Trend of the Mean of NE Scores of MTs.

Table 2. Model Fit Indices for the Four Tested Models

	2 Profiles	3 Profiles	4 Profiles	5 Profiles
Log likelihood	-30550.824	-29768.638	-29303.236	-28854.721
Free parameters	37	50	63	76
AIC	61175.648	59637.275	58732.472	57861.442
BIC	61340.140	59859.561	59012.553	58199.316
SABIC	61222.669	59700.817	58812.535	57958.025
aLMR (<i>p</i>)	-	1545.924 (.19)	919.825 (.64)	919.721 (.53)
BLRT (<i>p</i>)	930.697 (.00)	930.803 (.00)	1564.373 (.00)	2182.034 (.00)
Entropy	.986	.993	.993	.975

Note: AIC = Akaike's information criteria; BIC = Bayesian information criteria; SABIC = sample-adjusted Bayesian information criteria; aLMR = adjusted Lo-Mendell-Rubin LR test; BLRT = bootstrap likelihood ratio test.

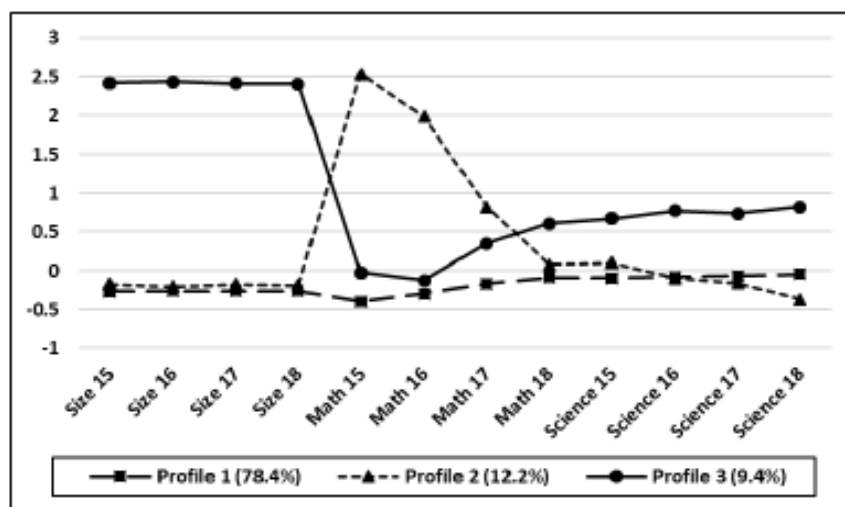


Figure 2. The Final Latent Profile Analysis Result in Z-score

Discussion

MTs Student Achievement (State-Private) in Mathematics and Science in the NE in 2015-2018

The results of the analysis show that in general, there is a trend of decline in MTs student mathematics and science scores in the NE. The decline across the year was proven to be multivariately significant. Analyzed from the status of the madrasahs, there were also differences in achievement between state and private madrasahs. In addition, the interaction between the status of the madrasahs (state-private) and the scores of the NE in the four years (2015, 2016, 2017, 2018) was also proven to be multivariately significant.

These results highlight that there is a significant influence on the status of the madrasahs (state-private) and the year of the examination toward student's average score of

mathematics and science in the NE. That is, the ability of MTs students to achieve mathematics and science is not yet maximal. Their lack of totality in learning mathematics and science may be because most of them have less ability to associate the benefits of learning the two subjects toward their religious knowledge even in their daily lives. Finally, they may have an idea that studying non-religious material such as mathematics and science can be ruled out. External factors can also contribute to and influence such a perspective. For example, most madrasahs still prioritize religious learning rather than learning other subjects. It brings an impact on the lack of maximum learning in mathematics and science (Akhwan, 2008). If the madrasahs override it all, madrasah students may not know and even master the material presented in both mathematics and science subjects.

The Profile of MTs Student Taking the NE based on the Status (State–Private), Number of Students, Year of Examination (2015, 2016, 2017, 2018), Scores in Mathematics and Science

The results of the analysis shows three patterns, namely: (1) Profile 1: small madrasahs, the scores in mathematics and science tend to rise, (2) Profile 2: medium madrasahs, mathematics score decreases, science score tends to decrease, and (3) Profile 3: big madrasahs, mathematics score decreases and then rises, science rises. Profile 2 deserves particular attention because it shows a consistent decrease in both mathematics and natural science. In profile 2, 75 of the 77 are private MTs. The NE mathematics score of the MTs students in this profile is much higher than those in other profiles. The possible explanation is as follows.

In the beginning, the NE became the basis for student graduation. At that time, MTs were competing so that all of their students could pass the NE, although sometimes they used less commendable ways (Kusaeri et al., 2018). For the sake of the good reputation of MTs, finally, any possible ways were taken to get the best NE results. Private MTs with a small and unstable number of students each year play a significant role in contributing to the drop in the NE scores in two subjects. The results of the study of Ali et al. (2011) show a strong correlation between the type of MTs and the number of students. The limited number of students and the lack of stability of the number of students enrolling to private MTs each year has implications to the limited facilities, infrastructure, and the number of teachers. This condition strongly affects the lack of maximum quality learning process (Nurhamzah, 2016).

Private MTs also lack the focus of learning. In addition to their main missions for religious education and character development, many private MTs, mainly those under the auspices of pesantren, have the curriculum highlighting aspects of skills and knowledge to find work after they graduate (Lukens-Bull, 2010). The teachers also contribute to low student achievement. MTs of-

ten prioritize primordial over professional aspects. As a result, quality teachers (especially for mathematics and natural science) are very rare.

The learning process in MTs also tends to encourage children to remember or memorize the contents of textbooks. This condition does not only occur in Indonesia, but also in madrasahs located in Pakistan and Bangladesh (Asadullah & Chaudhury, 2016). When such condition happens in the long term, according to Putra and Kumano (2018), teachers focusing students' memorization will not sufficiently develop students' readiness in science, technology, engineering, and mathematics (STEM) materials.

Alternative Solution and Policy to Improve Education Quality, Especially Mathematics and Science for MTs Students

There are several alternative solutions. (1) The government has the obligation to make policies that can change the mindset on the dichotomy of general knowledge (such as mathematics and science) and the Islamic subjects in MTs. Rao and Hossain (2011) also state that the views of the teachers or managers are dichotomous, and they distinguish Islamic subjects from between general science. (2) The improvement of MTs student mathematics and science learning outcomes can be achieved through the use of media, teaching materials, good quality (valid, practical, and effective) and innovative teaching and learning including: innovative learning models based on inquiry to improve learning outcomes, process performance and also high-level skills (Astutik & Prahani, 2018; Jatmiko et al., 2016; Pandiangan et al., 2017; Prayogi, Yuanita, & Wasis, 2018; Siswanto et al., 2018; Suyidno et al., 2018; Yunus, Sanjaya, & Jatmiko, 2013; Zulkarnaen, Supardi, & Jatmiko, 2017).

Conclusion

This study found a consistent decrease in the achievement in the mathematics and natural sciences in the NE from 2015 to 2018 for MTs students in Profile 2. The fundamental implications of the results of this

study are: (1) innovation in learning development is needed to improve learning outcomes in mathematics and science for the 21st century learning; (2) the data can serve as the main data base for making policy revisions to the curriculum, especially mathematics and science for 21st century learning; (3) the data provides empirical evidence for developed countries in order to improve the quality of mathematics and science learning outcomes for 21st century learning and support teaching and learning in the developing countries. Further research can be carried out, especially in the study of the development of innovative teaching and learning in mathematics and science subjects.

Acknowledgments

The authors would like to express sincere appreciation to the Center for Educational Assessment, the Ministry of Education and Culture of Indonesia for the valuable support to this research.

References

- Ahid, N. (2010). Problem Pengelolaan Madrasah Aliyah dan solusinya [The problems and solutions of Islamic High Schools management]. *Islamica: Jurnal Studi Keislaman*, 4(2), 336-353.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3), 317-332.
- Akhwan, M. (2008). Pengembangan Madrasah sebagai pendidikan untuk semua [Madrasa development as education for all]. *El-Tarbawi*, 1(1), 41-54.
- Al-Hasani, S. M. A., Ismail, A. R., Kazeemkayode, B., & Elegu, D. A. Q. (2017). Creating a practicing Muslim: A study of Qawmi Madrasah in Bangladesh. *British Journal of Education, Society & Behavioural Science*, 20(3), 1-9.
- Ali, M., & Furqon. (2016). Madrasah students' achievement study in Indonesia. *Global and Stochastic Analysis*, 3(3), 181-190.
- Ali, M., Kos, J., Lietz, P., Nugroho, D., Furqon, Zainul, A., & Emilia, E. (2011). *Quality of education in madrasah: Main study*. Jakarta: Ministry of Religious Affair.
- Asadullah, M. N., & Chaudhury, N. (2016). To madrasahs or not to madrasahs: The question and correlates of enrolment in Islamic schools in Bangladesh. *International Journal of Educational Development*, 49, 55-69.
- Astutik, S., & Prahani, B. K. (2018). The practicality and effectiveness of collaborative creativity learning (CCL) model by using PhET simulation to increase students' scientific creativity. *International Journal of Instruction*, 11(4), 409-424.
- Barkey, J. P. (2007). Madrasah medieval and modern: Politics, education, and the problem of Muslim identity. In R. W. Hefner & M. Q. Zaman (Eds.), *Schooling Islam: The Culture and Politics of Modern Muslim Education* (pp. 40-60). Princeton, NJ: Princeton University Press.
- Care, E., Griffin, P., & Wilson, M. (Eds.). (2018). *Assessment and teaching of 21st century skills: Research and applications*. New York, NY: Springer.
- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195-212.
- Flaherty, B. P., & Kiff, C. J. (2012). Latent class and latent profile models. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 3: Data analysis and research publication* (pp. 391-404). Washington, DC: American Psychological Association.
- Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis and latent profile analysis. *Psychometrika*, 24, 229-252.

- Ibda, H. (2018). Penguatan literasi baru pada guru Madrasah Ibtidaiyah dalam menjawab tantangan era revolusi industri 4.0 [Strengthening new literacy for Madrasah Ibtidaiyah teachers in responding to the challenges of the 4.0 industrial revolution era]. *JRTIE: Journal of Research and Thought of Islamic Education*, 1(1), 1-21.
- Isnawati, Indana, S., & Susantini, E. (2018). Using metacognitive strategy to teach learning strategies: A study of Indonesian pre-service biology teachers. *The New Educational Review*, 52(2), 258-268.
- Jatmiko, B., Prahani, B. K., Supardi, Z., Wicaksono, I., Erlina, N., Pandiangan, P., & Althaf, R. (2018). The comparison of OR-IPA teaching model and problem based learning model effectiveness to improve critical thinking skills of pre-service physics teachers. *Journal of Baltic Science Education*, 17(2), 1-22.
- Jatmiko, B., Widodo, W., Martini, Budiyanto, M., Wicaksono, I., & Pandiangan, P. (2016). Effectiveness of the INQF-based learning on a general physics for improving student's learning outcomes. *Journal of Baltic Science Education*, 15(4), 441-451.
- Kusaeri, K. (2016). Studi perilaku cheating siswa madrasah dan sekolah Islam ketika ujian nasional [Study of cheating behavior of madrasah students and Islamic schools on national exams]. *Edukasia: Jurnal Penelitian Pendidikan Islam*, 11(2), 331-354.
- Kusaeri, K. (2018). The portrait of Madrasah Aliyah in Indonesia: A critical evaluation of the mathematics score in the national examination. *Journal of Physics: Conference Series*, 1028, 1-7.
- Kusaeri, K., & Aditomo, A. (2019). Pedagogical beliefs about critical thinking among Indonesian mathematics pre-service teachers. *International Journal of Instruction*, 12(1), 573-590.
- Kusaeri, K., Aditomo, A., Ridho, A., & Fuad, A. Z. (2018). Socioeconomic status, parental involvement in learning and student' mathematics achievement in Indonesian senior high school. *Jurnal Cakrawala Pendidikan*, 37(3), 333-344. <https://doi.org/10.21831/cp.v38i3.21100>
- Kusaeri, K., Hamdani, A. S., Suparto, S., & Irmanila, E. (2018). Komparasi kredibilitas penyelenggaraan UNBK dan UNKP pada pelajaran matematika [Comparison of the credibility of the implementation of UNBK and UNKP in mathematics]. *Jurnal Ilmu Pendidikan*, 24(1), 10-18.
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767-778.
- Lukens-Bull, R. (2010). Madrasa by any other name: Pondok, pesantren, and Islamic schools in Indonesia and larger Southeast Asian region. *Journal of Indonesian Islam*, 4(1), 1-21.
- McClure, K. R. (2009). Madrasahs and Pakistan's education agenda: Western media misrepresentation and policy recommendations. *International Journal of Educational Development*, 29(4), 334-341.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models: Wiley series in probability and statistics*. New York, NY: Wiley.
- Meyer, J. P., Stanley, L. J., & Parfyonova, N. M. (2012). Employee commitment in context: The nature and implication of commitment profiles. *Journal of Vocational Behavior*, 80(1), 1-16.
- Ministry of Education and Culture. (2018). *Laporan hasil ujian nasional tahun pelajaran 2017/2018 [Report on the results of the 2017/2018 national examinations]*. Jakarta: Centre for Educational Assessment, Ministry of Education and Culture of Republic of Indonesia.

- Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7 ed.). Los Angeles, CA: Muthén & Muthén.
- Nizah, N. (2016). Dinamika Madrasah Diniyah: Suatu tinjauan historis [Dynamics of Madrasa Diniyah: A historical review]. *Edukasia: Jurnal Penelitian Pendidikan Islam*, 11(1), 181-202.
- Nurhamzah, N. (2016). The analysis of determinants factors in improving the quality of madrasah. *International Journal of Scientific & Technology Research*, 5(1), 1-4.
- Pandiangan, P., Sanjaya, I. G. M., & Jatmiko, B. (2017). The validity and effectiveness of physics independent learning model to improve physics problem solving and selfdirected learning skills of students in open and distance education systems. *Journal of Baltic Science Education*, 16(5), 651-665.
- Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences* (6 ed.). New York, NY: Routledge.
- Prayogi, S., Yuanita, L., & Wasis. (2018). Critical inquiry based learning: A model of learning to promote critical thinking among prospective teachers of physics. *Journal of Turkish Science Education*, 15(1), 43-56.
- Putra, P. D. A., & Kumano, Y. (2018). Energy learning progression and STEM conceptualization among pre-service science teachers in Japan and Indonesia. *The New Educational Review*, 53(3), 153-162.
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111-163.
- Ramaswamy, V., Desarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, 12(1), 103-124.
- Rao, N., & Hossain, M. I. (2011). Confronting poverty and educational inequalities: Madrasahs as a strategy for contesting dominant literacy in rural Bangladesh. *International Journal of Educational Development*, 31(6), 623-633.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Sclove, S. L. J. P. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333-343.
- Siswanto, J., Susantini, E., & Jatmiko, B. (2018). Multi-representation based on scientific investigation for enhancing students' representation skills. *Journal Physics: Conference Series*, 983, 012034.
- Subekti, H., Taufiq, M., Susilo, H., Ibrohim, & Suwono, H. (2018). Mengembangkan literasi informasi melalui belajar berbasis kehidupan terintegrasi STEM untuk menyiapkan calon guru sains dalam menghadapi era revolusi industri 4.0: Review literatur. *Education and Human Development Journal*, 3(1), 81-90.
- Suyidno, S., Nur, M., Yuanita, L., Prahani, B. K., & Jatmiko, B. (2018). Effectiveness of creative responsibility based teaching (CRBT) model on basic physics learning to increase student's scientific creativity and responsibility. *Journal of Baltic Science Education*, 17(1), 136-151.
- Wahab, A., Mahmud, A., & Tiro, M. A. (2018). The effectiveness of a learning module for statistical literacy. *The New Educational Review*, 53(3), 187-199.
- Woodward, K. E. (2015). Indonesian schools: Shaping the future of Islam and democracy in a democartics

- Muslim country. *Journal of International Education and Leadership*, 5(1), 1-14.
- Yunus, S., Sanjaya, I. G. M., & Jatmiko, B. (2013). Implementasi pembelajaran fisika berbasis guided inquiry untuk meningkatkan hasil belajar siswa auditorik [Implementation of physics learning based on guided inquiry to improve auditory learners outcomes]. *Jurnal Pendidikan IPA Indonesia*, 2(1), 48-52.
- Zulkarnaen, Supardi, Z. A. I., & Jatmiko, B. (2017). Feasibility of creative exploration, creative elaboration, creative modeling, practice scientific creativity, discussion, reflection (C3PDR) teaching model to improve students' scientific creativity of junior high school. *Journal of Baltic Science Education*, 16(6), 1020-1034.