REiD (RESEARCH AND EVALUATION IN EDUCATION)
Vol. 4, No. 1, June 2018

Indexed in:

Research and Evaluation in Education

Vol. 4, No. 1, June 2018

Vol. 4, No. 1, June 2018 **ISSN 2460-6995**

Research and Evaluation in Education

Publisher:
PROGRAM PASCASARJANA
UNIVERSITAS NEGERI YOGYAKARTA

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

Publisher
Program Pascasarjana Universitas Negeri Yogyakarta

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

Editorial
Department of Educational Research and Evaluation, Graduate School of Yogyakarta State University
3rd Floor Pascasarjana UNY New Building, Colombo Street No. 1, Karangmalang, Yogyakarta 55281
Telephone: 0274 586168 ext. 229 or 0274 550836, Facsimile: 0274 520326
E-mail: reid.ppsuny@uny.ac.id, reid.ppsuny@gmail.com

Foreword

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

REiD (Research and Evaluation in Education) contains and spreads out the results of research which is not limited to the area of common education, but also comprises the results of research in education in a broader coverage, such as moral education, social attitude in certain educational level, language education, mathematics education, teacher competence, and academic performance, with focuses on assessment and evaluation.

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

Yogyakarta, June 2018

Editor in Chief

TABLE OF CONTENT

REiD (Research and Evaluation in Education), 4(1), 2018, 1-11
Available online at: http://journal.uny.ac.id/index.php/reid

An evaluation of Islamic moral teaching for students of *Madrasah Aliyah Negeri* (MAN)

[*1]Siti Amanah; [2]Haryanto
[1]Graduate School of Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[2]Faculty of Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: sitia7001@gmail.com

Submitted: 14 March 2018 | Revised: 09 April 2018 | Accepted: 09 April 2018

Abstract

This research is aimed at evaluating: the preparation, implementation, and outcome of the moral teaching program using Stake Countenance evaluation model (antecedent, implementation, and outcome). The study was conducted at *Madrasah Aliyah Negeri* (MAN) Cilacap, MAN Kroya, and MAN Majenang. The subjects were the principals, teachers of *Aqidah Akhlak*, chairpersons of the *Madrasah* Committee, and 276 grade XII students. Interview, observation, questionnaire, and documentation were used as data collection techniques. The data analysis method used was the quantitative-qualitative descriptive analysis. The result of the evaluation shows that: (1) the preparation of the moral teaching is in 'good' category; (2) the implementation of moral teaching in terms of time and methods is in 'good' category, but the model of moral judgments used is not in 'good' category; (3) the result of moral teaching in the madrasah and outside the madrasah as a whole is in 'good' category. Thus, Islamic moral teaching evaluation of MAN in Cilacap Regency viewed from the preparation, implementation, and the result is in accordance with the evaluation criteria. In addition, there is a need for further action to examine the effectiveness of moral teaching in *madrasahs*.

Keywords: *evaluation, moral, stake countenance*

Introduction

According to Surah Al Baqarah [2]: 30, Allah Subḥānahu wa-taʿālā (SWT) created men to be the viceroy (leader) of the earth (Indonesian Department of Religious Affairs, 2011, p. 6). A good leader is a leader who provides examples through goodness and nobility of his/her soul and manner. Since the very beginning, education (both formal and non-formal ones) has played a fundamental role for human because it is the mean to bolster his/her level of knowledge and mannerism – or in Arabic, *akhlaq*. According to the National Education System Law, the goal of education is to enlighten a nation, develop the potentials, and build a healthy national civilization possessing manners and God-fearing attitude. Therefore, men of manners are symbols of success in education.

Etymologically, *akhlaq* comes from an Arabic word *al akhlaq* which is the plural form of *khuluqun* or ethical conduct, good attitude, good manner, or disposition (Ya'qub, 1991, p. 11). There are many verses of the Holy Quran that focus on *akhlaq*. One of them is Verse 4 in Chapter 68 which says:

وَإِنَّكَ لَعَلَى خُلُقٍ عَظِيْمٍ (القلم: ٤)

In English: 'And verily, you (Muhammad Sallallaahu Alaihi Wasallam) are on an exalted standard of character'. (Al-Qalam [68]: 4).

Al-Qalam [68]: 4 indicates that Allah SWT has chosen the best man, Prophet Muhammad *Sallallaahu Alaihi Wasallam* (SAW), as the example for mankind. There-

ISSN 2460-6995

fore, it is mandatory to follow his teachings as indicated in the following hadith.

إنما بعثت لأتمم مكارم الأخلاق

Prophet Muhammad SAW said: '*I have been sent to perfect good character*' (Abas, 1437, p. 7).

Perfect character is an achievement that is only possible through tireless work from both parents and other family members in educating their closest relatives (Daradjat, 2000, p. 35). Education starting from the family is essential in developing a child's character. Further, other parts of the community, such as close friends, colleagues, and other acquaintances, have to take part in the character (*akhlaq*) education (Zuchdi, 2013, p. 20).

Focusing on achieving the key functions and goals of education, schools or madrasas play an important role in producing human resource possessing elevated level of intelligence, manner and morality, providing excellent example to be followed. Darmayanti and Wibowo (2014, p. 227) state that schools are means to deliver strategic program and tackle the existing moral problems. At schools or madrasas, the process to achieve students' elevated level of mannerism might involve several curricular and extracurricular activities. The curricular activities are the activities in Islamic moral teaching and learning activities, while the extracurricular activities are religious discussion forum, Islamic holiday celebration, and also pilgrimage to Mecca. The importance of those activities lays in the fact that the learning process and religious atmosphere affect the students' learning behavior and the fact that the result of the study is heavily affected by the religious atmosphere at the schools (Kartowagiran & Maddini, 2015, p. 995).

Furthermore, the teachers' tenacity and leadership are also the key factors to the success of Islamic moral teaching (*akhlaq* teaching). As professional educators, teachers perform their main tasks including teaching, educating, fostering, guiding, directing, counseling, training, assessing and evaluating students in the levels of early-childhood, elementary, and secondary education (Kartowagiran, 2011, p. 464). Therefore, the professional teachers will be able to facilitate the students to construct good manner and habit.

As an Islamic education institution, *Madrasah Aliyah Negeri* (State Islamic Senior High School/MAN) in Cilacap Regency, Central Java, Indonesia has put forward religious education by focusing on producing students with good manners (good *akhlaq*) and high achievement so that they are excellent at science and technology. As a follow up, madrasas have taken many actions. However, in my opinion, many of the factors have not been properly assessed. These factors are (1) available resources, (2) religious curricular and extracurricular activities, (3) applicable learning methods, (4) an applicable evaluation model, (5) the goal and scope of the mannerism being built. Additionally, even-though the strategic plan had been arranged, the achievement of the activities enforcement supporting the programs had not been well arranged. Based on the observation, it was known that (1) the teachers were not consistent in preparing the lesson plan device for Islamic moral teaching (*akhlaq*/character building), (2) the teachers did not rely on the standards of character (*akhlaq*) evaluation in evaluating the students, and (3) the process of evaluation conducted by the teacher was not systematic. Therefore, the evaluation on Islamic moral teaching at MAN was highly needed to find out the imperfect parts of the implementation of Islamic moral teaching in Cilacap Regency.

In this context, evaluation is a set of practices to determine the quality, performance, and productivity of an institution in the implementation of its programs (Mardapi, 2012, p. 4). It is also noted that *'Evaluation is the determination of the worth of the thing. It includes obtaining information for use in judging the worth of a program, product, procedure, or objective, or the potential utility of alternatives approaches designed to attain specified objectives'* (Worthen & Sanders, 1973, p. 19). The evaluation aims to answer the following questions: (1) How are the Islamic moral teaching programs prepared; (2) How are the Islamic moral teaching programs executed; and (3) What is the result of Islamic moral teaching at MAN in Cilacap Regency like?

Method

The research was conducted from February to August 2016. The evaluation of the Islamic moral teaching was conducted in three MAN's in Cilacap Regency: MAN Cilacap, MAN Kroya, and MAN Majenang. The research setting was determined based on the similarities among these schools in terms of the implementation of the Islamic moral teaching programs.

The subjects of the evaluation were grade XII students, Madrasas' principals, teachers of *Aqidah Akhlak* (Islamic Creed and Mannerism) and the heads of madrasa committee. The research sample was established using purposive sampling technique by considering the competence and role. Table 1 shows the number of the involved respondents. Random sampling technique with Slovin's formula was applied to proportionally select 276 respondents out of 806 students (see Table 2.)

Data Collection Techniques

This was an evaluation research implementing quantitative and qualitative approach. This research employed *Stake Countenance* evaluation model consisting of three stages of evaluation: preparatory stage (*antecedent*), implementation stage (*transaction*), and result stage (*outcome*).

In this research, there were two types of data: quantitative and qualitative data. The quantitative data were collected using a questionnaire, whereas the qualitative data were gathered through interviews with the madrasa principals, teachers, and the head of the madrasa committees. The data collected through observation and documentation were used to support the result of the analysis on quantitative and qualitative data. Table 3 shows the data collection techniques used.

Table 1. The number of respondents consisting of the madrasa principals, teachers and the head of the madrasa committees

MAN	The Madrasa principals	Teachers	The Head of Committee
MAN Cilacap	1	1	1
MAN Kroya	1	1	1
MAN Majenang	1	1	1
Total	**3**	**3**	**3**

Table 2. The number of respondents from students

MAN	Number of Students	%	Number of Respondents
MAN Cilacap	238	30	70
MAN Kroya	231	28	65
MAN Majenang	337	42	141
Total	**806**	**100**	**276**

Table 3. Data collection techniques

Aspects	Indicators	Techniques
Preparatory (*antecedent*)	- Resources - Goals and scopes of Islamic moral teaching (*akhlaq teaching*) - The management of the infrastructures	Interviews Documentation Observation
Implementation (*transaction*)	- Implementation time - Islamic moral teaching methods - Character evaluation models	Interviews Documentation Observation Questionnaire
Results (*outcome*)	- The application of Islamic morality by the students in the area of the madrasas - The application of Islamic morality by the students outside the area of the madrasas	Interviews Documentation Observation Questionnaire

The goals of the preparatory stage (*antecedent*) were to gain insights into resources, purposes, scope of the material, and management of the infrastructures. The implementation stage (*transaction*) included evaluation on the implementation time, methods used, and evaluation models used. The result stage (*outcome*) aimed at understanding the application of Islamic morality by the students in the madrasas area. The data concerning these stages were collected through interviews, observations, document analysis, and questionnaire.

In order to better understand the above stages, the researchers conducted interviews with the madrasa principals, the head of the madrasa committees, and the teacher of *Aqidah Akhlak* for the twelfth grade students. The preparatory aspect was essential in revealing resources, goals, scopes of the materials, and the management of the infrastructure. Moreover, the indicators for resources were the strategic plans and competence of the teachers in delivering Islamic moral teaching (*akhlaq* teaching). In the aspect of implementation, the interviews aimed at revealing the implementation time of the Islamic moral teaching and the methods and evaluation of the education applied. The last stage (evaluation stage) was designed to reveal the application of Islamic morality by the students inside and outside the area of the madrasas.

There were three evaluators in three different MAN's to understand the infrastructures, process of teaching and learning of the subject of *Aqidah Akhlaq* in madrasas, time of the Islamic moral teaching, and the methods used to build good characters in the students. The evaluators conducted documentation to gather supporting data for the preparation, implementation, and result of the Islamic moral teaching at MAN's in Cilacap Regency. The documentation specifically gathered the data concerning resources, materials, infrastructures of the madrasas, and curricular and extracurricular activities related to Islamic moral teaching in the madrasas.

As a part of the evaluation process, the authors distributed questionnaire to 276 students. The goal was to gather the students' responses to the components of the implementation of the Islamic moral teaching, the methods of the Islamic moral teaching, the models of the evaluation, and the application of Islamic morality by the students inside and outside the area of the madrasas.

Validity of the Instruments

The validity of the content was a means to understand the accuracy of the instruments of the observation, interviews, and questionnaire conducted by three experts in the field of education research and evaluation. The validity of the content was measured using Aiken V formula. Table 4 shows the content validity of the observation sheet.

Table 5 shows the analysis result of the content validity of the interview instruments. Meanwhile, Table 6 shows the analysis result of the content validity of the questionnaire.

Table 4. The result of V value in the content validity of the observation sheet

V Value	V Value in Table	Items No.	Items Numbers	Description
1	0.81	2, 7, 9	3	Valid
0.89	0.81	1, 3, 4, 5, 7, 8	6	Valid
	Total		**9**	**Valid**

Table 5. The result of V value in the content validity of the interviews instruments

V Values	V Value in the Table	Items No.	Items Numbers	Desc.
1	0.68	5, 7, 12, 16, 19, 27, 34	7	Valid
0.89	0.68	1, 2, 3, 4, 6, 9, 10, 11, 13,14, 15, 17,18, 20, 22,23, 26, 29, 30,31,32,33,36,37,38,40,42, 43, 44, 46, 47, 48, 50,51,53, 54,55,56,57	39	Valid
0.78	0.68	8, 21, 24,25, 28, 35, 39, 41, 45, 49, 52	11	Valid
		Total	**57**	**Valid**

Construct Validity was used to reveal the accuracy of the construction of the questionnaire, which was measured with exploratory factor analysis (Retnawati, 2016, p. 42). The result of the construct validity analysis is shown in Table 7.

Table 7. KMO and Barlett's test result

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.614
Bartlett's Test of Sphericity	Approx. Chi-Square	1475.660
	Df	780
	Sig.	0.000

It can be concluded that the instruments of the Islamic moral teaching evaluation were valid for data collection. Out of 50 items in the questionnaire, there were 10 items with *Anti-image Correlation* value < 0.5. The items were items 10, 12, 16, 20, 30, 43, 44, 45, 46 and 49. Thus, there were 40 valid items in the questionnaire.

Reliability of the Instruments

The reliability index of the instruments was considered acceptable if the reliability value was > 0.7 (Linn, 1989, p. 106). The reliability of the observation sheets was estimated using the ICC (*Intraclass Correlation Coefficient*) formula. Generally, the result of observation sheet ICC analysis based on the rater was at 0.895 and the result of the evaluation on each rater was at 0.740. Based on the estimation, the instruments were deemed to be reliable and valid for conducting the research. Table 8 shows the reliability of the observation sheets.

The estimation of the questionnaire reliability was conducted with *Alpha Cronbach* coefficient formula supported with the SPSS program. Based on the estimation result, the coefficient value of the *Cronbach's Alpha* was at 0.867 or higher than 0.7. Therefore, the questionnaire was deemed reliable (see Table 9).

Table 9. Reliability of the *Alpha Cronbach*

Reliability Statistics	
Cronbach's Alpha	N of s
0.867	40

Data Analysis Techniques

This research employed quantitative-qualitative descriptive analysis. Each technique is described in the following sections.

Quantitative Data Analysis

Quantitative data analysis was used to describe the data collected through the questionnaire based on the score. The scores were categorized using normal distribution. The categorization was conducted with normal curve as the reference with the measurement of mean ideal (Mi) and standard deviation (SDi) (Mardapi, 2008, p. 123). The score categorization of the students is shown in Table 10.

Table 6. The result of V values in the content validity of the questionnaire

V Value	V Value in Table	Item no.	Item Numbers	Desc.
1	0.68	1, 3, 5, 7, 8, 10, 16, 17, 19, 21, 22, 24, 25, 28, 29, 30, 31, 34, 35, 37, 40, 42, 43, 47, 50	25	Valid
0.89	0.68	2, 4, 6, 9, 11, 12, 13, 14, 18, 23, 26, 27, 32, 33, 36, 38, 39, 41, 45, 46, 48, 49	22	Valid
0.78	0.68	15, 20, 44	3	Valid
		Total	**50**	**Valid**

Table 8. The result of observation sheet ICC

Intraclass Correlation Coefficient							
	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.740a	.406	.927	9.538	8	16	.000
Average Measures	.895c	.672	.974	9.538	8	16	.000

Table 10. The categorization of the Islamic moral teaching result

No.	Score	Category
1.	$\bar{X} + 1{,}5\ SBx \leq X$	Very Good
2.	$\bar{X} \leq X < \bar{X} + 1{,}5\ SBx$	Good
3.	$\bar{X} - 1{,}5\ SBx \leq X < \bar{X}$	Poor
4.	$X < \bar{X} - 1{,}5\ SBx$	Very Poor

Description:
$\bar{X}$: average of total score
SBx : standard deviation of total score
X : score achieved

Qualitative Data Analysis

The qualitative data analysis was used by implementing the interactive and sustainable analysis model of Miles and Huberman which consisted of four stages: data collection, data description, data reduction, and data verification/conclusion. Figure 1 schematically shows the analysis process of the qualitative data.

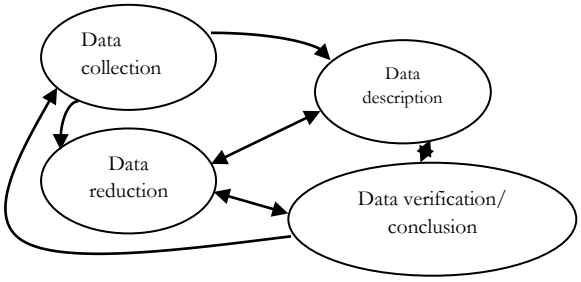

Figure 1. Analysis model of qualitative data (Miles & Huberman, 1994)

Findings and Discussion

The Preparation for Islamic Moral Teaching (*Akhlaq* Teaching)

The Resource of Education

From the interviews, it is known that the resources of the Islamic moral teaching at MAN's in Cilacap Regency were supported by the vision and mission of the Islamic moral teaching as defined in the strategic plan. Rusniati and Haq (2014, p. 102) state that a strategic plan is a plan to utilize available resources in order to achieve the goals set by an organization. MAN's in Cilacap Regency position the strategic plan as the implementation guidance for gaining academic achievement and as the vision and mission of the madrasas for the years to come.

In addition to the strategic plan, from the interviews, it is also known that the teachers of the madrasas had been trained to facilitate the students to succeed in the cognitive, psychomotor, and also affective areas. Moreover, Kartowagiran (2011, p. 465) states that teachers are the spearhead of the effort to level up the quality of the service and result of education. Other important factors, in addition to strategic plan and the competence of the teachers, were material and non-material supports from the parents/guardian in the Islamic moral teaching. This is supported by Jalaluddin (2011, p. 291) who argues that, in building Islamic attitude within the children, parents have to provide supports in the elementary education.

The Goals and Scope of the Materials

Islamic moral teaching aims at providing the students with motivation not only to study *aqidah* but also to put it into everyday practices. In Islamic studies, morality includes our morality in relation to God and His Prophet, other people, ourselves, and nature. This is in line with previous studies on the similar subjects by Prihatini, Mardapi and Sutrisno (2013, p. 347) who discovered that the construction of Islamic morality encompassed our morality in our relation to Allah *subhannahu wa ta'ala*, Prophet Muhammad *salallahu alaihi wasallam*, parents, ourselves, friends, family, community, and nature.

Specifically, in the Islamic moral teaching, the teaching materials cover the materials on good morality, poor morality, stories and examples, and *Aqidah* (human relationship with Allah SWT). In madrasas, these materials are transformed into materials of Islamic moral teaching.

Facilities and Infrastructure

The evaluation of the facilities and infrastructure of the madrasas focused on the facilities and infrastructures used in curricular and extracurricular activities. In the curricular activities, there were (1) lesson plan (RPP), (2) the implementation of the lesson plan in the classroom; (3) learning activity manual and

referential books; and (4) evaluation on the Islamic moral teaching. Whereas, in the extra-curricular activities, there were: (1) praying room; (2) praying equipment; (3) facilities and infrastructure for ablution before prayers; and (4) toilets. The condition of each facility and infrastructure is shown in Figure 2.

As shown in Figure 2, overall, the facilities and infrastructures at MAN's in Cilacap are in a good condition. Here are the percentages at each MAN: MAN Cilacap is 88% (very good), MAN Kroya is 63% (good) and MAN Majenang is 75% (good).

The Transaction of Islamic Moral Teaching

The result of the implementation of the Islamic moral teaching in those madrasas is categorized based on the score of each madrasa (see Table 11).

Table 11. The categorization of the result of the transaction of the Islamic moral teaching

No.	Score	Categories
1.	$3.25 \leq X$	Very Good
2.	$2.50 \leq X < 3.25$	Good
3.	$1.75 \leq X < 2.50$	Poor
4.	$X < 1.75$	Very Poor

Generally, the implementation of the Islamic moral teaching at MAN's in Cilacap Regency is in 'good' category. It can be seen from the score of 2.79 with the percentage as high as 70%. The scores for each MAN in Cilacap Regency are presented in Figure 3. As shown by the result, the Islamic moral teaching has been well implemented, following the schedule set and applying an appropriate learning method and technique.

Implementation Time of Islamic Moral teaching

In terms of the implementation time, the Islamic moral teaching is in a good category scoring 3.08 with percentage of 77%. The details of the implementation at each MAN are shown in Table 12.

Table 12. Scores on the indicator of the implementation time of the Islamic moral teaching

Indicator	Madrasa	Score	%	Categories
Implementation Time	MAN Cilacap	3.07	77	Good
	MAN Kroya	3.58	90	Very Good
	MAN Majenang	2.86	72	Good
Total		**3.08**	**77**	**Good**

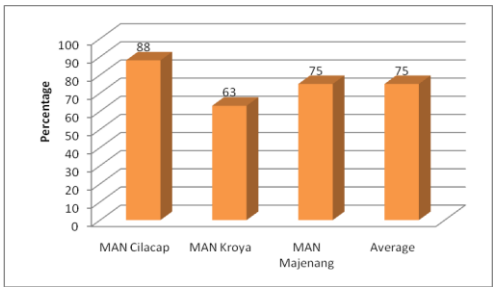

Figure 2. Percentage of facilities and infrastructures for Islamic moral teaching at MAN's in Cilacap Regency

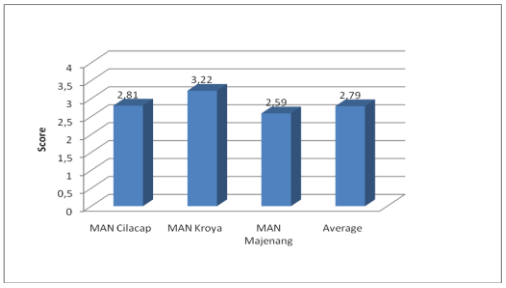

Figure 3. The scores for each MAN in the implementation stage of the Islamic moral teaching

In this research, the Islamic moral teaching covers curricular activities (learning *Aqidah Akhlaq* in the classroom) and also extracurricular activities. The extracurricular activities in the madrasas are HIMDA'IS (Student's Preachers Association) at MAN Cilacap, *An Nisa* Study Club at MAN Majenang, and Rohis (Islamic Spiritual and Religious Students Group) at MAN Kroya. Other religious extracurricular activities include: praying together, reading the 99 holy names of Allah, community service, reading the Holy Quran, organizing Islamic Holidays, distributing charity, hajj training, short course on Islamic studies during Ramadhan, *Musabaqah Tilawatil Quran* (Quran recitation festival), *Hadroh Sholawatan* (Gathering to praise the Prophet Muhammad SAW), socialization, and also *istighozah* (gathering to ask Allah for help).

Methods of the Islamic Moral teaching

The indicator of the method applied in the Islamic moral teaching at MAN's in Cilacap Regency is in a good category. The details are presented in Table 13.

Table 13. Score on the indicator of the methods applied in the Islamic moral teaching

Indicator	Madrasa	Score	%	Category
Evaluation methods	MAN Cilacap	2.88	72	Good
	MAN Kroya	3.17	79	Good
	MAN Majenang	2.66	67	Good
Average		**2.84**	**71**	**Good**

In the Islamic Moral teaching, preaching, providing example, habit formation, giving advice, motivating, and admonishing are the methods used by the teachers. Daradjat (1984, p. 262) and Thoha, et al. (2004, p. 122) state that there are some effective methods in delivering the Islamic moral teaching, including preaching, conducting question-and-answer session, opening discussions, habit formation, providing exemplary actions, and providing advice related to the teaching materials.

Evaluation Model of Islamic Moral Teaching

The evaluation model of the Islamic moral teaching at these three MAN's in Cilacap Regency is in a poor category. The reason is that most of the evaluation was conducted only on cognitive aspects. In general, the evaluation put aside the affective aspects. This is in line with previous research which was conducted by Syamsudin, Budiyono and Sutrisno (2016, p. 40) who argue that: '*The elicitation of data from the objects of research has not been as easy as it has been thought because of the behavioral dynamics of the human individuals involved as research subjects*..'. The details of the scores of the evaluation model of the Islamic moral teaching in each school are elaborated in Table 14.

Table 14. Scores on the indicators of the evaluation model of the Islamic moral teaching

Indicator	Madrasa	Score	%	Category
Evaluation Model	MAN Cilacap	1.99	50	Poor
	MAN Kroya	2.42	61	Poor
	MAN Majenang	1.77	44	Poor
Average		**1.98**	**50**	**Poor**

In the next phase, it is known that the difficulty to measure the affective achievement laid on the fact that the indicator of affective aspects is hard to measure directly. It is possible, but it requires more time spent for observation. In addition, Khuriyah (2003, p. 60) in her research states that the construct in the measurement of morality has not been fully developed.

The Outcome of Islamic Moral Teaching

The outcome of Islamic moral teaching at MAN's in Cilacap Regency is categorized based on the score that each school gained. The details of the categorization are presented in Table 15.

Table 15. The categorization of the outcome of the Islamic moral teaching at MAN's in Cilacap Regency

No.	Score	Category
1.	$3.25 \leq X$	Very Good
2.	$2.50 \leq X < 3.25$	Good
3.	$1.75 \leq X < 2.50$	Poor
4.	$X < 1.75$	Very Poor

In general, based on the data which were collected through the questionnaire, the outcome of the Islamic moral teaching at MAN's in Cilacap Regency reaches the score of 2.99, or it is in a good category. This shows that the students have partially applied good morality in their relation with Allah SWT, the Prophet, other people (social interaction), themselves, and also the nature. The score which is gained by each MAN is shown in Table 16.

Table 16. Scores on the outcome stage of the Islamic moral teaching

Indicator	Madrasa	Score	%	Category
Evaluation Outcome	MAN Cilacap	3.04	76	Good
	MAN Kroya	3.15	79	Good
	MAN Majenang	2.89	72	Good
Average		**2.99**	**75**	**Good**

There are two indicators in the outcome of the Islamic moral teaching at MAN's in Cilacap Regency, namely: (1) the application of the Islamic morality inside the area of the madrasas, and (2) the application of the Islamic morality outside the area of the madrasas.

The Implementation of Islamic Morality by Students inside the Area of the Madrasas

In applying Islamic morality in the area of the madrasas, the students of MAN Cilacap gained the score of 3.07 (77%). The students of MAN Kroya scored 3.18 (80%), and the students of MAN Majenang scored 2.90 (73%). In agreement with these scores, during the interviews, the madrasa principals and the teachers stated that, in general, the students applied Islamic morality around the schools well. It is portrayed, evidently, in the behavior and habits of the students. They are friendly, disciplined, soft-spoken, top achievers, and also active in worship and other religious activities. Table 17 and Figure 4 show the score on the implementation of Islamic morality in each *Madrasah Aliyah Negeri* (MAN) in Cilacap Regency.

Table 17. The score and percentage of the students' application of the Islamic morality in the area of the madrasas

Indicator	Madrasa	Score	%	Category
Application of Islamic morality inside the madrasas	MAN Cilacap	3.07	77	Good
	MAN Kroya	3.18	80	Good
	MAN Majenang	2.9	73	Good
Average		**3.01**	**75**	**Good**

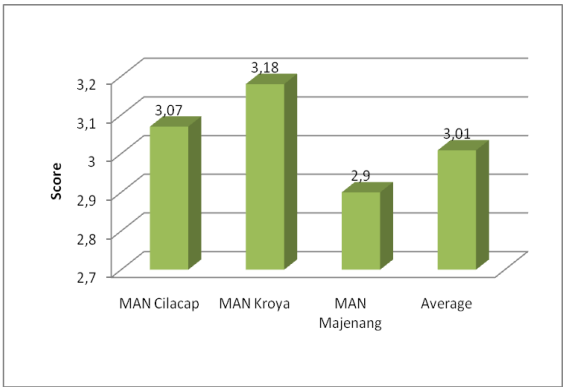

Figure 4. The score of the students' application of Islamic morality in the area of the madrasas

The good category means that Islamic Morality Education in the madrasas had driven the students to be better human beings. This is in line with the definition of education by Marzuki (2009, p. 1) which states that the process of education is a part of the agents of change which shall possess the power to improve the characters of the nation through the improvement of the characters of the students in the education institutions.

The Implementation of Islamic Morality by Students outside the Area of the Madrasas

There were two indicators in defining the application of Islamic morality – based on the orders of Allah SWT and the teachings of Prophet Muhammad SAW - outside the area of the madrasas: (1) the students' ability to control themselves and shield themselves from promiscuity and negative impacts of technology development, and (2) the students' ability to foster awareness on social issues. The indicator is in a good category with the score of 2.88 and the percentage of 72%. The scores are shown in Table 18 and Figure 5.

Table 18. The score and percentage of the students' application of the Islamic morality outside the area of the madrasas

Indicator	Madrasa	Score	%	Category
Application of Islamic morality outside the madrasas	MAN Cilacap	2.88	72	Good
	MAN Kroya	2.99	75	Good
	MAN Majenang	2.83	71	Good
Average		**2.88**	**72**	**Good**

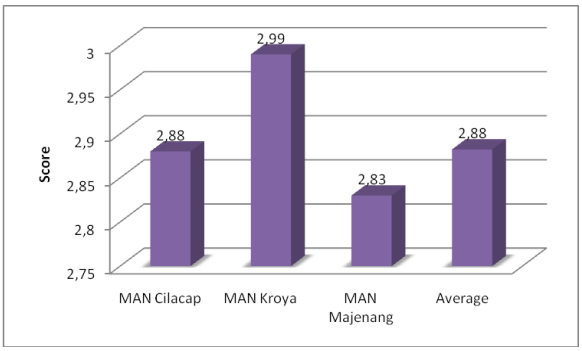

Figure 5. The score of the students' application of the Islamic morality outside the area of the madrasas

The good category means that the students had applied what they had learnt from the Islamic moral teaching outside the area of the madrasas. It was evidenced in their worship activities, their ability to stay away from bad habits and promiscuity, and the negative impacts of technology. They were also able to apply the Islamic morality in the social aspects. The finding in this research is in line with Zuchdi (2013, p. 20) who affirms that the community, represented by colleagues, close friends, co-workers and other parties in the community, has to take part in the development and education of morality of the students as the heirs of the nation.

Conclusions and Recommendations

Conclusions

Based on the result of the Islamic moral teaching at MAN's in Cilacap Regency, it can be concluded that: (1) the preparatory stage (antecedent) of the Islamic moral teaching, covering resources, goals, material scope, and facilities and infrastructures, scored 75% or is in a good category; (2) the implementation stage (transaction) of the Islamic moral teaching, covering the implementation time and the method, is in a good category (except for the evaluation model which is in a good category); (3) in general, the outcome of the Islamic moral teaching – the application of Islamic morality inside and outside the area of the madrasas - at MAN's in Cilacap Regency scored 75% or is in a good category.

Recommendations

The result of the analysis, i.e the indicator of the evaluation model of the programs of Islamic moral teaching is in a poor category. Therefore, the evaluation aspects of the education need improvement and deeper analysis. The goal is to formulate students' morality criteria objectively. Additionally, the research on the effectiveness of the Islamic moral teaching has to be conducted. It will serve as the follow up of this evaluation, which should serve as the basis for researchers to conduct systematic research on the aspects of Islamic moral teaching, specifically on the morality evaluation techniques and models, the development of the morality evaluation instruments, the effectiveness of the morality teaching methods and the effectiveness of the programs of the Islamic moral teaching.

References

Abas, Z. Z. (1437). *Makarimal Akhlak*. Cairo: Al Qomar.

Daradjat, Z. (1984). *Dasar-dasar pendidikan agama Islam: Buku teks pendidikan agama Islam pada perguruan tinggi umum*. Jakarta: Bulan Bintang.

Daradjat, Z. (2000). *Ilmu pendidikan Islam*. Jakarta: Bumi Aksara.

Darmayanti, S. E., & Wibowo, U. B. (2014). Evaluasi program pendidikan karakter di sekolah dasar Kabupaten Kulonprogo. *Jurnal Prima Edukasia*, *2*(2), 223–234. https://doi.org/10.21831/jpe.v2i2.2721

Indonesian Department of Religious Affairs. (2011). *Al Qur'an dan terjemahnya: Syaamil al Qur'an special for women*. Bandung: PT. Sigma Exa Grafika.

Jalaluddin. (2011). *Psikologi agama*. Jakarta: Raja Grafindo Persada.

Kartowagiran, B. (2011). Kinerja guru profesional (guru pasca sertifikasi). *Cakrawala Pendidikan*, *30*(3), 463–473. https://doi.org/10.21831/cp.v3i3.4208

Kartowagiran, B., & Maddini, H. (2015). Evaluation model for Islamic education learning in junior high school and its significance to students' behaviours. *American Journal of Educational Research*, *3*(8), 990–995. https://doi.org/10.12691/education-3-8-7

Khuriyah, K. (2003). Pengembangan instrumen evaluasi ranah afektif untuk pedidikan agama Islam. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *5*(6), 59–73. https://doi.org/10.21831/pep.v5i6.2058

Linn, R. L. (1989). *Educational measurement*. New York, NY: Macmillan.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekia.

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

Marzuki. (2009). *Prinsip dasar akhlak mulia: Pengantar studi konsep-konsep dasar etika dalam Islam*. Yogyakarta: Debut Wahana Press.

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

Prihatini, S., Mardapi, D., & Sutrisno, S. (2013). Pengembangan model penilaian akhlak peserta didik madrasah aliyah. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(2), 347–368. https://doi.org/10.21831/pep.v17i2.1705

Retnawati, H. (2016). *Validitas reliabilitas & karakteristik butir (Panduan untuk peneliti, mahasiswa, dan psikometrian)*. Yogyakarta: Nuha Medika.

Rusniati, & Haq, A. (2014). Perencanaan strategis dalam perspektif organisasi. *Jurnal INTEKNA: Informasi Teknik Dan Niaga*, *14*(2), 102–209. Retrieved from http://ejurnal.poliban.ac.id/index.php/intekna/article/view/178

Syamsudin, A., Budiyono, B., & Sutrisno, S. (2016). Model of affective assessment of primary school students. *REiD (Research and Evaluation in Education)*, *2*(1), 25–41. https://doi.org/10.21831/reid.v2i1.8307

Thoha, C. (2004). *Metodologi pengajaran agama* (2nd ed.). Yogyakarta: Pustaka Pelajar.

Worthen, B. R., & Sanders, J. R. (1973). *Educational evaluation: Theory and practice*. Worthington, OH: Longman.

Ya'qub, H. (1991). *Etika Islam: Pembinaan akhlaqulkarimah (suatu pengantar)*. Bandung: Diponegoro.

Zuchdi, D. (2013). *Pendidikan karakter: Konsep dasar dan implementasi di perguruan tinggi*. Yogyakarta: UNY Press.

REiD (Research and Evaluation in Education), 4(1), 2018, 12-21
Available online at: http://journal.uny.ac.id/index.php/reid

Assessment of the social attitude of primary school students

[*1]Ari Setiawan; [2]Siti Partini Suardiman
[1]Universitas Sarjanawiyata Tamansiswa
Jl. Kusumanegara 157 Yogyakarta 55165, Indonesia
[2]Graduate School of Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: ari.setiawan@ustjogja.ac.id

Submitted: 11 April 2018 | Revised: 28 May 2018 | Accepted: 17 July 2018

Abstract

The implementation of Curriculum 2013 at primary school level brings about its own problems to teachers. A serious problem emerges in the assessment, especially the assessment of core competence for the social attitude aspect. This problem arises because social attitude has many dimensions and requires judgments in diverse forms. In addition, the assessment of social attitude is focused on the affective sphere. The objective of this research is to assess the social attitude of grade IV and/or V students of primary school using three integrated instrument models: self-assessment (SA), peer assessment (PA), and observational assessment (OA). This research employed qualitative approach. The respondents were 58 students chosen by using cluster random sampling and purposive sampling techniques. The data were collected through direct disclosure questionnaire and observation, and analyzed descriptive quantitatively. The results of the research are as follows: (1) the component of honesty attitude is in category A (entrusted); (2) the component of discipline is in category A (entrusted); (3) the responsibility component is in category B (developing); (4) the politeness component is in category B (developing); (5) caring component is in category B (developing); (6) confidence component is category A (entrusted); and (7) students' social attitude is mainly in category B (good) which indicates that most students have good social attitude.

Keywords: *assessment, social attitude, primary school*

Introduction

There are three domains of learning outcomes that a student achieves in a learning process, namely: cognitive, affective, and psychomotor domains (Krathwohl, Bloom, & Masia, 1973, pp. 6–7). Cognitive domain is the result of learning that has something to do with memory, ability to think, or intelligence. In addition, affective domain refers to learning outcomes in the form of sensitivity and emotion that deals with attitude, values, and interests, meanwhile, psychomotor domain is related to a certain skill or ability of motion (Kurniawan, 2014, pp. 10–12). As a result of learning, these three domains require assessment, including integrated thematic approach model. A successful learning is defined by behavior (affective) as well as environment (Retnawati, 2016).

One aspect that requires assessment is affective domain. The characteristics of the affective domain are attitude, values and interests (McCoach, Gable, & Madura, 2013, pp. 7–24). The attitude referred to in this study is the social attitude of elementary school students. Social attitude is an affective domain that needs to be assessed using an appropriate instrument.

Social attitude can be seen as something associated to the attitude which is related to social conditions. It is an acquired tendency to evaluate social things in a specific way. It is characterized by positive or negative beliefs in, feelings of, and behaviors on a particular entity. It has three main components: emo-

ISSN 2460-6995

tional, cognitive, and behavioral components. The emotional component is the feeling experienced in evaluating a particular entity. The cognitive component implies thoughts and beliefs adopted towards the subject, while the behavioral component is the action that results from a social attitude (Bernann, 2015, p. 13).

LaPierre in Azwar (2015, p. 5) contends his idea that social situation is an anticipatory pattern of behavior, tendency or readiness, predisposition to conformity in social situations, or simply social attitude is a response to conditioned social stimuli. In other words, social attitude is a pattern of behavior regarding conditioned social situations.

Ahmadi (2002, p. 163) writes that social attitude is the consciousness of an individual who determines the real, repetitive actions of the social object. Thus, social attitude represents a person's response to social objects. In line with this idea, Gerungan (2004, p. 161) proposes that social attitude is the same and repeated ways of responding to social objects. It leads to the repeated ways of behaving toward a social object. As stated by Soekanto (Supardan, 2011), social objects relate to interpersonal behavior or social processes. It involves relationships between people or groups in social situations.

Social attitude is a tendency to evaluate social things in a certain way. It plays an important role in children's development, because it shapes children's perceptions of the social environment and has a significant effect on behavior (Crano & Prislin, 2011, p. 19). Children who start interacting with the social environment will begin to have social attitude, and this also occurs in primary school-aged children.

Considering the various understandings above, the writer concludes that social attitude is the awareness of a person in acting repetitively in real life to determine the response to social objects in his or her relation with others. Social attitude encourages a person to do things in a certain way as a form of his or her reaction to social objects.

The evidence of children's behaviors these days is quite concerning. Primary school students are now generally less disciplined than they used to, and they have low care and responsibility. It is not in accordance with the ideal affective development of primary students. Ekowarni (2009) contends that there are some values related to social condition that should be instilled in primary school students, including: politeness, caring, cooperativeness, discipline, humility, even-temperedness, tolerance, independence, honesty, confidence, toughness, positivity, fairness, peacefulness, perseverance, creativity, citizenship, responsibility, and sincerity.

In today's education practice, where social attitude actually becomes the core of education, the assessment has not yet been conducted. This is due to the teachers' limitations, especially in the assessment process. Teachers are more likely to spend their time on teaching regardless of the importance of making appropriate assessment. Stiggins's study shows that teachers should spend a third to a half of their available time to engage in assessment activities (Stiggins, 1999, p. 3). They are constantly making decisions about how to interact with their students, and decisions that are based on part of information that they collect about their students through classroom appraisals. In fact, they do not spend much time on assessment.

The results of a study conducted by Zuchdi, Prasetyo, and Masruri (2012, p. 68) show that the practice of assessing the learning outcomes especially in elementary schools, so far, is mainly focused on the cognitive assessment. The students' appreciation is shown by the rank and score in their examination. Although all educators know that the realm of education is cognitive, affective, and psychomotor (behavioral) aspects, in practice, the affective and psychomotor aspects are not given adequate attention, especially in assessing students (Khilmiyah, Sumarno, & Zuchdi, 2015). Teachers are not accustomed to assessing changes in the social attitude (affective spheres) of students of primary schools. This happens not because of the unwillingness of the educator, but because of the lack of educators' ability to describe the affective field of achievement indicators. As a consequence, the assessment does not reflect the students' overall abilities.

It is clear that the assessment of social attitude cannot be done in the same way as that of the cognitive domain (such as by giving questions). Assessment of social attitude is more directed to recording physical activities related to social interaction, not merely the ability to answer a number of questions given.

In the primary school education system which applies thematic approach, the social attitude aspect that is part of the affective domain must be assessed. This refers to the content standards in elementary schools that contain competence in social attitude reflected by the students showing honesty, discipline, responsibility, politeness, care, and confidence in interacting with family, friends, teachers, and neighbors and showing love to their own nation.

The existing assessment system is simple without sufficient indicators. The teachers have put more focus on the assessment of the cognitive aspect which has clearer construct and criteria, while the affective aspect has more complicated construct and the teachers have insufficient competence in designing the instruments of the assessment. Another obstacle is the fact that designing learning objectives in terms of affective aspects is more difficult than designing the cognitive and psychomotor aspects (Mardapi, 2012). In other words, the affective domain is difficult to define and assess because it is latent.

Based on the data collected by the researchers related to the assessment employed to assess the existing social attitudes, the models include observation methods (Syamsudin, 2015, p. 109; Waryadi, 2013, pp. 1–5), self-assessment of social attitude at the end of learning, and assessment developed by the teacher by referring to the technical guidance. These three assessments focus only on one method and tend to assess the apparent aspect of the student based on one point of view (teacher or student). This assessment also does not cover all of the aspects suggested in the core competencies of the social attitudes that the curriculum suggests. In addition, assessment which uses only one method will produce inaccurate conclusions on the social attitudes assessed.

Assessment of social attitude is often done at the end of an instruction, regardless of the process. This is done by the teacher as a routine and an attempt to execute the obligation. This kind of assessment produces only a visible social attitude at the end of learning. This will result in insufficient information, in which the results obtained are only viewed from one section of the lesson. Assessment should be done during the teaching-learning process, from the start to the end based on real or authentic condition.

In addition, an assessment applying three assessment methods (integrated) has not been conducted. Thus, this research is very important to do because by doing the assessment integrating self-assessment, peer assessment, and observational assessment, the results will be more adequate.

Method

This research is explorative descriptive research that describes the social attitude of elementary school students using three forms of self-assessment (SA), peer assessment (PA), and observational assessment (OA) instruments. The instrument validity was done using the confirmatory factor analysis (CFA), seen from the estimated loading factor per item. The result of the grain loading factor is between 0.31-0.99 (> 0.30) which means that the item in social attitude instrument (SA, PA, and OA) is valid. The use of validity criteria was seen at the loading factor of at least 0.30 as the consideration referring to Azwar (2015, p. 143). The Alpha Cronbach approach was used to estimate the reliability of the instrument, obtaining the reliability value between 0.788 and 0.886 (> 0.70). This requirement refers to Nunally (1981), Sunyoto (2012), and Mardapi (2017) who state that an instrument is said to be reliable when the combined coefficient of grains (alpha reliability) is 0.70 or more.

The population in this research was the students of elementary schools in Yogyakarta which have been implementing Curriculum 2013 for two years. A sample of 58 students of Kaliagung Elementary School in Sentolo, Kulonprogo Regency and Pakel Elementary School was established using the cluster

random sampling technique. The two schools were chosen because they have been implementing Curriculum 2013 based on thematic learning and conducting affective assessment.

The data were collected using questionnaires for SA and PA, and observation sheets for OA. The questionnaire and observation data were complementary and integrated. The data obtained were analyzed to describe the students' achievement in social attitude. The achievement in social attitude was divided into two parts: (1) achievement based on honesty, discipline, responsibility, politeness, care, and confidence components, and (2) the achievement of social attitude as the combination of all social attitude components, referring to the results of the social attitude of elementary students. There is also a categorization of social attitude as a whole by combining all of the three forms of assessment used in this research.

The data analysis was done through categorization of assessment results using score, average, and standard deviation. The data were derived from overall scores obtained by the respondents. The data obtained were analyzed using the categorization suggested by Mardapi (2012) as stated in Table 1.

This categorization was used to assess the social attitude in detail based on the honesty, discipline, responsibility, politeness, care, and confidence components. This categorization also helps the teacher in monitoring the students' ability to absorb thematic learning outcomes especially in the affective aspect. The assessment result of each component was then continued with the assessment of the students' social attitude, which was the integration of all components.

To understand and interpret the assessment results of the social attitude using the three models in this research, the researcher made a description to get the understanding of the social attitude components performed by the students. The description helped the teacher to reveal the achievement of social attitude, as stated in Table 2.

Table 1. Categorization of components of students' social attitude

No.	Student's Score	Categorization of students' social attitude
1.	$X \geq \bar{x} + 1.SBx$	Entrust (A)
2.	$\bar{x} + 1.SBx > X \geq \bar{x}$	Developing (B)
3.	$\bar{x} > X \geq \bar{x} - 1.SBx$	Seen (C)
4.	$X < \bar{x} - 1.SBx$	Not yet seen (D)

Notes:
$\bar{x}$: the average score of all students in a class
SBx: standard deviation of the overall score of students in one class
X: score achieved by students

Table 2. Description of students' social attitude achievement

No	Assessed Aspect	Achievement	Description
1.	Social attitude components (honesty, discipline, responsibility, politeness, care, and confidence)	Entrust	Students consistently show social attitude (honesty, discipline, responsibility, politeness, care and confidence*) in daily life and interaction at school.
2.		Developing	Students often show social attitude (honesty, discipline, responsibility, politeness, care and confidence*) in daily life and interaction at school.
3.		Seen	Students start to show social attitude (honesty, discipline, responsibility, politeness, care and confidence*) in daily life and interaction at school.
4.		Not yet seen	Students have not yet shown show social attitude (honesty, discipline, responsibility, politeness, care and confidence*) in daily life and interaction at school.

*choose one based on the component being assessed.

Table 3. Categorization of students' social attitude

No.	Student's Score	Category of Social Attitude Achievement
1.	$X \geq X + 1.SBx$	SB (*Sangat Baik*/ Very Good)
2.	$X + 1. SBx > X \geq X$	B (*Baik*/ Good)
3.	$X > X \geq X - 1. SBx$	CB (*cukup baik*/ Fair)
4.	$X < X - 1. SBx$	KB (*kurang baik*/ Poor)

Notes:
$\bar{x}$: the average score of all students in a class
SBx: standard deviation of the overall score of students in one class
X: score achieved by students

Table 4. Description of the achievement of students' social attitude

No.	Assessed Aspect	Achievement	Description
1.	Social Attitude	SB (very good)	Students are always honest during the learning process and social interaction, disciplined in daily life, show responsibility for the tasks and duties. The students are also polite to the teachers and friends, show care to others and environment, and also show high confidence in the class. All of those aspects are entrusted.
2.		B (good)	Students are often honest during the learning process and social interaction, disciplined in daily life, show responsibility for the tasks and duties. The students often show polite behavior to the teachers and friends, show care to others and environment, and also show high confidence in the class. All of those aspects are developed.
3.		CB (fair)	Students sometimes show honesty during the learning process and social interaction, discipline in daily life, and responsibility for the tasks and duties. The students are sometimes polite to the teachers and friends, show care to others and environment, and also show high confidence in the class. All of those aspects start to emerge and be seen.
4.		KB (poor)	Students have not shown honesty during the learning process and social interaction, have not been disciplined in daily life, and have not shown responsibility for the tasks and duties. The students are less polite to the teachers and friends. They also have not given care to others and environment or performed high confidence in the class. All of those aspects are not yet seen or observed.

Students' social attitude (honesty, discipline, responsibility, politeness, care, and confidence) was derived from the categorization presented in Table 3. To figure out the meaning of the results of the social attitude assessment, Table 4 presents the description of each achievement.

The next assessment was a test to know the effectiveness of the assessment done. The effectiveness is based on the criteria suggested by four experts at psychometrics, assessment, thematic learning of primary education, and psychological counselor. The consultation also involved three primary teachers. The data obtained were categorized and presented in Table 5 (Mardapi, 2012).

Table 5. Categorization of the instruments effectiveness

No.	Respondent's Score	Effectiveness Categorization
1.	$X \geq \bar{x} + 1.SBx$	Very Effective
2.	$\bar{x} + 1. SBx > X \geq \bar{x}$	Effective
3.	$\bar{x} > X \geq X - 1. SBx$	Fairly Effective
4.	$X < \bar{x} - 1. SBx$	Less Effective

Notes:
$\bar{x}$: the average score of respondents
SBx : standard deviation of the overall score of respondents
X : score achieved by the respondents

Findings and Discussion

The assessments were conducted in two qualified primary schools; they were Pakel Elementary School and Kaliagung Elementary School, involving 58 students. The data obtained were analyzed using the descriptive method and categorization. The assessment of these values was done using SA, PA, and OA instrument models. The results were analyzed to know the description of the assessment.

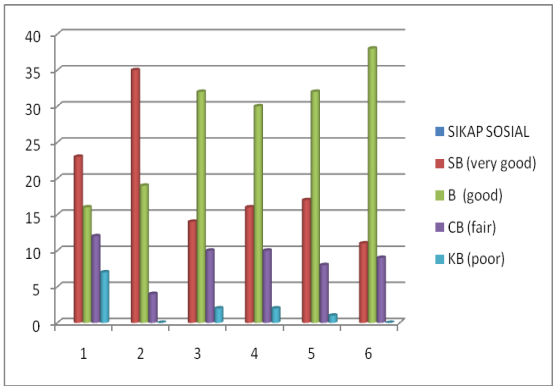

Figure 1. Social attitude viewed from six components

The results of the assessment were analyzed in two phases. The first phase presents each component. The *honesty* component or value of the primary school students is presented in Table 6.

Table 6. Social attitude value: Honesty

No.	Value	Number of Student	Percentage
1.	A (entrust)	23	39.66%
2.	B (developing)	16	27.59%
3.	C (seen)	12	20.68%
4.	D (not yet seen)	7	12.07%
Total		58	100%

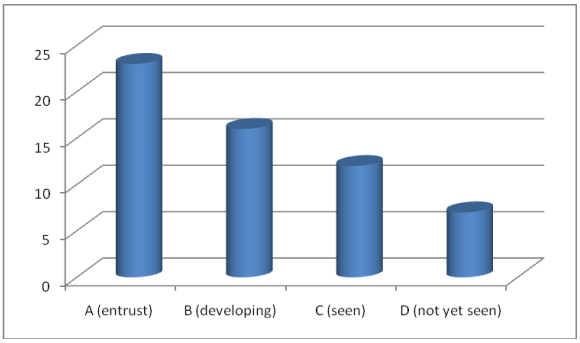

Figure 2. Histogram of results of the students' honesty assessment

Table 6 and Figure 2 show that generally the value of honesty in thematic learning from the sample of 58 students is as follows: there are 23 students (39.66%) who are in category A or entrust, 16 students (27.58%) who are in category B or honesty is developing, 12 students (20.68%) in category C or honesty starts to be observed, and seven students (12.07%) in category D which means that their honesty has not been shown.

The next value is discipline. The detailed results can be seen in Table 7.

Table 7. Social attitude value: Discipline

No.	Value	Number of Student	Percentage
1	A (entrust)	35	60.34%
2	B (developing)	19	32.76%
3	C (seen)	4	6.90%
4	D (not yet seen)	0	0%
Total		58	100%

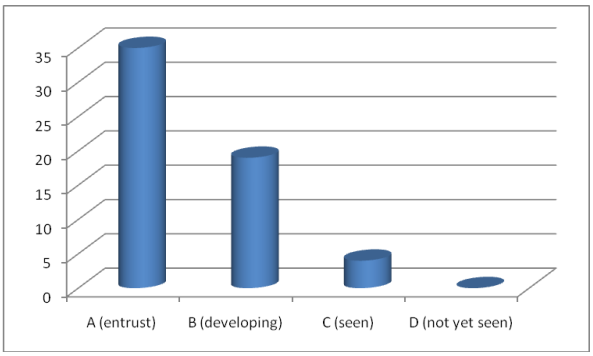

Figure 3. Histogram of results of the student's discipline assessment

Table 7 and Figure 3 indicate that from the sample students, their discipline in thematic learning is categorized as follows: there are 35 students (60.34%) who are in category A or entrust, 19 students (32.76%) who are in category B or developing, four students (6.90%) who are in category C, and no student in category D.

Table 8. Social attitude value: Responsibility

No	Value	Number of Student	Percentage
1	A (entrust)	14	24.14%
2	B (developing)	32	55.17%
3	C (seen)	10	17.25%
4	D (not yet seen)	2	3.44%
Total		58	100%

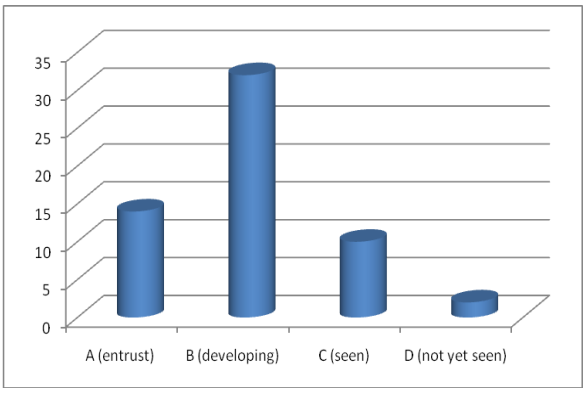

Figure 4. Histogram of results of the students' responsibility assessment

Table 8 and Figure 4 indicate that generally from the sample students, it can be seen that there are: 32 students (55.17%) in category B (responsibility is developing), 14 students (24.14%) in category A which means that responsibility is entrusted, 10 students (17.25%) in category C where responsibility starts to emerge, and two students (3.44%) in category D.

Table 9. Social attitude value: Politeness

No	Value	Number of student	Percentage
1	A (entrust)	16	27.58%
2	B (developing)	30	51.73%
3	C (seen)	10	17.24%
4	D (not yet seen)	2	3.45%
Total		58	100%

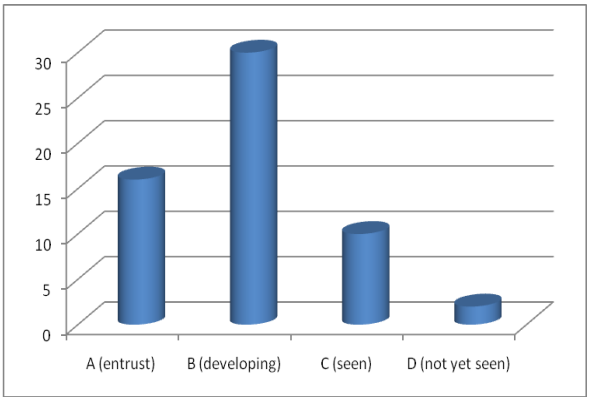

Figure 5. Histogram of the results of students' politeness assessment

Table 9 and Figure 5 indicate that from the sample students involved, it can be seen that there are 30 students (51.73%) who are in category B or developing, 16 students (27.58%) who are in category A which means that politeness is already instilled. In addition, there are 10 students (17.25%) who are in category C, and two students (3.44%) who are in category D, which means that the students have not shown polite behavior in thematic learning.

Table 10. Social attitude value: Care

No	Value	Number of student	Percentage
1	A (entrust)	17	29,31%
2	B (developing)	32	55,17%
3	C (seen)	8	13,79%
4	D (not yet seen)	1	1,73%
Total		58	100%

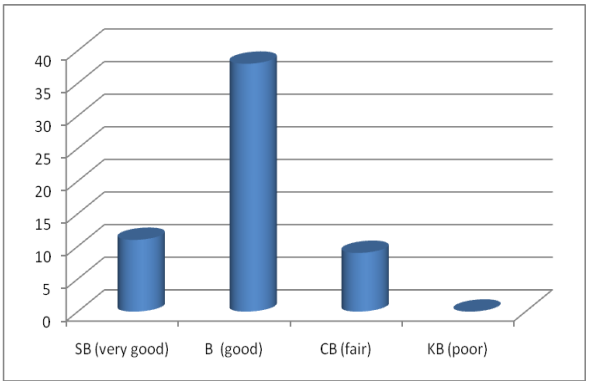

Figure 6. Histogram of the results of students' care assessment

Table 10 and Figure 6 show that the results of the assessment of students' care are as follows: 32 students (55.17%) are in category B, 17 students (29.31%) are in category A, eight students (13.79%) are in category C, and one student (1.73%) is in category D. In addition, Table 11 and Figure 7 indicate that from the sample students involved, the results of the confidence assessment are as follows: 46 students (79.31%) are in category A or instilled, nine students (53%) are in category B or developing, one student (1.72%) is in category C, and two students (3.44%) are in category D or not showing self-confidence.

Table 11. Social attitude value: Confidence

No	Value	Number of student	Percentage
1	A (entrust)	46	79.31%
2	B (developing)	9	15.53%
3	C (seen)	1	1.72%
4	D (not yet seen)	2	3.44%
Total		58	100%

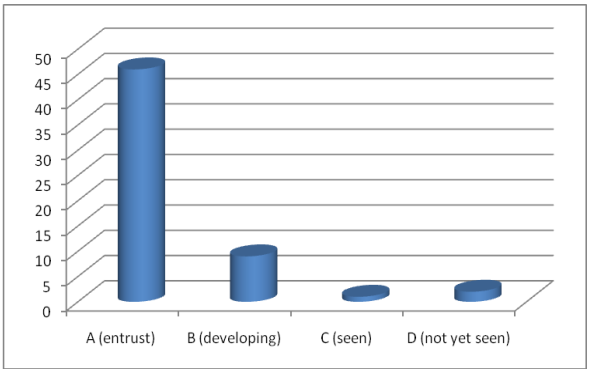

Figure 7. Histogram of the results of the students' confidence assessment

The second phase of analysis in this research dealt with the description of the assessment results of students' social attitude in the thematic learning. The results are the integration of the three assessment models employed in this research (SA, PA and OA). The results are presented in Table 12.

Table 12. Description of the students' social attitude assessment

No	Value	Number of student	Percentage
1	SB (very good)	11	18,96%
2	B (good)	38	65,52%
3	CB (fair)	9	15,52%
4	KB (poor)	0	0
Total		58	100%

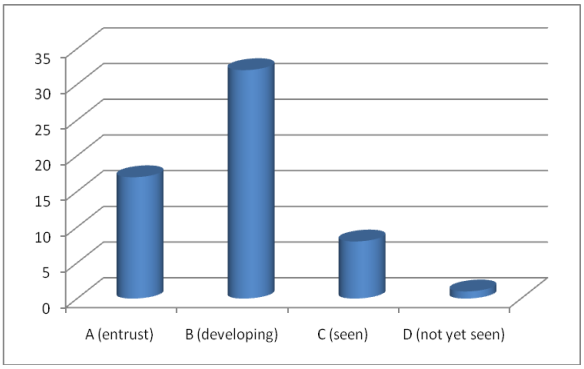

Figure 8. Histogram of the results of students' social attitude assessment

From Table 12 and Figure 8, it can be seen that the students' social attitude in thematic learning is as follows. Eleven students (18.96%) are categorized as SB or very good. There are 38 students (65.52%) included in category B or good. There are nine students (15.52%) considered as CB or fair in terms of their social attitude. There is no student categorized in category D or poor. An example of SB (very good) category is when the students are able to show honesty during a teaching-learning process and social interaction, they are disciplined in daily activities at school, they show responsibility for their tasks and duties, they show polite behavior to their teachers and peers, they care about others and environment, and they show confidence in class. All those aspects have already entrusted and instilled in students' daily life.

As previously mentioned, the results of this research are divided into two parts. The first result is the assessment based on the social attitude components, covering honesty, discipline, responsibility, politeness, care, and confidence. The second result deals with the social attitude value along with the description which can be used to fill out the report of the learning outcome. Based on the components of assessment results, it can be generally said that confidence is included in category A or *entrust* (46 out of 58 students or 79.31%). In addition, 35 students show discipline as how it is described in category A, while honesty is reflected by 23 students and is considered as being instilled. There are 32 students showing responsibility, 30 students showing care, and 32 students reflecting politeness. These three values are in category B (developing).

Another interesting result is that there are seven students (12.06%) who are categorized in category D. They have not shown honesty in their daily life and social interaction at school. The dishonesty is shown when they copied other students' work. It is in line with the idea of Koellhoffer (2009, p. 27) that honesty deals with avoiding plagiarism, including taking others' idea or answers without permission during the learning process, test, etc.

The results also present that the social attitude assessment is integrated components developing the attitudes such as honesty, discipline, responsibility, politeness, care, and also confidence. From the sample of 58 students, 11 (18.96%) are included in SB, or, in other words, their social attitude is very good. In addition, 38 students (65.52%) are considered to be good. The social attitude is the

result of responses to the social stimuli contained in thematic learning. This is supported by LaPierre in Azwar (2015, p. 5) who proposes that social situation is a pattern of behavior, anticipative tendency or readiness, predisposition to adapt to social situation, or, simply social attitude is a response towards conditioned social stimulus.

From the assessment results of the students' social attitude, it can also be inferred that their social attitude turns out to be varied. There are 36 (65.52%) students in SB (very good) category and 11 students (18.96%) in B (good) category. From that result, SB (very good) category has deep meaning.

The results can also be used in the report of the learning outcomes of core competence in social attitude aspect or *Kompetensi Inti* (KI)–2 (Core-Competence 2) and become the evaluation material for thematic learning. The assessment results obtained are also used by teachers to fill out the report of the learning outcomes in the mid semester and the end of the semester.

This research also yields effectiveness from the assessment conducted. There are 79% of the teachers who claim that the assessment involving three different models in this research is effective. This indicates that more varied and integrated methods can result in more accurate assessment results. This shows that this instrument is useful in helping teachers to assess social attitude as an affective component of integrated thematic learning outcomes in primary school.

Conclusion and Suggestion

Conclusion

The results of this research are divided into two parts. The first result is the assessment based on the components of social attitude covering honesty, discipline, responsebility, politeness, care, and confidence. The second result deals with the social attitude value along with the description which can be used to fill out the report of the learning outcome.

For teachers, this assessment can be used to fill in the report of students' learning outcomes in the affective domain or KI 2 (Core-Competence 2). For parents and students, the assessment results are helpful in finding out the description of social attitude that has been achieved by students. This description can be used as an introspection and improvement of students' social attitude.

Suggestion

The comprehensive results of this research may become a guidance for the teachers to assess students' social attitude. The existing assessment can also become an evaluation towards the learning practice. The future research should reveal other components of social attitude as the results of learning process.

References

Ahmadi, H. A. (2002). *Psikologi sosial*. Jakarta: Rineka Cipta.

Azwar, S. (2009). *Penyusunan skala psikologi*. Yogyakarta: Pustaka Pelajar.

Azwar, S. (2015). *Skala pengukuran sikap manusia*. Yogyakarta: Pustaka Pelajar.

Bernann, S. L. (2015). *Pengetahuan, sikap, dan perilaku manusia*. Yogyakarta: Parama.

Crano, W. D., & Prislin, R. (2011). *Attitudes and attitude change*. New York, NY: Psychology Press.

Ekowarni. (2009). *Pedoman pendidikan akhlak mulia siswa sekolah dasar*. Jakarta: Departemen Pendidikan Nasional, Direktorat Pendidikan Dasar dan Menengah.

Gerungan, W. A. (2004). *Psikologi sosial*. Bandung: Refika Aditama.

Khilmiyah, A., Sumarno, S., & Zuchdi, D. (2015). Pengembangan model penilaian keterampilan intrapribadi dan antarpribadi dalam pendidikan karakter di sekolah dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 1–12. https://doi.org/10.21831/pep.v19i1.4550

Koellhoffer, T. T. (2009). Character education being fair and honest. New York: Infobase Publishing.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1973). *Taxonomy of educational objectives Book 2/Affective domain*. New York, NY: Longmans, Green.

Kurniawan, D. (2014). *Pembelajaran terpadu tematik (Teori, praktik, dan penilaian)*. Bandung: Alfabeta.

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

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

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications*. New York, NY: Springer.

Nunally, J. C. (1981). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *REiD (Research and Evaluation in Education)*, *2*(2), 155–164. https://doi.org/10.21831/reid.v2i2.11029

Stiggins, R. J. (1999). Assessment, student confidence, and school success. *The Phi Delta Kappan*, *81*(3), 191–198.

Sunyoto, D. (2012). *Validitas dan reliabilitas*. Yogyakarta: Nuha Medika.

Supardan, D. (2011). *Pengantar ilmu sosial: Sebuah kajian pendekatan struktural*. Jakarta: Bumi Aksara.

Syamsudin, A. (2015). *Model penilaian afektif siswa sekolah dasar*. Doctoral dissertation, Universitas Negeri Yogyakarta, Yogyakarta.

Waryadi. (2013). *Menyiasati pelaksanaan penilaian sikap dalam implementasi kurikulum 2013*. Jakarta: Balitbang Kemenag.

Zuchdi, D., Prasetyo, Z. K., & Masruri, M. S. (2012). *Model pendidikan karakter terintegrasi dalam pembelajaran dan pengembangan kultur sekolah*. Yogyakarta: UNY Press.

REiD (Research and Evaluation in Education), 4(1), 2018, 22-34
Available online at: http://journal.uny.ac.id/index.php/reid

A factor analysis of an instrument for measuring physical abuse experience of students at school

[*1]Safrudin Amin; [2]Badrun Kartowagiran; [3]Pracha Inang
[1]Faculty of Literature and Culture, Universitas Khairun
[1]Jl. Pertamina Kampus II Unkhair Gambesi Kota Ternate Selatan, 97719, Indonesia
[2]Faculty of Engineering, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman 55281, Yogyakarta, Indonesia
[3]Faculty of Education, Burapha University
169 Longhaad Bangsaen Road, Saensook, Mueang, ChonBuri 20131, Thailand
[*]Corresponding Author. E-mail: safrudinamin@gmail.com

Submitted: 24 May 2018 | Revised: 04 August 2018 | Accepted: 08 August 2018

Abstract

Violence in schools is increasingly reported by the mass media. It indicates that its prevalence is escalating. An instrument which has a proper psychometric property is needed to investigate the phenomenon. The study aims to develop an instrument for measuring physical abuse experienced by students in schools and explore the construct of the instrument. To pursue those objectives, the content validitity, construct validity, and reliability analysis on the developed instrument were measured. Its content validity was confirmed through expert judgment, construct validity was proven through exploratory factor analysis, and reliability was estimated through Cronbach's alpha coefficient. Experts considered that the content of all items were relevant, though they also suggested some improvement in wordings for greater clarity. The exploratory factor analysis on 31 items indicates that seven items need to be dropped and 24 items are divided into three factors called (1) victimized by friends with the loading factor ranging from 0.44 to 0.69, (2) victimizing friends with the loading factor ranging from 0.45 to 0.66, and (3) being victimized by teachers with the loading factor ranging from 0.57 to 0.68. The reliability of the test was 0.874. Based on this result, the developed instruments consist of three factors with good validity and reliability.

Keywords: *physical abuse, student, school, validity, reliability*

Introduction

Studies on abuses against children have been conducted by individuals as well as organizations. Research findings show the increasing incidents of violence against children in this country. This alarming trend, however, has not attracted serious attention from those in power (Idris, 2015). One of the most important concerns of violence against children is the violence which takes place at school.

A survey conducted by *Plan International* and *International Center for Research on Women* (ICRW) shows 84% of Indonesian children experience abuse in school. This result is higher than the trend in Asian region which is 70% (Qodar, 2015). The data released in June 2015 by the Commission of Indonesian Children Protection (*Komisi Perlindungan Anak Indonesia* – KPAI) show that from 2011 to April 2015, violence against children grew significantly. In 2012, a survey in nine provinces demonstrated that 87.6% students experienced abuse in school (Setyawan, 2015). The data published by KPAI in November 2017 show that violence against children in school is mounting. As many as 84% students, or eight out of ten students, have ever experienced abuse in school. Among them, 45% male students report that their teachers or school staff are the persecutors (Setyawan, 2017).

ISSN 2460-6995

Before proceeding further, it is important to assert the terms 'violence', 'abuse', 'maltreatment', 'bullying', and the like. Many studies by organizations or individuals have used the terms interchangeably although they refer to the same phenomena, or some concepts are treated as part of other concepts. The UN Secretary-General's Study defines violence against children in line with article 19 of the CRC which treats 'abuse', 'maltreatment', and 'exploitation' as parts of violence (UNICEF, 2014b, p. 2). World Health Organization (WHO) equates the concept of 'abuse' and 'maltreatment' (UNICEF, 2014a, p. 19). It defines child abuse or maltreatment:

> *'…constitutes all forms of physical and/or emotional ill-treatment, sexual abuse, neglect or negligent treatment or commercial or other exploitation, resulting in actual or potential harm to the child's health, survival, development or dignity in the context of a relationship of responsibility, trust or power'*

UNICEF (2014a, p. 21) made inventory of studies on violence against children and grouped together studies using different terms such as 'physical violence', 'physical abuse', and 'physical maltreatment' into one category that is physical dimension of violence for the reason that they are dealing with roughly the same phenomena. Here, 'violence', 'abuse', and 'maltreatment' are regarded as identical.

UNICEF (2014a, p. 21) also includes 'bullying' as part of 'violence'. Nansel et al. (2001) define bullying as aggressive behavior which is intended to harm or disturb, committed by a more powerful person or group to those who are powerless, and it occurs repeatedly over time. The characters of intention to harm others and asymmetric power between the persecutors and their victims overlaps with the definition of 'violence' held by WHO which also emphasizes the intention to harm others by using power (UNICEF, 2014a, p. 19).

Nansel et al. (2001, p. 2094) also state that bullying behavior could be verbal, psychological, or even physical. Rivers and Smith (1994, p. 362) find that physical abuse in bullying could be in the forms of '..direct-physical behaviours such as hitting, kicking, and stealing'. NSPCC (2016, p. 7) uses the term 'physical bullying' to refer to kicking, hitting, biting, pinching, hair pulling, and making threats'. This clearly shows that 'bullying' is equal to 'physical abuse', and it is reasonable that UNICEF includes 'bullying' as a part of 'violence'.

This research adopts WHO's definition of child abuse mentioned earlier since it offers a notion that abuse or maltreatment does not always result in actual harm but could also be in a form of potential harm. However, as we go further to discuss physical abuse in this section, it will become clearer that our point of emphasis is not on the effects of violence acts as asserted by WHO, but on the acts of violence themselves.

Apart from the conceptual problem, in general, many experts come to a conclusion that any kinds of abuse against children committed either by teachers or fellow students in school, or abuse taking place outside school, has a destructive impact on children's academic performance in school, in addition to other forms of negative impacts faced by the children. Hyman and Perone (1998, p. 19) explain that many studies have found that children who experience psychological maltreatment during their preschool and school age have lower academic performance. Likewise, their ability and social competence are also low, compared to those students who have not experienced such maltreatment. This is in line with Ajema, Muraya, Karuga, and Kiruki (2016, p. 2) who conclude that violence in school and associated fear, anxiety, and injuries contribute to poor education and health outcomes. According to them, violence in school can lead to the destruction of children's capacity and potentials to take advantages maximally during their education processes because they tend to be absent, unwilling to continue their study, and weakly motivated to get academic achievement.

Nansel et al. (2001) summarize that bullying has a significant correlation with academic achievement. Both the persecutors and victims show low academic achievement compared to those students who are not involved in abuse. Quoting some studies, Simpson (2015, p. 18) also confirms that abuse such as

bullying can badly affect student's academic performances. Cohn and Canter (2003) find that *bullying* causes the victims to face difficulties in dealing with academic challenges in school, and both perpetrators and victims have strong correlation with drop-out incidence. In addition, based on some studies, United Nations Secretary-General's Study (2006, p. 130) also synthesizes that 'physical and psychological punishment, verbal abuse, bullying and sexual violence in schools are repeatedly reported as the reasons for absenteeism, dropping-out, and lack of motivation for academic achievement'.

So far, violence against children has been reflected in various terms, such as 'child abuse', 'violence against children', 'maltreatment', 'bullying', and some more. However, the aspects of abuse are rather well-accepted by different organizations and scholars. Choo, Dunne, Marret, Fleming, and Wong (2011) divide child abuses or the victimization of children into four categories: physical abuse, sexual abuse, emotional abuse, and neglect. In line with that notion, Law No. 35 of 2014 of Republic of Indonesia also states that the aspects of child abuse are physical, psychological, sexual, and negligent.

Moreover, experts provide the detailed aspect of physical abuse. Muthmainnah (2014, p. 446) states that 'physical abuse occurs when an adult (parent, educator, caregiver, etc.) injures a child physically such as hitting, pinching, kicking, slapping, etc.' Clark, Clark, and Adamec (2007, p. 203) define physical abuse as 'an act of commission by a parent or other persons that may or may not be accidental and that results in physical injury.' Besides, WHO claims that:

> *'physical abuse of a child is that which results in actual or potential harm from an interaction or lack of an interaction, which is reasonably within the control of a parent or person in a position of responsibility, power or trust. There may be a sigle or repeated incidents' (UNICEF, 2014a, p. 20).*

The adjective term 'physical' in 'physical abuse' has allowed the birth of various derivative terms such as physical violence, physical assault, physical harassment, physical victimization, physical maltreatment, physical bullying, and the like, but all refer to the *threats or harmful actions that make the victim's physicality a target, whether it causes physical injury or not*. The definition emphasizes on the *acts* of violence or abuses rather than the *results* of the acts. This position is fully reflected in the instrument developed in this study.

In the context of research on child abuse in school, this instrument development is considered to be crucial for two reasons. First, there is a clear evidence of the increasing number of child abuse in school, including physical abuse, which has potential destructive impacts on students. Second, studies on violence in school frequently do not make public the detailed psychometric properties of their instruments. In order to be useful, the instrument developed must have good validity and reliability to ensure its accuracy and internal consistency.

Method

This study selected 584 respondents, who were grade IX students of three junior high schools in Ternate, North Maluku, Indonesia. They were asked to fill out a questionnaire concerning their experiences of physical abuse in their schools. Out of the total sample, 577 responses were feasible to be analyzed. This research adopted several items from previous studies (Choo et al., 2011; Straus, Hamby, Boney-McCoy, & Sugarman, 1996; Straus, Hamby, Finkelhor, Moore, & Runyan, 1998; UNICEF, 2014a), tailored and modified them to meet its specific objectives.

The items in the questionnaire were ranked and scored using a modified Likert scale. The respondents were asked to choose one of the responses offered. The response categories were: *never* = 1, *seldom* = 2, *sometimes* = 3, *frequently* = 4, and *always* = 5. All items are cast in positive terms. The 31 items addressed three different aspects assumed to be the aspects of child physical abuse. The three aspects are abuses committed by teachers, abuse committed by fellow students, and abuse committed by respondents to other students.

The content validity was confirmed through expert judgment to ensure its rele-

vance to the construct to be measured. Three reviewers reviewed the first draft of the instrument and provided their input to improve the quality of the instrument. Each item was accompanied by five alternative responses and each reviewer had to score the item by choosing an alternative answer ranging from 1 = irrelevant, 2 = rather relevant, 3 = relevant enough, 4 = relevant, and 5 = very relevant. Experts considered the content of all 31 items were either relevant or very relevant, and suggested some improvements in wordings for more clarity.

The construct validity was ensured through the exploratory factor analysis (EFA). The exploratory analysis employed orthogonal rotation carried out with the varimax approach. The reliability of the instrument was estimated using coefficient alpha. Both analyses were performed using SPSS 23 for Windows.

Findings and Discussion

Exploratory Factor Analysis

Basically, similar to many previous studies, construct validity can be proven by employing confirmatory factor analysis (CFA) (Widdiharto, Kartowagiran, & Sugiman, 2017) and or exploratory factor analysis (Clemens, Carey, & Harrington, 2010). This study employed exploratory factor analysis (EFA) in order to explore the dimensions or factors in the instrument based on the empirically collected data (Kartowagiran, 2008, p. 188).

The results of initial check show that the instrument has the value of 0.89 in Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. This value is bigger than the minimum required score of 0.5. The significance indicated by the value of sig. is $0.000 < 0.05$. All the values suggest that the data collected by using this physical abuse instrument were suitable for factor analysis. The next analysis involved factor extraction, factor rotation, interpretation of the result, reliability estimation, and naming the factors.

The purpose of factor extraction is 'to determine the number of initial subsets or *factors* that appear to represent the dimensions of the construct which is being measured' (Pett, Lackey, & Sullivan, 2003, p. 85). There are some factor extraction methods available for factor analysis, and some of them have been available in statistical soft wares such as Statistical Package for the Social Sciences (SPSS) or Statistical Analysis System (SAS). This article prefers the principal component analysis (PCA) to other methods, although some methodologists are not convinced of the use of PCA for various reasons. Costello and Osborne (2005, p. 2), for example, write 'component analysis is only a data reduction method', and it does not 'regard to any underlying structure caused by latent variables', etc.

Although many criticisms stand against the use of principal component analysis, Kline (2008, p. 74) sees that 'principle factor analysis seems to be a sensible choice' in factor analysis. Besides, the use of principle component analysis is the most popular one probably due to the fact that some statistics software packages use it as their default (Costello & Osborne, 2005, p. 2), and also its result is easier to interpret compared to other methods (Pett et al., 2003, p. 102).

There are some common approaches to determination of the number of extracted factors to be retained (Fabrigar & Wegener, 2012, pp. 53–67). This study, however, applied three of them which were considered to be the most common procedures, namely, eigenvalues greater than 1, percentage of variance explained, and the use of scree plot. These methods are the most frequently used in determining factor solution in the form of unrotated factor solutions. Although these approaches sometimes 'do not provide meaningful and easily interpretable clusters of items' (Pett et al., 2003, p. 131), they are most commonly used in the stage of factor extraction before processing factor rotation.

One of the results of factor extraction is table of Total Variance Explained. In this study (see Table 1), the formation of seven factors or components with eigenvalues > 1. Factor one has the eigenvalue of 7.550, factor two has the eigenvalue of 1.983, and factor three has eigenvalue of 1.827. The fourth factor has the eigenvalue of 1.528, the fifth factor has 1.218, sixth factor has 1.064, and the

Table 1. Total variance explained in seven-factor model

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.550	24.354	24.354	7.550	24.354	24.354	3.783	12.203	12.203
2	1.983	6.395	30.749	1.983	6.395	30.749	2.913	9.398	21.601
3	1.827	5.894	36.643	1.827	5.894	36.643	2.700	8.708	30.309
4	1.528	4.929	41.572	1.528	4.929	41.572	2.129	6.868	37.177
5	1.218	3.928	45.499	1.218	3.928	45.499	1.801	5.808	42.985
6	1.064	3.431	48.931	1.064	3.431	48.931	1.528	4.928	47.913
7	1.019	3.286	52.216	1.019	3.286	52.216	1.334	4.303	52.216
8	.978	3.156	55.372						
9	.917	2.958	58.331						
10	.888	2.865	61.196						
11	.843	2.719	63.915						
12	.800	2.581	66.496						
13	.790	2.549	69.045						
14	.759	2.448	71.493						
15	.720	2.322	73.815						
16	.665	2.146	75.961						
17	.653	2.107	78.069						
18	.628	2.026	80.095						
19	.610	1.967	82.062						
20	.587	1.893	83.955						
21	.562	1.813	85.769						
22	.534	1.721	87.490						
23	.527	1.700	89.190						
24	.509	1.641	90.831						
25	.484	1.561	92.392						
26	.475	1.531	93.923						
27	.433	1.395	95.318						
28	.391	1.261	96.580						
29	.376	1.213	97.793						
30	.368	1.186	98.979						
31	.317	1.021	100.000						

Extraction Method: Principal Component Analysis.

seventh factor has 1.019. Using this procedure, the number of components with eigenvalues > 1 would be counted as the number of the extracted factors which later are specified into the model (Fabrigar & Wegener, 2012, p. 55).

The second approach is the percentage of variance explained by each component. The table of total variance explained shows seven components, each of which has different values of the variance explained. Component one accounted for 24% of variance, component two accounted for 6.3% of variance, factor three explained 5.8% of the variance. The rest four factors explained 4.929%, 3.928%, 3.431%, and 3.286% consecutively of the variance. This seven-factor model solution explained 52.216% of the variance in the table.

The last approach used to determine the number of extracted factors was scree plot (see Figure 1). The scree test basically examines 'the graph of the eigenvalues and looking for the natural bend or breaks point in the data where the curve flattens out' (Costello & Osborne, 2005, p. 3). Although the interpretation of the scree plot is subjective in nature (Fabrigar & Wegener, 2012, p. 58; Kline, 2008, p. 75), Gorsuch proposes scree plots over the Eigen-value > 1 as the criteria (Pett et al., 2003, p. 120). With reference to that criterion, it is difficult to assume the formation of seven factors, since only four factors have eigenvalues > 1.

In terms of the variance explained, although many researchers stop the factor extraction process when the total variance explained reaches 50-80%, there are no definite guidelines for a particular threshold (Pett et al., 2003, p. 116). Hair et al. (1995) give criteria of the last factor no less than 5% of the explained variance (Pett et al., 2003, p. 116). Although it is intended for natural science, in this study, it is quite relevant as can be seen in

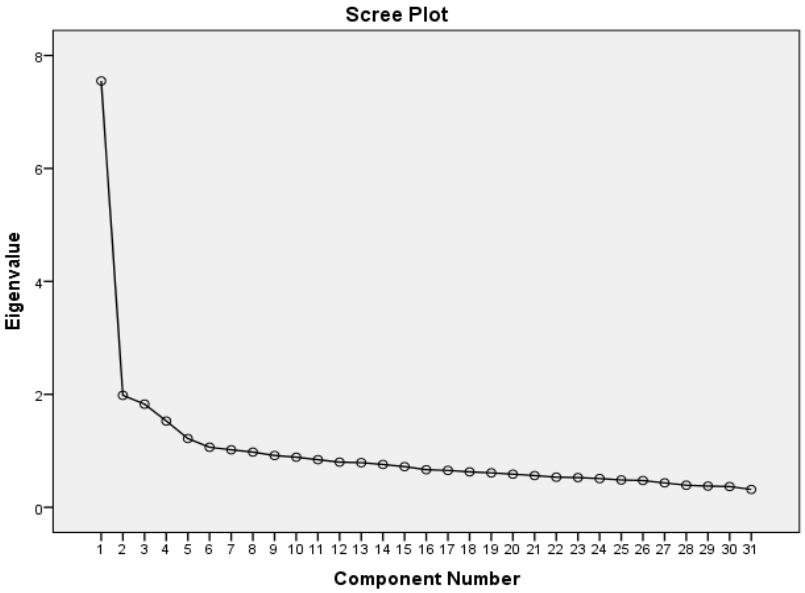

Figure 1. Scree plot of seven-factor model

the rest of this article. By applying that criterion to the above seven-factor model, it appeared that only the first three factors met the 5% criterion, despite the fact that seven factors had eigenvalue > 1. Another problem was that the scree plot of this seven-factor model also could not show clearly the appearance of seven factors, instead, it showed only four factors. Using a straight line drawn with a ruler through the lower values of the plotted eigenvalues, it identified only four factors formed above the line.

Dealing with the inconsistent outcomes of the above three approaches in determining the number of factors, and the difficulties in interpreting their results, like many other researchers, we relied on rotation to improve the meaningfulness and to have better interpretation of the factors generated. This study used orthogonal rotation carried out with the varimax approach. Varimax maximizes the variances of the loading in the factors.

The results of the factor rotation are component matrix and rotated component matrix. In component matrix, all factor loadings of each item in each factor are shown without discriminating them based on high loadings only. Consequently, it includes all items-to-factor correlations. It, therefore, becomes overwhelming and rather difficult to interpret. In rotated component matrix, only high loading factors appear in each factor or component. At this point, a researcher can decide the value of factor loading allowed in the factor solution. In our seven-factor model, following Sadtyadi and Kartowagiran (2014, p. 295), we suppressed the absolute values of factor loadings to less than 0.50 and maintained > 0.5. This helped to provide only high factor loadings (> 0.5) in each item. As a result, the loadings appeared were not overwhelming. Pedhazur and Schmelkin state that ideally, each item has high and meaningful factor loading on one factor only and each factor has high or meaningful loadings for only some of the items (quoted in Pett et al., 2003, pp. 132–133).

The output of the rotated solution clearly showed that items were grouped into seven components or factors. Except for three items with loadings factor less than 0.5, the items were distributed to seven components or factors. There was no crossloading item in this solution but several problems occurred. The first problem emerged because some different items carrying different conceptual meanings were grouped together, particularly in components 4 and 5. In terms of conceptual inappropriateness, some items were loaded on irrelevant factors or, in other words, some items failed to load on conceptually appropriate factors. This was considered as an indication of incorrect factor structure (Costello & Osborne, 2005, p. 5).

The second problem came to light from the fact that two components, 6 and 7, were supported only by 2 and 1 items consecutively. Costello and Osborne (2005, p. 5) state that 'a factor with fewer than three items is generally weak and unstable'.

The unsatisfying appearance of the seven-factor solution led us to seek other solution which was expected to be meaningful conceptually and easy to interpret. Looking back to some indications shown in the seven-factor model, particularly in the rotated component matrix, only the first three components or factors were easy to interpret and turned up to be conceptually more appropriate. In addition, by using the 5% criteria of variance extracted proposed by Hair et al (Pett et al., 2003, pp. 116–118), we found only three first factors met the criterion of 5%. This is a strong indication of the existence of a three-factor solution.

In addition, we also linked this indication of three factors to the initial constructs in the physical abuse questionnaire and mapped the main issues in it. The instrument, in fact, contains three main issues i.e. abuse committed by teachers to respondents, abuse committed by fellow students to respondents, and abuse committed by the respondents of the survey to their fellow students. From these considerations, the factor analysis with three factors to extract was conducted. The lowest factor loading allowed was also determined to ≥ 0.40 by suppressing the items that have factor loadings of less than 0.40.

The results showed that the three-factor analysis met the 5% criterion for each factor (as proposed by Hair in earlier discussion). The variance explained by the three factors, however, was only 36.643%, lower than the variance explained by the seven-factor model. To solve this low variance explained, we tried to accommodate more items by cutting down the lowest factor loading to 0.30. The solution resulted from that decision, however, became more difficult to interpret. The analysis, therefore, was dragged back to ≥ 0.40. With this threshold, the result revealed that the loadings of some four items disappeared due to having factor loadings of less than 0.40 and three items loaded in inappropriate factors (this was fewer than the number of items loaded in inappropriate factors in the seven-factor model). Those problematic seven items were then eliminated. Therefore, the number of items declined from 31 to 24 items.

After dropping these problematic items, we changed the sampling adequacy measured by Kaiser-Meyer-Olkin (KMO) into 0.891. The Barlett's test was still significant $0.000 < 0.5$. The elimination of some items did not give negative impact on the data as a whole because both values of KMO and significance indicated that the data were suitable for factor analysis. The decision to eliminate those problematic items, in fact, improved the factor structure given that the variance explained increased from 36.643% to 41.428%.

Another effect of eliminating some problematic items was that the number of the factors with eigenvalues > 1 decreased to five factors (previously seven factors). Although the decreased number of factors was accompanied by an increase in variance explained, the criterion of determining the number of factors to retain was based more on the criterion that the factor has no less than 5% accounted for variance as proposed by Hair et al (Pett et al., 2003, p. 116). Besides, the appearance of the scree plot and theoretical considerations of the original constructs contained in the questionnaire were also the basis for our decision. With regard to the criterion of > 5%, the data in Table 2 clearly show the formation of three factors.

In terms of the 5% criterion, the three model solutions prove that only the first three factors have higher than 5% of the variance extracted. Factor one accounts for 26.288% of variance and has an eigenvalue of 6.309, factor two accounts for 8.016% of the variance and its eigenvalue is 1.924, and the third factor's eigenvalue is 1.710 and it accounts for 7.123% of the variance. The fourth and fifth factors, although have eigenvalues of 1.165 and 1.050 respectively, which are higher than 1, each of their contributions to the explained variance is only 4.853% and 4.374%, less than 5%. These lead to their exclusion from the factors retained. As a whole, the three factors account for 41.428% of variance.

Table 2. Total variance explained in three-factor model

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.309	26.288	26.288	6.309	26.288	26.288	4.038	16.823	16.823
2	1.924	8.016	34.305	1.924	8.016	34.305	3.004	12.518	29.341
3	1.710	7.123	41.428	1.710	7.123	41.428	2.901	12.087	41.428
4	1.165	4.853	46.281						
5	1.050	4.374	50.655						
6	.937	3.904	54.559						
7	.884	3.683	58.243						
8	.851	3.545	61.788						
9	.807	3.362	65.150						
10	.781	3.253	68.403						
11	.740	3.082	71.485						
12	.690	2.876	74.361						
13	.634	2.640	77.001						
14	.608	2.532	79.533						
15	.598	2.494	82.027						
16	.581	2.419	84.446						
17	.570	2.375	86.821						
18	.529	2.206	89.027						
19	.522	2.175	91.201						
20	.489	2.036	93.238						
21	.440	1.831	95.069						
22	.435	1.811	96.881						
23	.378	1.574	98.454						
24	.371	1.546	100.000						

Extraction Method: Principal Component Analysis.

Another method which is used to help making decision on the number of factor to keep is scree plot. Although some methodologists criticize the use of scree plot (Fabrigar, Wegener, Maccallum, & Strahan, 1999, pp. 278–279), it is one of the most widely used approaches in the exploratory factor analysis. Costello and Osborne (2005, p. 3) even state that 'the best choice for researchers is the scree test'. It is admitted that one of main problems related to the use of scree plot is that researchers tend to use their subjective nature in interpreting them. Some researchers, however, provide guidelines. Costello and Osborne (2005, p. 3) assert that 'the number of data points *above* the "break" (i.e., not including the point at which the break occurs) is usually the number of factors to retain'. Pett et al. (2003, p. 119) advise 'that point where the factors curve above the straight line drawn [with a ruler] through the smaller eigenvalues identifies the number of factors'. The scree plot presented in Figure 2 is an output based on the data processed through SPSS package.

Following the afore-mentioned guidelines of interpreting scree plot, the scree plot presented in Figure 2 clearly presents three factors above the break or above the straight line drawn from the lowest eigenvalue horizontally. In other words, the scree output shows a similar result with the 5% criterion and is also relevant to the original constructs containing three main themes in the questionnaire of physical abuses. Except for the criteria of eigenvalues > 1, all of these other criteria confirm the formation of three-factor solution model in the factor extraction.

The decision to involve theoretical considerations or original construct in determining the number of factors to retain referred to the recommendations provided by Nunnally and Bernstein (1994) paraphrased by Pett et al. (2003, p. 125) as follows:

> *How many factors should we extract... two... three... four? There is no easy solution to this decision. Nunnally and Bernstein (1994) caution the researcher against using rigid guidelines for determining the ultimate number of factors to extract. Whatever solution we arrive at should not be solely based on statistical criteria; it also needs to make theoretical sense. The ultimate criteria for determining the number of factors are factor interpretability and usefulness both during the initial extraction procedures and after the factors have been rotated to achieve more clarity.*

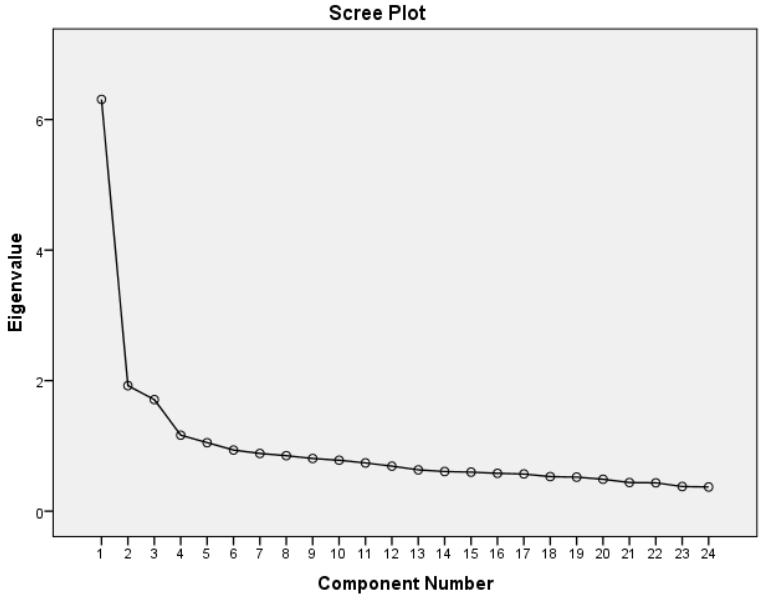

Figure 2. Scree plot of three-factor model

It was the purpose to reach factor interpretability and usefulness that led us to involve our original construct in determining the factors beside statistical inputs. The purpose also became the basis for repeatedly refining the solution and examine them to find a more suitable solution which best explained the data and disclosed the structure of constructs behind the measurable variables. The use of orthogonal rotation with varimax had generated factor loading matrix in which the items were grouped together neatly to each factor. This is, therefore, more interpretable.

There are two things worth noting here. First, the requirement of adequate numbers of items load in each factor is fulfilled. Referring to views proposed by Nunnally and Bernstein (1994), Pett et al. (2003, p. 125) write 'if the extracted factors serve to describe characteristics that variables have in common, then, by definition, there need to be at least two items for each extracted factor'. Further, Costello and Osborne (2005, p. 3) propose at least three items for each factor. Second, the factor loadings of the items ranging from 0.44 to 0.69 are good enough and even very good (Comrey & Lee, 2009, p. 243). The detailed illustration of the matrix of factor structure and item loadings can be found in Table 3.

There are some guidelines to interpret the construct validity of this instrument based on the information presented in the factor structure matrix. Some researchers employ factor loading of each item ≥ 0.30 (McCauley, Ruderman, Ohlott, & Morrow, 1994, p. 548). Comrey and Lee (2009, p. 243) propose higher than 0.30, by saying 'whereas loadings of 0.30 and above have commonly been listed among those high enough to provide some interpretive value, such loadings certainly cannot be relied upon to provide a very good basis for factor interpretation'. In addition, some even use factor loading > 0.50 (Kartowagiran & Jaedun, 2016, p. 133; Wijanto, 2008, p. 193).

According to Costello and Osborne (2005, p. 3), item loading table 'has the best fit to data' if item loadings above 0.30, no or few items cross-loadings, and no factors with fewer than three items. To meet those criteria, this study uses factor loading ≥ 0.40. In more detail, out of 24 valid items with factor loadings above 0.40, 14 of them are > 0.60, six are > 0.50, and the rest four items are > 0.40. There is no cross-loading item in the matrix which means each item is unidimensional. In addition, there are more than three items load in each factor. Due to all requirements proposed by the above methodologists, which were fulfilled well, it can be confidently affirmed that the construct validity of this instrument has been reached satisfactorily.

Table 3. Factor loading matrix of physical abuse experience among school students

No	Item Wordings	Factor 1	Factor 2	Factor 3
Fisik17	Has any student pushed your body or head harshly?	.690		
Fisik11	Has any student hit you by using any blunt objects (examples: wood, rattan, or others)?	.688		
Fisik16	Has any student tweaked your ears?	.662		
Fisik13	Has any student thrown something solid at you? (such as book or other stuff)	.646		
Fisik10	Has any student slapped you?	.622		
Fisik15	Has any student pulled your hair harshly?	.607		
Fisik14	Has any student pinched you because he/she got angry to you?	.581		
Fisik12	Has any student kicked you? (not in a jock or sport).	.555		
Fisik18	Has any student injured you?	.468		
Fisik19	Has any student scratched you?	.441		
Fisik30	Have you scratched other students?		.665	
Fisik31	Have you bitten other students?		.658	
Fisik26	Have you pulled another student's hair harshly?		.627	
Fisik27	Have you tweaked another student's ears?		.548	
Fisik25	Have you pinched other students because you are angry to him/her?		.529	
Fisik28	Have you pushed other student's body or head harshly?		.528	
Fisik29	Have you injured other students?		.499	
Fisik24	Have you thrown something solid at other students (such as book or other stuff)		.452	
Fisik5	Has your teacher pinched you because he/she is angry?			.687
Fisik2	Has your teacher hit you by using any blunt objects (examples: wood, rattan, or others)?			.676
Fisik8	Has any teacher punished you by asking you to position your body in a way that made you are physically unpleasant?			.636
Fisik7	Has your teacher tweaked your ears?			.632
Fisik1	Has your teacher hit or slapped you?			.628
Fisik4	Have your teachers thrown something (such as book or other stuff) at you?			.578

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Usually, the factor's name is drawn from the name of the item with the highest factor loading. In the case of this study, however, it is much easier to give the name since the items grouped in each factor have some common themes. Ten items that load on factor 1 appear to have one common theme in spite of having different contents from one another. Every item contains a specific act of abuse such as pushing body, hitting, and tweaking, but all refer to the same topic, that is, abuse committed by fellow students. In factor 2, each of the eight items deals with specific content, but the main theme assembling the items' similarities within this factor is that the persecutors committing the abuse are the respondents who abuse other students. With the same pattern of interpretation, the six items loaded in factor 3 hold the same common theme, apart from their differences, namely abuse committed by school teachers.

Based on the mapping of the common themes reflected by the groups of items in each factor, it is reasonable to name the first factor containing items on abuses by fellow students as *victimized by friends,* the second factor covering items on abuses by respondents towards other students as *victimizing friends*, and the third factor carrying items containing abuses by teachers as *being victimized by teachers*. These names become new identities of each factor while the identity of each item is not important anymore. These identities, according to Kachigan, can be used to communicate to other people who are interested in using the instrument for their own research or in applying the results of the studies that have used the instrument (Pett et al., 2003, p. 210).

Reliability

Beside instrument validity, the instrument reliability is also important to estimate. Reliability test is part of instrument construc-

tion to make sure the instrument composed by the retained factors has good internal consistency. Reliability helps to know to what extent an instrument is free from measurement error.

In order to ensure the reliability of an instrument that has some subscales (factors), some methodologists and also researchers emphasize to estimate the coefficient alpha of each factor or subscale (Amir, 2015, p. 227; Pett et al., 2003, p. 188). Other methodologists, however, recommended to estimate the reliability of each scale as well as the entire scale. Parsian and AM (2009, p. 5), referring to Nunally and Bernstein (1994) and DeVon et al (2007), state that 'if an instrument contains two or more subscales, Cronbach's alpha should be computed for each subscale as well as the entire scale.' For this reason, in order to estimate the instrument reliability of the student's experience of physical abuses in school, first, the researchers generated the coefficient alpha for the whole items involving the three factors together, entire scale, then we generated coefficient alphas of each of the three derived factors independently. The lowest but still acceptable reliability coefficient used here is ≥ 0.65 (Cohen & Swerdlik, 2009, p. 151; Nurmin & Kartowagiran, 2013, p. 189).

The result of the reliability estimation shows that the reliability for the overall physical abuse scale (when the 24 items combined) is 0.874, which is satisfactory. Coefficient alpha will not be significantly affected by any drop of item. If any item were deleted, the coefficient of the entire scale would remain higher than 0.80. Coefficient alpha for factor one with the whole 10 items is 0.830. This is stable since any removal of any item will not seriously affect the coefficient for the reason that coefficient will remain above 0.80. Factor two with eight items has 0.735 coefficient alpha, and Cronbach's Alpha of factor three is 0.766. In short, the reliability estimation shows that both entire scale and each subscale of the instrument have a good reliability coefficient.

Conclusion and Suggestions

The exploration of construct of physical abuse or violence against children in schools and the development of instrument for measuringsuch abuse have revealed three factors behind the construct: (1) *victimized by friends, (2) victimizing friends, and (3) being victimized by teachers.* The factor loadings of the items grouped in the *victimized by friends* factor range from 0.44 to 0.69. The item loads in the *victimizing friends* factor have loadings ranging from 0.45 to 0.66. The items included in the factor of *being victimized by teachers* have factor loadings ranging from 0.57 to 0.68. All of these prove that this instrument has good construct validity.

The reliability of the instrument was estimated through Cronbach's Alpha coefficient. It is categorized as good since the reliability coefficient of the first factor is 0.830, that of the second factor is 0.735, and that of the third factor is 0.766. The Alpha coefficient of the entire instrument is 0.874. In short, the final result of this instrument development is the formation of an instrument for measuring students' experience of physical abuse in schools, which consists of three factors with 24 items, and it has good validity and reliability.

By providing this instrument for measuring physical abuse experienced by students in schools, any researchers who are interested in studying student's experience of physical abuse in schools can use this instrument. Likewise, those who want to evaluate policies concerning child-friendly schools or any related policies on the subject of preventing physical abuse in schools can make use of this instrument. Furthermore, this is also open for those who want to confirm this instrument through further analysis using the *confirmatory factor analysis* (CFA).

References

Ajema, C., Muraya, K., Karuga, R., & Kiruki, M. (2016). *Childhood experience of abuse in Kajiado County-Kenya*. Kenya: LVTC Health.

Amir, M. T. (2015). *Merancang kuesioner: Konsep dan panduan untuk penelitian sikap, kepribadian, dan perilaku*. Jakarta: Prenada Media Group.

Choo, W.-Y., Dunne, M. P., Marret, M. J.,

Fleming, M., & Wong, Y.-L. (2011). Victimization experiences of adolescents in Malaysia. *Journal of Adolescent Health*, *49*(6), 627–634. https://doi.org/10.1016/j.jadohealth.2011.04.020

Clark, R. E., Clark, J. F., & Adamec, C. A. (2007). *The encyclopedia of child abuse* (3rd ed.). New York, NY: Facts on File Library of Health and Living.

Clemens, E. V, Carey, J. C., & Harrington, K. M. (2010). The school counseling program implementation survey: Initial instrument development and exploratory factor analysis. *Professional School Counseling*, *14*(2), 125–134.

Cohen, R. J., & Swerdlik, M. (2009). *Psychological testing and assessment: An introduction to test and measurement*. New York, NY: McGraw-Hill.

Cohn, A., & Canter, A. (2003). Bullying: Facts for schools and parents. Retrieved from http://www.naspcenter.org/factsheets/bullying_fs.html

Comrey, A. L., & Lee, H. B. (2009). *A first course in factor analysis* (2nd ed.). New York, NY: Psychology Press.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., Maccallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

Hyman, I. A., & Perone, D. C. (1998). The other side of school violence: Educator policies and practices that may contribute to student misbehavior. *Journal of School Psychology*, *36*(1), 7–27.

Idris, F. (2015, December 30). 2015, tahun buram kekerasan anak. *Republika*. Retrieved from https://www.republika.co.id/berita/koran/opini-koran/15/12/30/o05vk812-2015-tahun-buram-kekerasan-anak

Kartowagiran, B. (2008). Validasi dimensionalitas perangkat tes ujian akhir nasional SMP mata pelajaran matematika 2003-2006. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(2), 177–195. https://doi.org/10.21831/pep.v12i2.1426

Kartowagiran, B., & Jaedun, A. (2016). Model asesmen autentik untuk menilai hasil belajar siswa sekolah menengah pertama (SMP): Implementasi asesmen autentik di SMP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(2), 131–141. https://doi.org/10.21831/pep.v20i2.10063

Kline, P. (2008). *An easy guide to factor analysis*. New York, NY: Routledge.

Law No. 35 of 2014 of Republic of Indonesia concerning Amendments to Law No. 3 of 2002 concerning Child Protection (2014).

McCauley, C. D., Ruderman, M. N., Ohlott, P. J., & Morrow, J. E. (1994). Assessing the developmental components of managerial jobs. *Journal of Applied Psychology*, *79*(4), 544–560. https://doi.org/10.1037/0021-9010.79.4.544

Muthmainnah, M. (2014). Membekali anak dengan keterampilan melindungi diri. *Jurnal Pendidikan Anak*, *3*(1). Retrieved from https://journal.uny.ac.id/index.php/jpa/article/view/3053

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *JAMA*, *285*(16), 2094–2100.

NSPCC. (2016). *What children are telling us about bullying. Child Bullying Report 2015/2016*.

Nurmin, N., & Kartowagiran, B. (2013). Evaluasi kemampuan guru dalam mengimplementasi pembelajaran tematik di SD kecamatan Salahutu Kabupaten Maluku Tengah. *Jurnal Prima Edukasia*,

1(2), 184–194. https://doi.org/10.21831/JPE.V1I2.2635

Parsian, N., & AM, T. D. (2009). Developing and validating a questionnaire to measure spirituality: A psychometric process. *Global Journal of Health Science*, *1*(1), 2–11. https://doi.org/10.5539/gjhs.v1n1p2

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications.

Qodar, N. (2015, March 15). Survei ICRW: 84% anak Indonesia alami kekerasan di sekolah. *Liputan6.Com*. Retrieved from https://www.liputan6.com/news/read/2191106/survei-icrw-84-anak-indonesia-alami-kekerasan-di-sekolah

Rivers, I., & Smith, P. K. (1994). Types of bullying behaviour and their correlates. *Aggressive Behavior*, *20*(5), 359–368.

Sadtyadi, H., & Kartowagiran, B. (2014). Pengembangan instrumen penilaian kinerja guru sekolah dasar berbasis tugas pokok dan fungsi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(2), 290–304. https://doi.org/10.21831/pep.v18i2.2867

Setyawan, D. (2015, June 14). KPAI: Pelaku kekerasan terhadap anak tiap tahun meningkat. *Komisi Perlindungan Anak Indonesia (KPAI)*. Retrieved from http://www.kpai.go.id/berita/kpai-pelaku-kekerasan-terhadap-anak-tiap-tahun-meningkat

Setyawan, D. (2017, November 23). Kekerasan anak di sekolah semakin memprihatinkan. *Komisi Perlindungan Anak Indonesia (KPAI)*. Retrieved from http://www.kpai.go.id/berita/kekerasan-anak-di-sekolah-semakin-memprihatinkan

Simpson, S. B. (2015). *Bullying perceptions: Understanding students with and without disabilities*. Retrieved from https://mds.marshall.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1974&context=etd

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The revised Conflict Tactics Scales (CTS2). *Journal of Family Issues*, *17*(3), 283–316. https://doi.org/10.1177/019251396017003001

Straus, M. A., Hamby, S. L., Finkelhor, D., Moore, D. W., & Runyan, D. (1998). Identification of child maltreatment with the parent-child conflict tactics scales: Development and psychometric data for a national sample of American parents. *Child Abuse & Neglect*, *22*(4), 249–270. https://doi.org/10.1016/S0145-2134(97)00174-9

UNICEF. (2014a). *Measuring violence against children: Inventory and assessment of quantitative studies*. New York, NY: Division of Data, Research and Policy.

UNICEF. (2014b). *Violence against children in East Asia and the Pacific: A regional review and synthesis of findings*. Bangkok: UNICEF EAPRO.

United Nations Secretary-General's Study. (2006). *Violence against children in schools and educational settings. Violence against children: United Nations Secretary-General's Study*. Geneva. Retrieved from https://www.unicef.org/violencestudy/4. World Report on Violence against Children.pdf

Widdiharto, R., Kartowagiran, B., & Sugiman, S. (2017). A construct of the instrument for measuring junior high school mathematics teacher's self-efficacy. *Research and Evaluation in Education*, *3*(1), 64–76. https://doi.org/10.21831/reid.v3i1.13559

Wijanto, S. H. (2008). *Structural equation modelling (SEM) dengan Lisrel 8.8*. Yogyakarta: Graha Ilmu.

REiD (Research and Evaluation in Education), 4(1), 2018, 35-44
Available online at: http://journal.uny.ac.id/index.php/reid

Developing an instrument for measuring the spiritual attitude of high school students

[*1]Safa'at Ariful Hudha; [2]Djemari Mardapi
[1,2]Graduate School of Universitas Negeri Yogyakarta
[1]Jl. Colombo No. 1, Karangmalang, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: safaat.a.huda@gmail.com

Submitted: 11 July 2018 | Revised: 31 August 2018 | Accepted: 18 September 2018

Abstract

Attitudinal competence is one the most fundamental concepts in social psychology. It is related to personal identity, moral, and ethics that gains popularity and becomes important in educational development. This research aims to develop an instrument to measure the spiritual attitude of high school students. The study was a research and development study consisting of four stages: (a) determining conceptual definition, (b) determining operational definition, (c) drawing indicators, and (d) constructing instrument. The quantitative data analysis was used to test the construct validity through Confirmatory Factor Analysis and the coefficient of construct reliability was used to estimate the instrument reliability. The results of the study show that: (1) the instrument to measure Moslems' spiritual attitude is an inventory model of summated rating scale containing 35 items; (2) the construct validity was proven by the value of the standardized loading factor and considered as significant. The instrument reliability regarded as the construct reliability coefficient is 0.890 and the average variance extracted is 0.542; (3) the construct of the instrument produces a fit statistical evidence indicated by the Goodness of Fit Index = 0.91 (≥0.90), and Root Mean Square Error of Approximation = 0.032 (≤0.08). The results indicate that the construct of the measurement is suitable with the data. In addition, this research has confirmed that the spiritual attitude of high school students is constructed by seven aspects, namely resignation *(tawakal),* sincerity *(ikhlas),* thankfulness *(syukur),* patience *(shabr),* fear *(khauf),* hopefulness *(raja'),* and righteousness *(takwa).*

Keywords: *spiritual attitude, validity, reliability*

Introduction

In the last decade, many people have been looking for the meaning and purpose of their lives as well as some spiritual experiences. It has been continuously emerged in the recent studies which have been presented by a number of researchers (Brown, 2007; Fisher, 2013). Although it has been discussed in many studies, the exact definition of spiritual experience has not been clearly explained yet. Further, the circumstance of spirituality itself can be indicated by the meaning of human life although how people intended and interpreted the meaning of life satisfaction is still being investigated (Smither & Khorsandi, 2009).

Spirituality can be interpreted as an understanding related to human identity, their ethic, and their way of life. Besides, it also explains a fundamental element that makes people full of energy and reveals the state of feeling which is integrated with overall internal human resources in the meaning beyond their religious belief (Min & Yun, 2015). In fact, spirituality dimension is almost always identified as being equal to the religious state. Furthermore, in order to support the previous statement, it is found that people with highly religious state are typically more spiritual, although it is somewhat at a lesser extent (Bryant, Choi, & Yasuno, 2003; Nikfarjam, Heidari-Soureshjani, Khoshdel, Asmand, & Ganji, 2017).

ISSN 2460-6995

The meaning of spirituality as a psychometric property has been variously defined. However, declaring the exact meaning of spirituality becomes a difficult thing (Fisher, 2016). There is no specific term which can describe how spirituality is explained. The most unclear discussion of the spiritual aspect is emphasized on the issue of the transcendental element (Koenig, 2009).

One such study implies that spirituality, as a complex construct, includes existential and also religious dimensions (Hungelmann, Kenkel-Rossi, Klassen, & Stollenwerk, 1996). It refers to the affective experiences of positive feelings from the person's ability to understand the purpose in life - related to personal, communal, and transcendental aspects (Soleimani et al., 2017). Religious dimension as the transcendental aspect in the construct of spirituality can be determined as a person's qualification and his/her ability to control his/her feelings related to how he/she interprets and makes a reflection of his/her religious belief. Furthermore, spirituality is not only evolved in terms of religious dimensions, but also becomes one of the most prominent subjects in the media and various disciplines, also in many salient factors especially in human health integrated with the internal forces (Azarsa, Davoodi, Markani, Gahramanian, & Vargaeei, 2015; Moberg, 2002).

Another outstanding theory explains spirituality as a personal belief in God or a higher power in the religious adherents (Good & Willoughby, 2006). In addition, Shodiq, Zamroni, and Kumaidi (2016) assert that as a transcendental element, spirituality in Islamic studies and in terms of Islamic faith has two dimensions, namely: belief *(tashdiq-al-qalb)* which is known as *rukun iman,* and also attitude or personal feeling *(amal-al-qalb)* which has seven aspects i.e. thankfulness *(syukur)*, fear *(khauf)*, love *(mahabbah)*, patience, resignation (*tawakkal)*, hopefulness *(raja)*, and sincerity *(ikhlas)*. In the same term of Islamic studies, spirituality based on a Moslem perspective centers on loving submission and closeness to God (Ghorbani, Watson, Geranmayepour, & Chen, 2014).

Spirituality and religiosity are often used interchangeably, but the two concepts are very different. Sheridan and Hemert (1999) define spirituality as a human search for the purpose and meaning of life experience, while Tanyi (2002) argues that spirituality is a personal search for the purpose and meaning in life. Spirituality entails connection to religious beliefs or self-chosen faith. The two previous definitions are almost the same thing, but there is a slight difference. Spirituality according to the first description is emphasized on the meaning of life experience, while the second is focused on the meaning in life.

According to Hill et al. (2000), the term 'spirituality' can be used to describe 'one's religious experiences,' while the term 'religiosity' is used to express 'the state of belief.' Spirituality in the general view seems more basic, positive, and sincere while religiosity implies the ritual and obedience in worship related to certain religious adherents.

One of the most important and fundamental concepts in social psychology is attitudinal competence (Bidjari, 2011). Fishbein and Ajzen (1975) define attitude as a person's location on a bipolar evaluation of affective dimension concerning some objects, viewed as predisposing the individual to do various overt behaviours. Likewise, attitude refers to the people's predisposition to respond consistently whether they like the object or not (Mardapi, 2017, p. 134). The term 'bipolar evaluation of affective dimension' can be described as the state of positive and negative feeling onto the particular object. The attitude in this way consists of the positive and negative direction.

According to Kusaeri and Suprananto (2012, p. 206), like the previous explanation, attitudinal competence is defined as a state of readiness to react to an object in a certain way as a form of evaluation and reflection of feeling. Furthermore, Sax (1980, p. 493) emphasizes the characteristics of attitude which contains some dimensions, i.e. direction, intensity, pervasiveness, consistency, and salience.

In relation to the spiritual term, attitude can be explained as a person's predisposition to choose his/her response to the prevalent situation with an internalization of specific dimension correlating with his/her religious understanding and spiritual conception. Spiritual

attitude is more often identified as the same as religious attitude. Hill et al. (2000) affirm that both spirituality and religiosity have been recognized as having a relationship with a person's mental health status and are relevant to the study of personality and in the genetic determinants of personality. Further, Huber and Huber (2012) state that the dimensions of spiritual attitude can be seen from the ideology, private practice, religious experience, and intellectual dimensions that are considered representing the totally religious life.

Although it is hardly practical to discuss the spirituality definition and its relation, which is a multidimensional concept (Cook, 2004; Hill et al., 2000) including such domains as personal, communal, environmental, and transcendental (Fisher, 2016), the measure of spirituality is more popular in the field of mental health, human existence, and social well-being research. However, this major property of psychometric related to the existential and religious dimensions is infrequently and less practiced in the scope of education, especially in student achievement and academic behaviour.

The spiritual attitude in terms of educational learning and curriculum is a student's qualification of ability to control him/herself and his/her description of spiritual self-coping. It is associated with the character building in education which is intended to build a moral, democratic, and religious student as the best outcome in educational learning. The spiritual attitude illustrates the increase of vertical interaction and the strong relationship with God (Ministry of Religious Affairs of Republic of Indonesia, 2014, p. 8). The spirituality and spiritual attitude are gaining popularity within educational curriculum and academics as the discussions regarding the prominence of spiritual attitude in education increase.

Based on the perspective of a Moslem, spiritual attitude is related to the faith, centered on loving submission and closeness to God which can be seen from his/her religious experience, private practice, and social relationship. This study is intended to develop an instrument to measure Moslems' spiritual attitude in education with the seven subscales drawn from the Islamic religious term named resignation (*tawakkal),* sincerity *(ikhlas),* thankfulness *(syukur)*, patience *(shabr)*, fear *(khauf)*, hopefulness *(raja')*, and righteousness *(takwa)*. This study is also intended to test the instrument construct validity and estimate the instrument reliability through quantitative analysis of the data obtained from the research sample.

Method

This study is a research and development (R & D) study employing the quantitative approach. It is aimed at developing an instrument to measure Moslems' spiritual attitude in education for high school students. The research procedure was carried out through four stages, namely: (a) determining conceptual definition, (b) determining operational definition, (c) drawing indicators, and (d) constructing instrument.

Population and Sample

This research was conducted at 11 public senior high schools in Yogyakarta, Indonesia. The population was the grade XI Moslem students, and the sample was 307 participants established by using the cluster random sampling technique by considering students' focus of study, MIA (Mathematics and natural science) and IS (Social science) as the cluster. The number of the sample respondents is shown in Table 1.

Table 1. The numbers of sample respondents

School Name	Amount
SMA Negeri 2 Yogyakarta	78
SMA Negeri 4 Yogyakarta	89
SMA Negeri 7 Yogyakarta	79
SMA Negeri 10 Yogyakarta	61
Total	**307**

Data Collecting Technique

The instrument to measure Moslems' spiritual attitude was developed by using the seven subscales drawn from the Islamic religious terms, and contained 24 indicators. Those indicators were developed into 35 items of questionnaire using three-point alternative response model (a, b, and c) of the

summated rating scale and designed to the multiple-choice form of questionnaires with the variant score of key answer (1 - 3). The conceptual framework of Moslems' spiritual attitude in this research is shown in Figure 1.

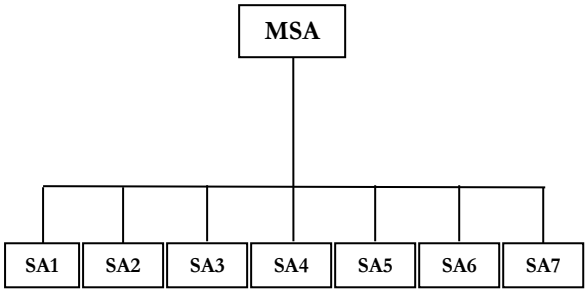

Figure 1. The conceptual framework of Moslems' spiritual attitude

Notes:
MSA : Moslems' Spiritual Attitude
SA1 : Resignation *(Tawakal)*
SA2 : Sincerity *(Ikhlas)*
SA3 : Thankfulness *(Syukur)*
SA4 : Patience *(Shabr)*
SA5 : Fear *(Khauf)*
SA6 : Hopefulness *(Raja')*
SA7 : Righteousness *(Takwa)*

The Moslems' spiritual attitude (MSA) has seven subscales. The first subscale is resignation *(Tawakal),* which refers to the state of self-resignation to obey in worship and to accept all Allah's decision. The second subscale is sincerity *(Ikhlas)*, the term which refers to being sincere to do a favor. The third subscale is thankfulness *(Syukur),* referring to admitting all Allah's best creatures and feeling happy to do His order and leaving His prohibition. The fourth subsclae is patience *(Shabr),* referring to the attitude of being consistent to refrain himself from ugliness. The fifth subscale is fear *(Khauf)*, being afraid of Allah. The sixth subscale is hope *(Raja'),* hoping and asking for His grace and forgiveness. The last subscale is righteousness *(Takwa),* which is the Islamic concept of having self-restraint.

Content Validity

The developed items in this research instrument were validated by the five panels of judges and regarded as the expert-judgement. The Aiken's V formula was used to assess the feasibility of the content validity. The lecturers of Educational Measurement and Islamic studies were involved in the panel. All of the experts were selected based on their experiences in the field of educational measurement, psychometrics, and Islamic studies.

The validators as the experts assess the whole instrument by giving scores to the developed items and give responses to the instrument's indicator through comments and suggestions. Subsequently, the validators' suggestions and comments become the basis for making a relevant improvement which will be used to rewrite the items of the research instrument.

Construct Validity

Construct validity needs a definition with the specified conceptual circumscription and more focused on particular attributes of the variable than concerned with the values or scores gained from the instrument (Salkind, 2000). Construct validity emphasizes on logical analysis and investigates the relationships of the data analysis based on theoretical consideration.

Construct validity explains the extent to which performance on the test is consistent with the constructs in a particular theoretical consideration. The present study is also concerned with investigating the construct validity for the research instrument to test how the instrument is consistent with the spiritual attitudes construct.

The result of the confirmatory factor analysis produced a standardized loading factor (SLF) and was determined as the construct validity. Once the SLF value of the certain indicator is over 0.30, the indicator is considered as significant (Igbaria, Zinatelli, Cragg, & Cavaye, 1997, p. 290). Another evidence of the construct validity is also determined by the significant *t-value* (*t-value*>1.96) which uses the confidence interval of 0.05.

Goodness of Fit Statistics

The fit statistics of the instrument in this study refers to the fulfilment of two of the three models of fit criteria, i.e. *Root Mean Square Error of Approximation* (RMSEA ≥*0.08*), *p-value* ≥*0.05* and *Goodness of Fit Index* (GFI

≥0.90) (Suranto, Muhyadi, & Mardapi, 2014, p. 102). Hair, Black, Babin, and Anderson (2010, p. 656) explain that RMSEA is the fittest statistics to be used in the confirmatory factor analysis. The Goodness of Fit statistics was used in this research to investigate the fit statistics between the primary data obtained from the research sample and the theoretical consideration. The fulfillment of the two models of fit criteria described that the construct of measurement was suitable to the data.

Table 2. Parameter of fit statistics

Goodness of Fit	Cut off Point	Notes
Chi-Square (p-value)	*p-value ≥0.05*	Model Fit
RMSEA	*RMSEA ≥0.08*	Model Fit
Goodness of Fit Index (GFI)	*GFI ≥0.90*	Model Fit

Data Analysis

The score given from the five experts' judgement for the total items in the research instrument was subsequently analyzed with the Aiken's V formula to investigate the content validity of the instrument. The content validity analysis was used prior to the dissemination of the research instrument. The primary data obtained from the research instrument were analyzed using Lisrel 8.80 software program.

To analyze the quantitative data, two statistical procedures were employed to answer the research question. *First*, the second-order Confirmatory Factor Analysis was applied to obtain the construct validity for the instrument based on the standardized loading factor and to investigate the fit statistics of the instrument construct. *Second*, the coefficient omega or construct reliability and Average Variance Extracted formula was applied to estimate the reliability coefficient of the instrument.

The fit statistics of the instrument was obtained from the output of the second-order Confirmatory Factor Analysis. RMSEA and GFI were used to determine the instrument fit statistics.

Findings and Discussion

This study is aimed to develop an instrument to measure Moslems' spiritual attitude as an inventory model. To achieve these goals, a number of respondents were involved as the research sample to obtain the quantitative data based on their responses to the questionnaires. The score gained by using the instrument was used to test the construct validity and the coefficient of instrument reliability through the data analysis.

The construct dimension of Moslems' spiritual attitude in this study includes seven aspects developed into 24 indicators. The seven aspects include resignation *(tawakkal)*, sincerity *(ikhlas)*, thankfulness *(syukur)*, patience *(shabr)*, fear *(khauf)*, hopefulness *(raja')*, and righteousness *(takwa)*. The establishment of the Moslems' spiritual attitude construction was based on the experts in Islamic studies, psychometry, and educational evaluation, as well as the general practitioners of Islamic education in several high schools.

The 35 items of the questionnaire were validated using Aiken's V formula to assess the feasibility of the content validity. The Aikens' V index ranged from 0.80 to 0.95 which can be interpreted that all the items which were developed from certain indicators in this research instrument have a good content validity. The validator's response reveals that the developed instrument in this research is a suitable instrument to measure Moslems' spiritual attitude in education.

Confirmatory Factor Analysis

The conceptual construct and the analysis result of the developed instrument with second-order CFA are presented in Figure 2. The analysis result of the second order CFA as indicated in Figure 2 shows that the model designed in this study complies with the goodness of fit statistics. The model fit of the instrument is indicated by the RMSEA = 0.032 and Goodness of Fit Index = 0.91.

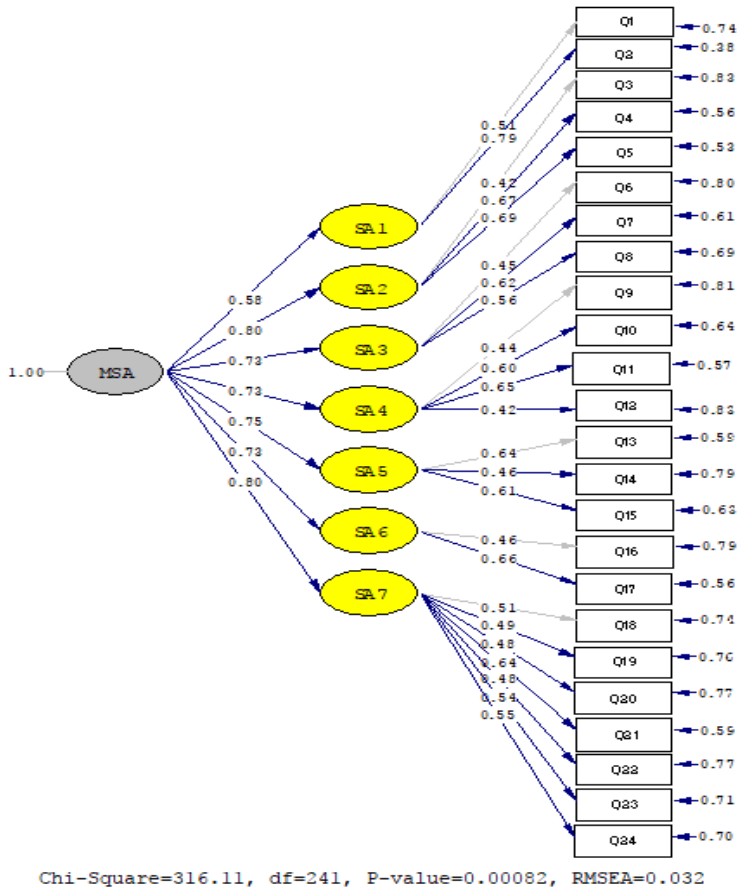

Figure 2. The result of CFA second order of Moslem's spiritual attitude

Notes:
Q1 : Self-resignation to obey in worship
Q2 : Recognizing human's limitation
Q3 : Not to expect any rewards
Q4 : Not to be careless in praise
Q5 : Not to be hopeless at failure
Q6 : Admitting all Allah's best creatures
Q7 : Using God's grace for good
Q8 : Not using the grace for ugliness
Q9 : Being consistent with Allah's commandment
Q10 : Being consistent with Allah's prohibition
Q11 : Being consistent to tell the good
Q12 : Being grateful for the tragedy and hardship
Q13 : Feeling guilty for disregarding Allah's commandment
Q14 : Feeling guilty for breaking Allah's prohibition
Q15 : Being afraid of His threat
Q16 : Hoping for Allah's grace
Q17 : Asking for His forgiveness
Q18 : Making *shalat* a priority in life
Q19 : Paying for *zakat*
Q20 : Being tolerant
Q21 : Being honest
Q22 : Rejecting adultery
Q23 : Not to use other's property
Q24 : Not breaking promises

The value of Standardized Loading Factor as the result of the second-order CFA is presented in Figure 2, while the *t-value* and R^2 of the instrument indicators are shown in Table 3.

The result of the second-order confirmatory factor analysis indicates that the 24 indicators in the conceptual construct of the Moslems' spiritual attitude are considered significant based on the t-value index. It is shown by the t-value >1.96 in which the lowest t-value is 3.83 (Q5) and the highest is 5.78 (Q21). Another evidence is shown by the value of standardized loading factor based on the result of the second-order CFA which is shown in Table 3. It indicates that the entire indicators have the value of SLF >0.30 (the lowest value of SLF is 0.42 while the highest is 0.79). The evidence of the t-value index and the standardized loading factor in this instrument research can be identified as an acceptable construct validity and suitable instrument to measure Moslems' spiritual attitude in education.

Table 3. The result of second order CFA of Moslems' spiritual attitude

Indicator	Loading Factor	t-value	R^2	Notes
Q1	0.51	-	0.26	*Reference Var*
Q2	0.79	4.70	0.62	*Indicator Fit*
Q3	0.42	-	0.17	*Reference Var*
Q4	0.67	4.59	0.44	*Indicator Fit*
Q5	0.69	3.83	0.47	*Indicator Fit*
Q6	0.45	-	0.20	*Reference Var*
Q7	0.62	5.13	0.39	*Indicator Fit*
Q8	0.56	4.90	0.31	*Indicator Fit*
Q9	0.44	-	0.19	*Reference Var*
Q10	0.60	4.16	0.36	*Indicator Fit*
Q11	0.65	4.86	0.43	*Indicator Fit*
Q12	0.42	4.34	0.17	*Indicator Fit*
Q13	0.64	-	0.41	*Reference Var*
Q14	0.46	4.19	0.21	*Indicator Fit*
Q15	0.61	4.77	0.37	*Indicator Fit*
Q16	0.46	-	0.21	*Reference Var*
Q17	0.66	4.04	0.44	*Indicator Fit*
Q18	0.51	-	0.26	*Reference Var*
Q19	0.49	5.28	0.24	*Indicator Fit*
Q20	0.48	4.76	0.23	*Indicator Fit*
Q21	0.64	5.78	0.41	*Indicator Fit*
Q22	0.48	4.82	0.23	*Indicator Fit*
Q23	0.54	5.07	0.29	*Indicator Fit*
Q24	0.55	5.14	0.30	*Indicator Fit*

Instrument Reliability

Reliability is an essential characteristic of a goodness between the test and the obtained scores. Reliability is required to obtain the instrument validity. The investigation of both validity evidence and reliability coefficient can be defined as the complementary aspects of identifying, estimating and interpreting different sources of variance in the scores (Bachman, Davidson, Ryan, & Choi, 1995).

The reliability coefficient of the instrument was employed to test the consistency of the measurement and was used as an estimation of how much the instrument would give the same result under the same conditions. The estimation of reliability in this research was evaluated with the construct reliability (CR) and the average variance extracted (AVE). The index values of the construct reliability coefficients are presented in Table 4, and the average variance extracted for the instrument is shown in Table 5.

Table 4. The coefficient of construct reliability of the instrument

Aspects	*SLF*	*Errorvar*
SA1	0.58	0.66
SA2	0.80	0.35
SA3	0.73	0.47
SA4	0.73	0.47
SA5	0.75	0.44
SA6	0.73	0.47
SA7	0.80	0.36
$(\Sigma SLF)^2$	26.21	
$\Sigma Errorvar$		*3.19*
CR	**0.890**	

The coefficient of construct reliability shown in Table 4 for the instrument is 0.890. The CR formula was used to perform the internal consistency of the instrument and to test the indicator in measuring the construct of the instrument. The result of the CR computation shows that the instrument has a high reliability index and is considered to have a good consistency to measure Moslems' spiritual attitude in education.

Table 5. The coefficient of average variance extracted (AVE) of the instrument

Aspects	SLF^2	*Errorvar*
SA1	0.34	0.66
SA2	0.64	0.35
SA3	0.53	0.47
SA4	0.53	0.47
SA5	0.56	0.44
SA6	0.53	0.47
SA7	0.64	0.36
$\Sigma(SLF^2)$	3.78	
$\Sigma Errorvar$		*3.19*
AVE	**0.542**	

The average variance extracted (AVE) is used to measure the number of variances that can be captured by certain constructs compared to the variances produced by the error of measurement. Table 5 shows that the developed instrument has a moderately good average variance extracted estimation and been proven by the computation of 0.542 (slightly above 0.50) for the entire subscales in the Moslems' spiritual attitude instrument.

Conclusion

The construct of Moslems' spiritual attitude is determined by Islamic religious terms. The instrument of the study is an inventory which is defined as a self-report model of the

summated rating scale and designed for the multiple-choice form of the questionnaire using three-point (1–3) alternative responses. The instrument contains 35 item questionnaire called Moslems' spiritual attitude scale.

The Moslems' spiritual attitude dimension in education means students' attitude or personal feeling in self-condition related to their religious experience, ideology, and private practice in terms of relation with God and interaction with Him. The Moslem's spiritual attitude consists of seven aspects as the latent variable and was developed into 24 indicators as the variable observed. .

The construct validity of the Moslems' spiritual attitude scale is considered as moderately high according to the standardized loading factor (SLF) value. The value of SLF as the result of the second-order confirmatory factor analysis for the 24 instrument indicators is above 0.30, ranging from 0.42 to 0.79. The computation result for the coefficient of construct reliability (CR) of the instrument is 0.890 while the average variance extracted (AVE) is 0.542.

The fit statistics produces a model fit as indicated by the Root Mean Square Error of Approximation (RMSEA) = 0.032 (<0.08), and Goodness of Fit Index (GFI) = 0.91. The result indicates that the construct of the measurement is suitable to the data. The model is also suitable for estimating the covariance matrix of the population which means that there is no difference from the sample respondents in this study. According to the reseach findings, it can be concluded that Moeslem's spiritual attitude is constructed by seven aspects, namely resignation *(tawakkal)*, sincerity *(ikhlas)*, thankfulness *(syukur)*, patience *(shabr)*, fear *(khauf)*, hopefulness *(raja')*, and righteousness *(takwa)*.

References

Azarsa, T., Davoodi, A., Markani, A. K., Gahramanian, A., & Vargaeei, A. (2015). Spiritual wellbeing, attitude toward spiritual care and its relationship with spiritual care competence among critical care nurses. *Journal of Caring Sciences*, *4*(4), 309–320. https://doi.org/10.15171/jcs.2015.031

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge TOEFL comparability study*. Cambridge: Cambridge University Press.

Bidjari, A. F. (2011). Attitude and social representation. *Procedia - Social and Behavioral Sciences*, *30*, 1593–1597. https://doi.org/10.1016/j.sbspro.2011.10.309

Brown, C. G. (2007). Secularization, the growth of militancy and the spiritual revolution: Religious change and gender power in Britain, 1901-2001. *Historical Research*, *80*, 393–418. https://doi.org/10.1111/j.1468-2281.2007.00417.x

Bryant, A. N., Choi, J. Y., & Yasuno, M. (2003). Understanding the religious and spiritual dimensions of students' lives in the first year of college. *Journal of College Student Development*, *44*(6), 723–745. https://doi.org/10.1353/csd.2003.0063

Cook, C. C. (2004). Addiction and spirituality. *Addiction*, *99*(5), 539–551. https://doi.org/10.1111/j.1360-0443.2004.00715.x

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Fisher, J. (2013). Assessing spiritual well-being: Relating with God explains greatest variance in spiritual well-being among Australian youth. *International Journal of Children's Spirituality*, *18*(4), 306–317. https://doi.org/10.1080/1364436X.2013.844106

Fisher, J. (2016). Selecting the best version of SHALOM to assess spiritual well-being. *Religions*, *7*(5), 45. https://doi.org/10.3390/rel7050045

Ghorbani, N., Watson, P. J., Geranmayepour, S., & Chen, Z. (2014). Measuring Muslim spirituality: Relationships of Muslim experiential religiousness with religious and psychological adjustment in Iran. *Journal of Muslim Mental Health*,

8(1). https://doi.org/http://dx.doi.org/10.3998/jmmh.10381607.0008.105

Good, M., & Willoughby, T. (2006). The role of spirituality versus religiosity in adolescent psychosocial adjustment. *Journal of Youth and Adolescence*, *35*(1), 39–53. https://doi.org/10.1007/s10964-005-9018-1

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Hill, P. C., Pargament, K. I., Hood, R. W., McCullough, J. M. E., Swyers, J. P., Larson, D. B., & Zinnbauer, B. J. (2000). Conceptualizing religion and spirituality: Points of commonality, points of departure. *Journal for the Theory of Social Behaviour*, *30*(1), 51–77. https://doi.org/10.1111/1468-5914.00119

Huber, S., & Huber, O. W. (2012). The centrality of religiosity scale (CRS). *Religions*, *3*(3), 710–724. https://doi.org/10.3390/rel3030710

Hungelmann, J., Kenkel-Rossi, E., Klassen, L., & Stollenwerk, R. (1996). Focus on spiritual well-being: Harmonious interconnectedness of mind-body-spirit—use of the JAREL spiritual well-being scale: Assessment of spiritual well-being is essential to the health of individuals. *Geriatric Nursing*, *17*(6), 262–266. https://doi.org/10.1016/S0197-4572(96)80238-2

Igbaria, M., Zinatelli, N., Cragg, P., & Cavaye, A. L. M. (1997). Personal computing acceptance factors in small firms: A structural equation model. *MIS Quarterly*, *21*(3), 279–305. https://doi.org/10.2307/249498

Koenig, H. G. (2009). Research on religion, spirituality, and mental health: A review. *The Canadian Journal of Psychiatry*, *54*(5), 283–291. https://doi.org/10.1177/070674370905400502

Kusaeri, & Suprananto. (2012). *Pengukuran dan penelitian pendidikan*. Yogyakarta: Graha Ilmu.

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

Min, S., & Yun, S. (2015). A study on the differences between spiritual wellbeing and sexual attitude considering the type of university. *Indian Journal of Science and Technology*, *8*(S1), 54–58. https://doi.org/10.17485/ijst/2015/v8iS1/57582

Ministry of Religious Affairs of Republic of Indonesia. (2014). *Model penilaian pencapaian kompetensi peserta didik madrasah tsanawiyah (MTs)*. Jakarta: Directorate General of Islamic Education.

Moberg, D. O. (2002). Assessing and measuring spirituality: Confronting dilemmas of universal and particular evaluative criteria. *Journal of Adult Development*, *9*(1), 47–60. https://doi.org/10.1023/A:1013877201375

Nikfarjam, M., Heidari-Soureshjani, S., Khoshdel, A., Asmand, P., & Ganji, F. (2017). Comparison of spiritual well-being and social health among the students attending group and individual religious rites. *World Family Medicine Journal*, *15*(8), 160–165. https://doi.org/10.5742/MEWFM.2017.93071

Salkind, N. J. (2000). *Exploring research*. Michigan, MI: Prentice Hall.

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation*. California, CA: Wadsworth.

Sheridan, M. J., & Hemert, K. A. (1999). The role of religion and spirituality in social work education and practice: A survey of student views and experiences. *Journal of Social Work Education*, *35*(1), 125–141. https://doi.org/10.1080/10437797.1999.10778952

Shodiq, S., Zamroni, Z., & Kumaidi, K. (2016). Developing an instrument for measuring the faith of the students of Islamic senior high school. *REiD (Research and Evaluation in Education)*, *2*(2), 181–193. https://doi.org/10.21831/reid.v2i2.11117

Smither, R., & Khorsandi, A. (2009). The implicit personality theory of Islam. *Psychology of Religion and Spirituality*, *1*(2), 81–96. https://doi.org/10.1037/a0015737

Soleimani, M. A., Sharif, S. P., Allen, K. A., Yaghoobzadeh, A., Nia, H. S., & Gorgulu, O. (2017). Psychometric properties of the Persian version of spiritual well-being scale in patients with acute myocardial infarction. *Journal of Religion and Health*, *56*(6), 1981–1997. https://doi.org/10.1007/s10943-016-0305-9

Suranto, S., Muhyadi, M., & Mardapi, D. (2014). Pengembangan instrumen evaluasi uji kompetensi keahlian (UKK) administrasi perkantoran di SMK. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 98–114. https://doi.org/10.21831/pep.v18i1.2127

Tanyi, R. A. (2002). Towards clarification of the meaning of spirituality. *Journal of Advanced Nursing*, *39*(5), 500–509. https://doi.org/10.1046/j.1365-2648.2002.02315.x

REiD (Research and Evaluation in Education), 4(1), 2018, 45-57
Available online at: http://journal.uny.ac.id/index.php/reid

Exploring the accuracy of school-based English test items for grade XI students of senior high schools

[*1]Martin Iryayo; [2]Agus Widyantoro
[1]University of Rwanda - College of Education
KG 11 Ave, Kigali, Rwanda
[2]Department of English Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: martiniryayo@gmail.com

Submitted: 08 June 2018 | Revised: 27 July 2018 | Accepted: 01 August 2018

Abstract

This study is set out to (1) explore the accuracy of school-based English test items developed by English teachers and (2) compare the relationship between the content covered by teacher and the students' success level. This research used the quantitative approach. The source of the data is all grade XI students' answers to the English test for the second semester of 2016/2017 academic year, and their English teachers' responses to the questionnaire. During this cross-sectional survey, 241 grade XI students and six English teachers were selected by using the total population sampling technique. To analyze the data, the IRT model was prioritized with BILOG MG 3.0, WINISTEPS 3.7. The findings of the study indicate that (1) the test is valid, (2) it is reliable, (3) majority of the items are moderately difficult, (4) more than a half of all items have power to discriminate the examinees, (5) some items show fully-effective distractors, and (6) the test gives much information at -.40 of theta which means that the test is difficult for the grade XI students. Moreover, there is a wide gap between the content covered and the level of success.

Keywords: *CTT, discrimination power, distractor, information function, IRT, theta, total population sampling*

Introduction

A well-constructed test is the best way to evaluate a student's mastery in a particular field. Gronlund (1993, pp. 205–206) stresses that tests do not only help teachers to make some instructional decisions with their direct influence on students' learning, but they also assist in a number of other ways. For instance, tests can increase students' motivation. The purpose of tests is to obtain an accurate and fair assessment of students' abilities. Nevertheless, it is impossible for a test to evaluate skills or knowledge bases if it is influenced by irrelevant factors that could undermine the results. These factors that potentially create bias can comprise of gender, ethnic, and cultural differences. In case there is no proper accounting for these biasing factors, the outcome of the test will unfairly represent the abilities of the examinees (Gronlund, 1993, p. 207). Alternatively, the results of a test are essentially meaningless if they are unfair for test takers due to the culture, gender, or ethnic origin biases.

A veracious picture of skills and knowledge that students have in either the subject area or domain tested should be presented by test results. The successful instructional, curriculum planning, and evaluation of linked programs cannot be accomplished without students' quality achievement data. Test scores that overestimate or underestimate students' actual knowledge and skills cannot serve these important purposes (Young, Cummings, & St-Onge, 2017). The accuracy of the achievement data cannot be procured since the composers of the test do not pay attention to the accuracy of the test components, because once the test is not well prepared, it obviously

ISSN 2460-6995

affects the students' achievements even if they understand the material well. Thus, the test must be as accurate as possible.

In standardized testing, there are several means for measuring students' cognitive abilities. Currently, multiple-choice tests are commonly used for measuring students' cognitive abilities (Galsworthy et al., 2005). Standardized scores is used by most schools to evaluate the educational quality and student performance (Brescia & Fortune, 1989, pp. 1–5). As long as it is believed that test scores are considered as an important factor to assess students' performance, teachers should develop the tests which are as fair as possible for examinees regardless of their races, genders, or any disability they may have (Joint Committee on Testing Practices of American Psychological Association, 2004). Reviewing all items of a test is the most fruitful way to ensure that they are free from all irrelevant sources of variances because item bias dirtily affects the examinees' scores. There must be empirical revision of the items before administering them to the examinees in order to ensure the quality of their characteristics.

In achievement testing, it is possible to use different formats. Multiple-choice (MC) items are broadly used for classroom assessment and they always account for a significant constituent of a student's grade in a course (DiBattista & Kurzawa, 2011). A normal MC item is made up of a question, known as stem, and a list of alternatives from which one becomes the right answer to the question. The test takers pick only the option they think fits to the question asked. The keyed option is the best name for the correct answer while the remaining alternatives refer to as distractors. For instructors, there is a variety of advantages to use the MC test format; scoring MC items takes a short time particularly when the examinees indicate their responses on a well-scanned optical MC answer sheet (universally used form). For teachers of subjects with large enrolment, easy grading can make MC tests very specifically appealing to them. Obviously, multiple choices tests are more advantageous even though there are some flaws still pending while measuring the students' performance.

Content validity, clarity, and reliability are the most crucial traits of achievement tests. The content validity of a test is always seen by how accurately the test samples the range of knowledge, skills, and abilities expected from the testees during an examination period. The reliability of a test depends on its grading stability and its power to discriminate students upon the basis of their different levels of performance (Kartowagiran, 2012). Well-developed multiple-choice test items are in general more valid, clearer, and more reliable than essay tests because they broadly represent content in the syllabus, able to distinguish all levels of performance, and scoring consistency is virtually guaranteed. Thus, validation is a starting point for dealing with multiple choice test item quality.

Content validity can be obtained in various ways. The content validity (relevance) by experts' judgement can be computed in different ways. The use of pre-established acceptability criterion, calculation of rating average upon each item relevance, quantification of item relevance (with three or more experts) by using coefficient alpha, and kappa coefficient computation are the most known techniques (Polit & Beck, 2006). With this approach, there should be a team of experts to judge whether an item on a scale is relevant to (or congruent with) the construct being measured. Each rater is free to compute the percentage of item relevance, then the average is taken across all raters (experts).

Another way of evaluating the accuracy of MC test items is concerned with studying the answers that the examinees make, in which within this research, this analysis approach was used. Precisely, teacher-developed test items administered to the examinees are basically analyzed on the basis of difficulty level, discrimination level, and effectiveness of the distractors (DiBattista & Kurzawa, 2011). In brief, before putting the items in their bank, the main characteristics stated above should be considered because any item which is either too difficult or easy, item that does not discriminate students, and item with ineffective distractors, does not qualify to be stored in the item bank.

Test takers should be differentiated by their abilities. The discrimination capacity of an MC test item is the most prevailing property because it reflects the extent to what more intelligent students are more likely than less knowledgeable students to select the keyed option (Abadyo & Bastari, 2015). MC test item discriminatory capacity can be measured with the computation of its index, which reflects the correlation between the examinees' total scores and the score received on the item to be considered (i.e. 1 stands for the keyed option, while 0 for the wrong answer). Even more, there are items which are problematic because they produce negative discriminatory indexes, maybe due to the unclear wording or the existence of two correct alternatives rather than one (DiBattista & Kurzawa, 2011). With the presence of such items, there is a detraction from the overall accuracy of the test as a whole, because the number of less knowledgeable examinees who select the keyed option outweighs that of the knowledgeable examinees.

With regard to the perspective of its functionality, there are two requirements for a distractor to be functioning: first, at least some examinees must select it, if they do not, the distractor is not plausible to them until they can be lured away from the correct answer, so such a distractor never contributes to the discrimination of the test takers. Abdulghani, Ahmad, Ponnamperuma, Khalil, and Aldrees (2014) have suggested that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for the effectiveness of the distractors. The second requirement refers to the power of a distractor to distinguish high achievers from low achievers (stronger from weaker students), considering that the power of discrimination is clear when the correct answer is more often chosen by the students with high scores than their counterparts.

Related to the statements, opinions, and views of different authors as fully explained in the previous section, the problems that always appear when developing school-based English test items with the format of multiple choices, are so many, such as the content of some multiple-choice tests, which does not cover the material taught in the classroom, and the main parts of multiple choice tests items; stem, key, and alternatives which are not built according to the criteria or guidelines; some teachers do not have enough skills to get by this problem; by analyzing the scores of the students obtained from multiple choice tests during a couple of academic years ago, there is inconsistency because there is a lack of item homogeneity; some individual items are not highly correlated to each other and even to the whole test; some English teachers are not cautious of the difficulty level of the items. At the end of a teaching session, they develop tests which are either too easy or difficult. The ideal index of difficulty should fall between -2 and 2 (Hambleton, Swaminathan, & Rogers, 1991, p. 13). It is quite problematic to have items with difficulty index of far less than -2 or more than 2, and some English teachers do not know how to develop multiple choice items which can discriminate the participants. Moreover, the distractors are powerless to attract the examinees because some are chosen by <5% of examinees (Mkrtchyan, 2011). It is a problem to have items which cannot discriminate the achievers (a_i less than 2).

Like other scientific studies, this study aims at exploring the accuracy of school-based English test items developed by English teachers through (1) validity index, (2) reliability coefficients, (3) difficulty level, (4) discrimination power, (5) distractor effectiveness, and (6) level of information given by the items and the whole test in general and at comparing between the content covered by teacher and student success level. The current study is expected to be beneficial. Practically and even theoretically, the results of this study should be used by English test administrators, moderators, and even supervisors in order to make adequate policies on how to fairly and professionally prepare a suitable English test. This is very important because some teachers and other school academicians who develop test items for testing students do not have enough skills yet to examine the primordial characteristics indicating a good item.

Many researchers worked on the accuracy or quality of achievement test items.

Charismana and Aman (2016) conducted a research about the quality of civic education final examination items, in the whole regency of Kudus, Indonesia. The students involved in the study were grade VIII students of junior high schools that apply Curriculum 2013. The data were analyzed both qualitatively and quantitatively. The qualitative results show that 31 items are good whereas are items are not. The quantitative results show that 24 items or 68.57% of all items are good, while 11 items or 31.42% of all items are not. As a result, approximately 15 items are recommended to be revised.

A study conducted by Osadebe (2015) with 100 items administered to 1000 students comes up with the results that the achievement test for the subject of Economics has a high face and content validities. The test item quality was evaluated through difficulty and discrimination indexes. A difficult index or *p-value* of 0.5 was referred to after the use of the formula for guessing correction. The index of discrimination was computed with point biserial statistics whereby the minimum boundary is .30. With the KR-20, the test was very highly reliable with the coefficient of .95. These findings support the use of this instrument to internally evaluate the students in order to be ready for the external testing (examination).

According to the study by Boopathiraj and Chellamani (2013), which was aimed at analyzing test items in the subject of Research with students enrolled in Master of Education (M.Ed) program, they wanted to ensure the difficulty and discrimination levels of MC test items. A sample of 200 students from different colleges of education was established. The sample consisted of both genders. The findings indicate that a big number of items are not accepted, and there is a good discrimination index for some items, but some of them are rejected due to poor discrimination indexes. Based on the statement above, most of the items have the difficulty level (*bi*) from -2 to 2 and discrimination index of (a_i) > 2.

Sabri (2013) worked on a comprehensive test at a university in Perak, involving 16 music students. With MS Excel, he computed the difficulty level of 41 items. The reliability coefficients and discriminatory indexes were computed using MS Excel and SPSS 17.7 respectively. The outcome of the research came up with the information that 44% of all items have the difficulty index of > .80, then 59% of the items have acceptable discriminatory power. There is no effective distractor. With KR-20, the coefficient of reliability is .717 while with KR-21 is .703. Hence, it is reasonable to conclude that the items are reliable, moderately easy, 80% discriminate high from low achievers, but some distractors were chosen by less than 5% of examinees (implausible).

Quaigrain and Arhin (2017) carried out a study about MC test items. The sample was made up of 247 students doing year-1 diploma in education at Cape Coast Polytechnics. A test of 50 MC items was given to them in the subject of educational measurement. The results of the study show that the whole test has an internal consistency reliability of .77 (KR-20), the mean score of 29.23, the standard deviation mean score of 6.36, difficulty level (*p-value*) and discrimination index (*DI*) of 58.46% (SD=21.23) and .22 (SD=.17), respectively, and the mean score of DE of 55.04 (SD=24.09). As to DI, 30 items (60%) are reasonably accepted. Every item with moderate difficulty level, high discriminatory power, and functioning distractors should still be part of the next testing to improve classroom assessment quality.

There is no study without innovation. The novelty of this study can be seen from data analysis section. Apart from the variables that look similar to the previous studies by other academic researchers, the current research involves a new way of giving grades to teachers on the basis of content covered after the learning term. As the majority of the previous studies used classical test theory to analyze item accuracy, the researchers in this study used the item response theory (IRT) to have clearer and more information on the item quality, so that the newly published IRT software was used.

This study is expected to come up with the answers to the questions in relation to the quality or accuracy of school-based English test items: (1) To what extent do English test

items represent the content or subject topics they intend to measure for grade-11?; (2) What proves that English test assesses the underlying theoretical construct it is purported to measure?; (3) How convergent are the items, making up English test, to be considered homogeneous? Do they complement each other?; (4) How reliable and informative are the English test items?; (5) What is the difficulty level of the items making up English test?; (6) At what level do English test items powerfully discriminate between high and low achievers?; How effective are the distractors to ensure that English test outcomes provide more credible and objective picture of the knowledge of the examinees?

Method

This study used the quantitative approach with a cross-sectional survey. It was carried out within the period of two months, from the end of May to the mid-June 2017. The study took place across all senior high schools under the management of Muhammadiyah foundation. The schools are situated in Bantul District, Special Region of Yogyakarta, Indonesia. In order to successfully reach the objectives of this study, the schools which are homogeneous were considered.

Population and Sample

The population of this study was all Muhammadiyah high school students of grade-11 in the whole district of Bantul, totalling 241 students. In order to have accurate results, all of the students were selected as participants. By the small community, it is possible to conduct a study with nearly the whole population and pay attention to whoever has moved through the network of the community (Guyette, 1983). Therefore, this study uses the purposive sampling technique with total population sampling.

Data Collection Techniques

The technique used for data collection is documentation whereby the researchers recorded the answers from all examinees. To have information on the content covered by each teacher during the learning session, a questionnaire was used. With regard to the validity and reliability of the instruments in this study, experts' judgement and Crobach's Alpha indexes were computed.

Data Analysis

Within the scope of this study, there are a lot of variables to be measured, including construct validity, internal consistency reliability, item level of difficulty, the level of discrimination, and the effectiveness of the distractors. It is, therefore, clear that both Classical Test Theory (CTT) and Item Response Theory (IRT) are necessary in this analysis. Table 1 displays the variables and related data analysis techniques.

Table 1. Data analysis techniques

No	Variables	Analysis Techniques
1	Validity: Content Validity	Expert judgments with Aiken indices
2	Reliability: Internal Consistency Information Function	 JASP 0.8.2.0 = SPSS24 IRT/BILOG-MG 3
3	Level of Difficulty	IRT/ BILOG-MG 3
4	Power of Discrimination	IRT/ BILOG-MG 3
5	Distractor Effectiveness	Rasch/WINISTEPS 3.73

Table 1 contains the variables of the study and the analysis related to them. The coefficient of reliability which can be accepted must have a minimum of .70. This value helps to determine the level of error within measurement. The higher the index of reliability is, the higher the level of errors within measurement decreases, and vice versa (Mardapi, 2012, p. 128). Item discrimination (a_i) is the power of an item, by which its score is used for differentiating the examinees whose level of understanding is high from those whose level of understanding is low. The discrimination index is called *slope* because it shows the extent to which the probability to change the correct response like the ability or increase of the trait exists. According to Hambleton and Swaminathan (1985, p. 36), discrimination index varies from 0 to 2.

The item difficulty is another important variable. Its index (b_i) is always measured from the scores of students or examinees which are

obtained from the answers of all participants in a test. Item difficulty depends on the ability of the examinees. The more the testees have correct answers on an item, the higher the difficulty level of that item flops or decreases and vice versa. The item which is good or accepted is always situated between the interval of $-2 \leq \theta \leq 2$ (Hambleton et al., 1991, p. 13). The level of difficulty decreases as the *b*-parameter value is close to -2, but when the *b*-parameter value is close to +2, the level of difficulty increases.

The item analysis by using IRT model must fulfills the prescribed assumptions. The general assumptions that always appear in Item Response Theory models are unidimensional, local independent, and invariant parameters. The proof of unidimensionality is proven by the plot called Scree Plot as presented in Figure 1.

Figure 1 shows that a unidimensional assumption is fulfilled for this study data analysis because there is one most dominant dimension. The only way to test the model fitness is statistical measurement with chi-square. The researchers chose the suitable model by considering the highest percentage as shown by Figure 2 (Stone, Ye, Zhu, & Lane, 2009).

Figure 2 shows that the data in this study fit more to the second parameter model because it contains 36% (18 of 50) of all items. This result also supports the invariance assumption because when the data fit a model, the invariance criteria are automatically fulfilled (Lord, 2012, p. 126).

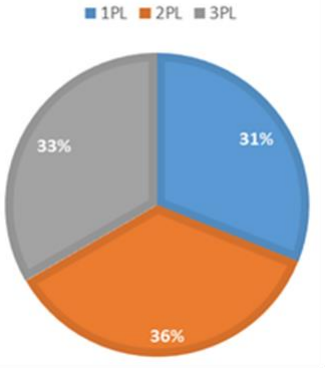

Figure 2. Goodness of fit (GoF)

The local independence has two facets: the local independence towards the test takers' answers and local independence towards the test items (Allen & Yen, 2001, p. 241). The first facet means that the wrong or right answer of a test taker does not depend on the wrong or right answer of his/her co-test taker on a given item. The second facet means that to be wrong or right on a test item does not affect the answer to another item. This study puts interest on the second facet of local independence because it is related to the test items. The results show that the correlation of residuals for all items is close to 0.

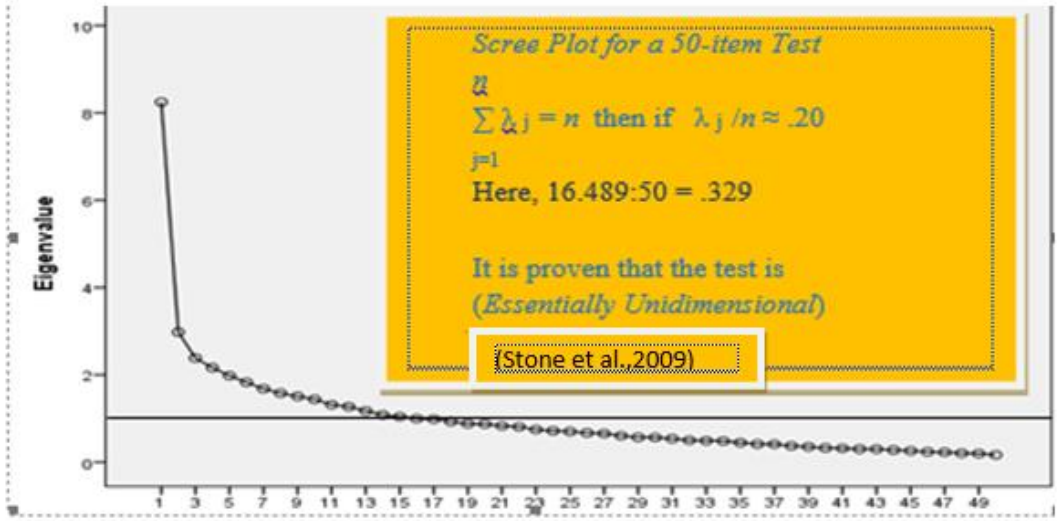

Figure 1. Unidimensionality proof by scree plot

Findings and Discussion

The findings of this study are discussed based on the variables to be measured. Validity index, reliability coefficient, discrimination index, difficulty level, distractor effectiveness, information function, and the success and content coverer weight were measured and the results can be found in this section. The findings about content validity with Aiken index are displayed in Figure 3.

Figure 3 contains information about the validity of items developed from the content expected to be covered by English teachers. It is supported that the English test represented the content taught because the Aiken Index for each indicator is accepted with the value bigger than .75. All items should be used because the overall index is .80. This result is supported by (Retnawati, 2016) who states that if the index is lower than or equal to .40, the validity is still low, if it is between .40 and .80, the validity is moderate, and if it is >.80, the validity is very high.

Reliability is another important criterion for item accuracy. Table 2 shows how reliable each item is. The Guttmann's Lambda$_7$ is the alternative of Cronbach's Alpha. Both coefficients were used to make a comparison.

The reliability coefficients are really good. Based on both Cronbach's and Guttmann's indices, the values range from .80 to .95. All items are perfectly reliable because any item's reliability greater than .70 is considered perfect, and the lowest and highest boundaries are .00 and 1.0 respectively. With this finding, there is no doubt that the students' answers to each item of the test are consistent. Hence, the test was measuring what it was purported to measure.

Table 2. Internal consistency reliability

Item	α	λ6	Item	α	λ7
Item1	0.88	0.92	Item26	0.89	0.93
Item2	0.88	0.92	Item27	0.88	0.92
Item3	0.88	0.92	Item28	0.88	0.92
Item4	0.88	0.92	Item29	0.88	0.92
Item5	0.88	0.92	Item30	0.88	0.92
Item6	0.88	0.93	Item31	0.88	0.92
Item7	0.88	0.92	Item32	0.88	0.92
Item8	0.88	0.92	Item33	0.88	0.92
Item9	0.88	0.92	Item34	0.88	0.92
Item10	0.88	0.92	Item35	0.88	0.92
Item11	0.88	0.92	Item36	0.88	0.92
Item12	0.88	0.92	Item37	0.88	0.92
Item13	0.88	0.92	Item38	0.88	0.92
Item14	0.88	0.92	Item39	0.88	0.92
Item15	0.88	0.92	Item40	0.88	0.92
Item16	0.88	0.92	Item41	0.88	0.92
Item17	0.88	0.92	Item42	0.88	0.92
Item18	0.88	0.92	Item43	0.88	0.93
Item19	0.88	0.92	Item44	0.88	0.92
Item20	0.88	0.92	Item45	0.88	0.92
Item21	0.88	0.92	Item46	0.88	0.92
Item22	0.88	0.92	Item47	0.88	0.92
Item23	0.87	0.92	Item48	0.88	0.92
Item24	0.88	0.92	Item49	0.88	0.92
Item25	0.88	0.92	Item50	0.88	0.92

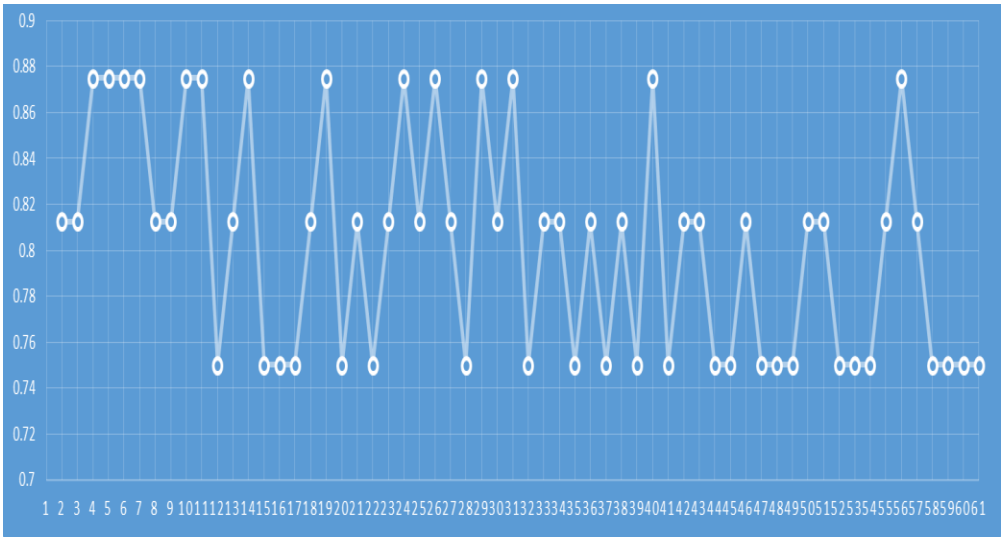

Figure 3. Aiken index (0.0 to 1.0)

The level of difficulty is very crucial to ensure the quality of test items. The results of *b*-parameter estimation for all English test items are summarized in Table 3.

Table 3. Difficulty index (*bi*)

Comment	Frequency	%
Good	47	94
Not good	3	6
Total	50	100

The parameter estimation for all 50 items shows that only three items (6%) are classified 'not good'. Those items are items 1, 40, and 46. The classification of item difficulty index relies on the range varying from -2 to 2 (good), and if it is out of the range, then it is not good. This result is in line with Mardapi (1991, p. 11) who states that the item difficulty level is the function of the ability of a test taker. An item is said to be good if it has the difficulty level (*bi)* between $-2 \leq b \leq +2$. An item with the difficulty level close or below -2 shows that the item is in an easy category. In contrary, an item with difficulty level (*bi*) close or above +2 shows an item that is in a difficult category. Figure 4 shows more about the accuracy of the test items based on *b*-parameter. The diagram in Figure 4 shows the test level of difficulty:

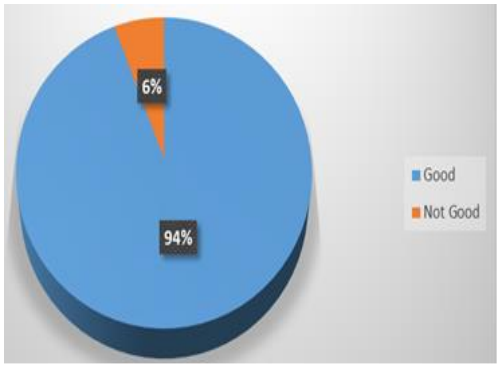

Figure 4. Item accuracy based on bi-index

Apart from the difficulty level, test items must be able to discriminate students by their abilities. The discrimination index for each item out of 50 items is well indicated in Figure 5. In terms of the discrimination index (*ai*), items 5, 10, 24, 35, 43, and 47, (12%) discriminate test takers at a low level because their *a-indexes* vary from between *.35 to -.64*. Items 6, 16, and 27, (6%) discriminate the examinees at a very low level because their *a-indexes* vary from .01 to .34.

However, the overall *a-index, 1.206,* shows that the English test moderately discriminates the examinees. Hence, all items with low discrimination indexes should be revised, while those with very low discrimination index should be replaced. The results are well shown in the diagram in Figure 5.

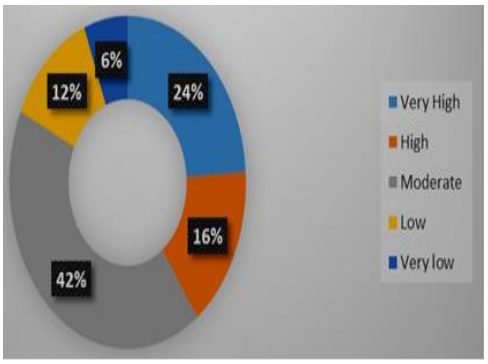

Figure 5. Discrimination power

The results presented in Figure 5 are supported by Baker (2001, p.34) that Discrimination Index (*ai*):

0.01 – 0.34 very low;
0.35 – 0.64 low;
0.65 – 1.34 moderate;
1.35 – 1.69 high;
1.70, and above very high.

Discrimination index (*ai*) is connected to the distractors' power to attract the examinees. The results can be seen in the diagram in Figure 6.

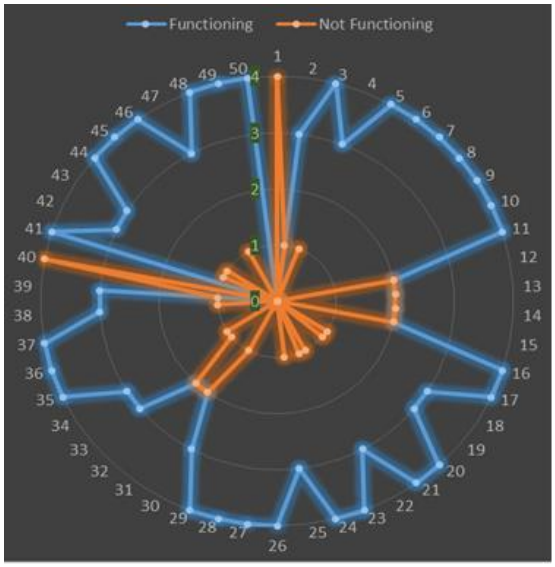

Figure 6. Distractor functionality

Notes:
0-4: Number of distractors (functioning ≥ 5% or not functioning ≤ 5%)
1-50: Number of items

It was found that items 1 and 40 (4%) do not have any functioning distractor, items 12, 13, 14, 15, 31, and 32 (6 items, 12%) have 50% of distractors that are not functioning effectively, items 2, 4, 18, 19, 22, 23, 25, 30, 33, 34, 38, 39, 42, 43, and 47 (15 item, 30%) have 25% of distractors that are not functioning, and items 3, 5, 6, 7, 8, 9, 10, 11, 16, 17, 20, 21, 24, 26, 27, 28, 29, 35, 36, 37, 41, 44, 45, 46, 48, 49, and 50 (27 items, 54%) have distractors that are functioning at 100%. In general, the English test for grade XI students during the second semester of the academic year of 2016/2017 has only 27 perfect items, two items that should be removed, and 21 items that should be repaired. Figure 6 represents the power of distractors within the test. These findings are supported by Abdulghani, Ahmad, Ponnamperuma, Khalil, and Aldrees (2014) who suggest that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for the effectiveness of distractors.

The information function is another indicator of test item accuracy. In the IRT, the information function stands for the reliability. In this study, the plot was used to easily see the amount of information the test could give, as presented in Figure 7.

The maximum information can be seen on the student's ability of -.04. On the other hand, the red line shows the error of measurement (SEM), the more information line picks, the fewer the error of measurement values drops. In fact, the majority of grade XI students have a low ability because the test gives much information on the left side from 0 on the latent trait. We can see that the test is fit for the students whose abilities vary from -.2.2 to 1.4. This is supported by Istiyono, Mardapi, and Suparno (2014).

As seen in Figure 8, around 70% (169 students) of all students (241) have a low ability to answer the questions. Therefore, there is no easy item for the students because their abilities are relatively low, -.40.

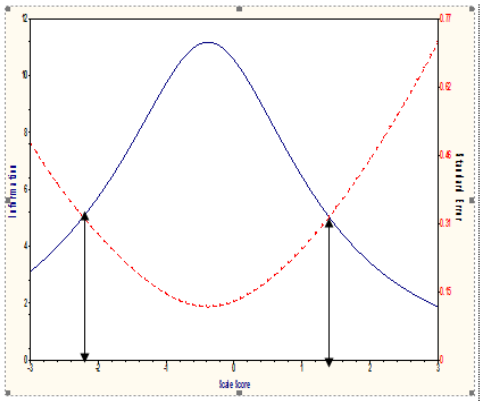

Figure 7. Information function (IF)

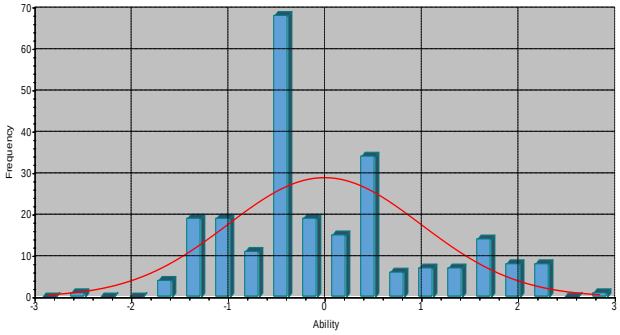

Figure 8. Proportition of students' abilities

As previously described, the content covered and the level of success were compared to see whether the students could understand that content well. Table 4 and Figure 9 have much information on the issue.

From Table 4, it is easy to evaluate, compare, and classify teachers at the end of a teaching term/period. It is known that every teacher has a syllabus that encloses the whole material. Every English teacher has the objectives to be achieved by the end of the term. A is given to an English teacher who reaches the target, B for a teacher who reaches an acceptable level, C for a teacher who needs to improve his/her teaching topics, D for a teacher who does not cover the content to the satisfaction, and F to a teacher that does cover a very minimum content. A= 82.5 to 100% of the content covered, B= 62.5 to 82.4% of the content covered, C= 42.5 to 62.4% of the content covered, D= 22.5 to 42.4% of the content covered, and F= 20% and below the content covered.

Apart from the teacher categorization criteria above, the new teacher project, as cited by Seidel, Stürmer, Blomberg, Kobarg, and Schwindt (2011), suggests a way to give scores to teachers. In the report called *Rating a Teacher Observational Tool,* the teachers can be put into categories, including: 'complete coverage' when the tool of evaluation covers all the elements in the curriculum, 'partial coverage' when the test does not cover some components of the syllabus, and 'inadequate coverage' when the evaluation tool covers lower than 50% of all indicators in the syllabus. Figure '3' stands for the first category, '2' for the second, and '1' for the third. Based on the answers of the teachers, all six teachers were categorized.

With Figure 9, it is easy to see the gap between the content covered and the success level of grade XI students. There are some English teachers, ENGT.BL, ENGT.SW, ENGT.PL, who show that content and success are in line, but the rest of the teachers, ENGT.PY, ENGT.IM, ENGT.KS, indicate a long gap between the content covered and success of students on the English test. Information from Figure 9 implies that there is a remarkable difference between rural and urban Muhammadiyah senior high schools. For the rural schools, the content covered by English teachers does not explain the success level of students on the test developed from that content, but for the urban schools, there is correlation between the content covered and the success level of the students.

Table 4. Classification of English teachers

ID CODE	Indicators Covered/61	Scale	Grade	Comment	Category	Comment
ENGT.BL	53.00	3.5	A	Reached Target	2	Partially Covered
ENGT.PY	53.00	3.5	A	Reached Target	2	Partially Covered
ENGT.IM	52.00	3.4	A	Reached Target	2	Partially Covered
ENGT.SW	49.00	3.2	B	Acceptable	2	Partially Covered
ENGT.KS	44.00	2.9	B	Acceptable	2	Partially Covered
ENGT.PL	37.00	2.4	C	Need Improvement	2	Partially Covered

Figure 9. Content covered vs success

Conclusion, Implications, and Suggestions

In connection with the results of this study and its discussion within the previous chapter on the accuracy of the multiple-choice items of the English test, different concluding statements can be made as follows: (1) The items represent the content taught to the students during the second semester of academic year 2016/2017. (2) All items are internally consistent. (3) Most of the items (47/50) have acceptable difficulty level, but there are two items which are very easy and one which is very difficult. (4) A big number of items (42/50) have good discrimination indexes, but nine items are unable to discriminate the high achievers from low achievers. (5) As many as 27 items have effective distractors, but 23 items still show powerless distractors. (6) Some English teachers tried their best to cover the content expected to be taught to the students, but some others did not cover at least 50% of the content, and therefore, there is still a gap (for some schools) between the content covered and the success level of the students on the test developed from that content. (7) The test is obviously difficult for more than 70% of the students who have the ability of -.40, and fits the students whose abilities range from -2.2 to 1.2.

Like in other scientific studies, some implications are put forward that the improvement in constructing and developing English test items for grade XI students of Muhammadiyah senior high schools in Bantul district needs both qualitative and quantitative review. It is necessary to test the quality of each item. This process contributes to the identification of some weaknesses within the test because the quality level of a test is completely determined by the quality of its items.

The results of the quantitative analysis of the English test, in general, are not accurate. The teachers should make some try outs of the items, then the results are analyzed with relevant and practical techniques, such as the item analysis with the classical test theory and item response theory as well. The determination of the technique of analysis depends on the purpose and number of examinees accompanied by other technical assessments. An analysis with the classical test theory needs a small sample (30 participants at minimum), but the item response is used for a big number of respondents.

For a better future school-based assessment, the following suggestions are given: (1) All items with medium quality should be revised, re-measured until they fulfill the criteria of a good item; the items with bad quality should be dropped or completely replaced. (2) It is much better for the teachers to conduct some tryouts and analysis of items before testing. (3) It is quite advisable for the teachers to develop items that are suitable to the content that is already taught to the students; they should also give the blueprint to them. (4) Before a set of items are chosen, it is necessary to conduct qualitative analysis with expert judgment. It can help English teachers to have information on the item characteristics in terms of construction, language, and content in general. (5) The item response theory is needed to identify the characteristics of items; IRT related programs should be trained to teachers of senior high schools. (6) It is suggested to make a test item bank at the district level (Bantul) for the English subject to help teachers practice in assessing students' achievements. (7) Schools should prepare some routine trainings on evaluation, assessment, and measurement. It will help to increase the ability English teachers in evaluating learning outcomes. The management office of Muhammadiyah schools should be vigilant to remote areas in terms of education and technology.

References

Abadyo, A., & Bastari, B. (2015). Estimation of ability and item parameters in mathematics testing by using the combination of 3PLM/ GRM and MCM/ GPCM scoring model. *REiD (Research and Evaluation in Education)*, *1*(1), 55–72. https://doi.org/10.21831/reid.v1i1.4898

Abdulghani, H. M., Ahmad, F., Ponnamperuma, G. G., Khalil, M. S., & Aldrees, A. (2014). The relationship between non-functioning distractors and

item difficulty of multiple choice questions: A descriptive analysis. *Journal of Health Specialties*, *2*(4), 148–151. https://doi.org/10.4103/1658-600X.142784

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory* (1st ed.). Long Grove, IL: Waveland Press.

Boopathiraj, C., & Chellamani, K. (2013). *Analysis of test items on difficulty level and discrimination index in the test for research in education*. *International Journal of Social Science & Interdisciplinary Research* (Vol. 2).

Brescia, W., & Fortune, J. C. (1989). Standardized testing of American Indian students. *College Student Journal*, *23*(2), 98–104.

Charismana, D. S., & Aman, A. (2016). Analisis kualitas tes ujian akhir semester PPKN SMP di Kabupaten Kudus. *Jurnal Evaluasi Pendidikan*, *4*(1), 1–9.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 1–23. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Galsworthy, M. J., Paya-Cano, J. L., Liu, L., Monleón, S., Gregoryan, G., Fernandes, C., … Plomin, R. (2005). Assessing reliability, heritability and general cognitive ability in a battery of cognitive tasks for laboratory mice. *Behavior Genetics*, *35*(5), 675–692. https://doi.org/10.1007/s10519-005-3423-9

Gronlund, N. E. (1993). *How to make achievement tests and measurements*. Needham Heights, MA: Allyn and Bacon.

Guyette, S. (1983). *Community-based research: A handbook for native Americans*. Los Angeles, CA: American Indian Studies Center, University of California.

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

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

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (PhysTHOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Joint Committee on Testing Practices of American Psychological Association. (2004). *Code of fair testing practices in education*. Washington, DC, United States of America.

Kartowagiran, B. (2012). *Penulisan butir soal*. A paper presented in the Seminar on Question Items Analysis and Writing for Civil Servant Resources of Dik-Rekinpeg, in Kawanua Aerotel Hotel.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Mardapi, D. (1991). Konsep dasar teori respons butir: Perkembangan dalam bidang pengukuran pendidikan. *Cakrawala Pendidikan*, *3*(X), 1–16.

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

Mkrtchyan, A. (2011). Distractor Quality Analyze In Multiple Choice Questions Based On Information Retrieval Model. *EDULEARN11 Proceedings*, 1624–1631.

Osadebe, P. U. (2015). Construction of valid and reliable test for assessment of students. *Journal of Education and Practice*, *6*(1), 51–56.

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

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent*

Education, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Yogyakarta: Parama Publishing.

Sabri, S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. *International Journal of Education and Research*, *1*(12), 1–14.

Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education: An International Journal of Research and Studies*, *27*(2), 259–267. https://doi.org/10.1016/j.tate.2010.08.009

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63–86. https://doi.org/10.1080/08957340903423651

Young, M., Cummings, B.-A., & St-Onge, C. (2017). Ensuring the quality of multiple-choice exams administered to small cohorts: A cautionary tale. *Perspectives on Medical Education*, *6*(1), 21–28. https://doi.org/10.1007/s40037-016-0322-0

REiD (Research and Evaluation in Education), 4(1), 2018, 58-69
Available online at: http://journal.uny.ac.id/index.php/reid

Developing an instrument of national examination of equivalency education Package C of mathematics subject

[*1]Ian Harum Prasasti; [2]Edi Istiyono
[1]Sekolah Tinggi Manajemen Informatika dan Komputer Kalirejo (STMIK Kalirejo)
Jl. Jend. Sudirman, Kalirejo, Kabupaten Lampung Tengah, Lampung 34174, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: ihp.harum8@gmail.com

Submitted: 30 August 2017 | Revised: 26 April 2018 | Accepted: 17 May 2018

Abstract

The national examination of equivalency education is a competency test to equalize non-formal education with formal education. Departing from the importance of the quality and expectations of the national examination of equivalency education package C mathematics subjects and because the results are inseparable from the implementation process, developing an evaluation instrument to assess the implementation of the national examination of equivalency education package C mathematics subjects is important. The purpose of this study is to develop a suitable instrument for conducting an evaluation of the national examination implementation of the equivalency education package C mathematics subject. The respondents in this research are package C test takers in Bantul Regency, Yogyakarta. The data were analyzed by using SPSS 20.0 and Lisrel 8.54. The result of the analysis shows (1) based on the data obtained from respondents of try-out, the developed instrument is valid, reliable and qualified as a fit model; (2) components in the instrument of test takers is learning time, socialization, test materials, and test venue; and (3) the instrument have the validity value of > 0.40 and reliability coefficient of > 0.70.

Keywords: *instrument, evaluation, national examination, equality education*

Introduction

In its effort to build and monitor the quality of education and to meet the needs of equity in the education aspect, the Indonesian government has continuously made a policy to develop the national standard competency test instruments. One of the efforts undertaken is through the provision of Government Regulation No. 19 of 2005, on National Education Standard in Article 3 that is a basis for the planning, implementation, and supervision of education in order to realize a quality national education.

The nationally standardized test of competence, or commonly known as national examination, aims to conduct coaching and provide assistance to schools in an effort to improve the quality of education (Mardapi & Kartowagiran, 2009). In addition, the toughest goal of the national examination is to improve clarity, efficiency, and also effectiveness in making decisions (Adow, Alio, & Thinguri, 2015). National examination also aims to measure students' learning achievement in certain subjects that are grouped into science and technology to assess the achievement of the national education standards (Mudjijanti, 2011).

The competency test is in the form of national examination and national examination of equivalency education. The use of the term national examination of equality education is due to the position of the exam results which can be accounted for an equivalent to the results of formal education exams. The purpose of the national examination is for the mapping of the quality of schools, the selec-

ISSN 2460-6995

tion of entry into the next level of education, and provision of schools in an effort to improve the quality of education. It can also be categorized as a diagnostic test (Setiadi, et al. 2011). The national examination of equivalency education is a competency test to equalize non-formal education in the form of Package A equivalent to elementary school, Package B equivalent to junior high school, and Package C equivalent to senior high schools.

Package C as one of the national equivalency education examination programs is aimed to solve educational problems that cannot be coped by formal education. Some factors in the non-formal education which have not been solved include a problem in senior high schools, traumatic experience, school drop-outs, and hyperactive and autism children. Thus, for equivalency of the non-formal education with formal education, the government runs programs of the national equivalency examination.

The term 'national equivalency examination' is used since the result of equivalency examination is credible and accountable, and its position is equivalent to the result of national examination of formal education. Likewise, one of the efforts undertaken by the government through the provision of Law No. 20 of 2003 of Republic of Indonesia on National Education System in article 26 verse 6 explains that the result of non-formal education can be equivalent with the result of formal education program after going through an equivalent assessment process by institutes selected by the government. Then, the national examination for equvalency education participants will automatically get a certificate from a non-formal educational institutions such as the learning group of Package C (Raharjo, 2012).

Every educational activity needs an evaluation activity to know the level of success of the implementation of the activity in accordance with the intended purpose. According to Sudjana (2006), evaluation is a necessity and fairness needed in the management of a program. According to Worthen and Sanders (1981, p. 20), 'evaluation is viewed as a process of identifying and collecting information to assist decision-makers in choosing between available decision alternatives'. Through different words, but with almost identical meanings, evaluation is described as a planned process to obtain information related to the achievement of a goal (Kartowagiran, 2013).

Evaluation is able to answer the variation of the statement and determine the success in viewing the quality of education. Rossi and Freeman (1985, p. 46) state that evaluations are conducted to answer a variety of questions related to what we have listed as the three foci of evaluation research: program conceptualization and design, program implementation, and program utility.

Weiss (1972, p. 4) writes, 'the purpose of evaluation research is to measure the effects of a program against the goals; it sets out to accomplish as a means of contributing to subsequent decision making about the program and improving future programming'. Rossi and Freeman (1985, p. 50) write that evaluation result, both from monitoring program implementation and from assessing impact and efficiency, can influence decisions on the expansion, continuation, or termination of the program and the organizations responsible for them.

This study examines the subjects of mathematics, a branch of science that has a very important role in various activities in everyday life, which can even be more than that. Thus, activities in everyday life cannot be separated from the use and application of concepts that exist in mathematics, so the unique characterization of mathematics learning is where the benefits are almost perceived in everyday life and become a key opportunity and have the contribution to other sciences.

Related to the process of its formation, mathematics is the knowledge that humans have. This knowledge arises because humans need to understand the natural world. Nature is used as a source of ideas for obtaining mathematical concepts through abstraction and idealization (Kartowagiran, 2008). If math skills can be well developed, then math can be an opportunity. This is in line with Mathematical Sciences Education Board of National Research Council (1993, p. 15) who states:

> *'. . . mathematics is the key to opportunity. No longer just the language of science, mathematics*

> *now contributes in direct and fundamentally to business, health, and defense. For the student, it opens doors to careers. For citizens, it enables informed decisions. For nations, it provides knowledge to compute in a technological community'.*

In addition, Hatfield, Edwards, Bitter, and Morrow (2008, p. 3) state that mathematics is nothing to be afraid of; it is our human heritage from all cultures. Clarifying the statement, Kahn and Kyle (2002, p. 15) explain, 'Mathematics is not fundamental too much of science and technology but needs an analytical model-building approach, whatever the discipline is'. Typically, it will be argued that mathematics claims a place in the curriculum because it can be seen as (1) contributing to the basic knowledge of any educated citizen; (2) contributing to the study and advancement of numerous disciplines, professions, and trades; (3) contributing to a student's general education through the inculcation of particular attitudes or approaches; (4) possessing an inherent interest and appeal (Christiansen, Howson, & Otte, 1986, p. 9).

The results of UNPK (*Ujian Nasional Pendidikan Kesetaraan* or National Examination of Equivalency Education) not only give results about the state of education but also provide information on improving students' learning achievement. This expectation is achieved when the data obtained are valid and reliable. In other words, the result has the smallest possible measurement error. The measurement error is divided into two: random, caused by the selection of exam materials and the condition of the examinees, and systematic, because the problem is too easy or too difficult and the implementation does not follow the guidelines, such as the regulations and operational standards of implementation (Mardapi, 2012).

The real examples of measurement error were taken from research by Kartowagiran (2008) about UAN (*Ujian Akhir Nasional* or National Examination) test device, that is, UAN Mathematics test device in 2003, 2005, and 2006 which measure three sub-dimensions of algebra, geometry, and measurement. The research found that the test devices are able to explain only 35% variance of math ability of the learners. In this regard, the test developer should attempt to increase the factor and variance of the difficulty level of the test items.

Starting from the importance of the quality and expectations of the National Examination of Equivalency Education of Package C mathematics subjects and because the results are inseparable from the implementation process, it is really important to develop evaluation instruments for the implementation of National Examination of Equivalency Education of Package C mathematics subjects. The purpose of this study is to develop a suitable instrument for conducting the evaluation of the national examination implementation of the equivalency education Package C mathematics subject.

Method

This research and development aims to produce a particular product, and test the effectiveness of the product. The product developed is a questionnaire of the implementation of National Examination of Equivalency Education Package C consisting of 25 items. The developmental procedure referred to the modified development steps which are proposed by Mardapi (2005, pp. 16–21), as follows: (1) base on theories about the concept to be measured, as construct variables, (2) develop dimensions and indicators, (3) make instrument gratings, (4) assign quantities or parameters, (5) list instrument items, (6) validate the process, (7) revise the draft, and (8) implement the test to the Package C takers as the participants.

The evaluation instrument of National Examination of Equivalency Education Package C of mathematics subjects evaluates the standard operational procedure including the preparation, implementation, and result of the national examination. The instrument used was Likert scale modification or summative rating with the highest score per item is 4 and the lowest score per item is 1. The modified Likert scale has four options: 4 (always/ strongly agree), 3 (often/agree), 2 (rarely/ rather disagree), and 1 (never/disagree). The four-point scale summative rating was used because according to Mardapi (2012), in the data retrieval if using Likert scale, a five-

alternative choice such as 5 (strongly agree), 4 (agree), 3 (doubt/neutral), 2 (agree), and 1 (strongly disagree) makes respondents often experience a tendency to choose category 3 (undecided/neutral).

The respondents in this research are Package C test participants of equality education in Bantul, Yogyakarta, Indonesia. In the try-out, the instrument was administrated to 190 participants of the examination participants. The analysis of the try-out data was to obtain evidence of construct validity and reliability of the instrument. The construct validity measurement in this study used factor analysis that serves to summarize or reduce observation variables into new dimension forms that present the main variables (factors). The proof of the construct validity used exploratory factor analysis which aims to investigate the factors in the observation, and confirmatory factor analysis with the aim to confirm a theory of measurement in order to compare theories with the empirical results. The data collecting instruments with the test participants as respondents were analyzed using the exploratory factor analysis with the help of SPSS 20.0 program and followed by the confirmatory factor analysis with the help of Lisrel 8.54 program.

This research used two main factor analysis techniques: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). The CFA attempted to confirm the hypotheses and used the path analysis diagrams to represent variables and factors, and then the EFA tried to uncover complex patterns by exploring the dataset and testing predictions (Child, 2006).

The criteria in the EFA analysis must meet the following criteria: Keyser Mayer Oikin (KMO) values greater than 0.5; and the significant value of Barlett's Test of Sphericity is less than 0.05 (Ghozali, 2005). In addition, the eigenvalues of the total variances explained is greater than 1.0 and the coefficient of the Rotated Component Matrix is greater than 0.40, and the loading value of the factor is greater than that of other factors with a difference of at least 0.10 indicating a correlation between test items with a factor formed (Azwar, 2015).

Furthermore, Hendryadi and Suryani (2014) state that the criteria in the CFA analysis that can determine the suitability of the model with the help of Lisrel 8.54 can be determined as follows: (1) chi-square with p-value > 0.05; (2) Root Mean Square Error of Approximation value ≤ 0.08. Root Mean Square Error of Approximation (RMSEA) is a value that attempts to correct the trend of chi-square statistics rejecting the model; (3) the value of Goodness of Fit Index (GFI) ≥ 0.90 means that the model tested has a good match. GFI is an index that describes the overall suitability of the model of the predicted model compared to the actual data; (4) T-value ≥ 1.96 at the significance level of 0.05; and (5) Standardized loading factor > 0.5.

Furthermore, the reliability of the instrument was measured by using Alpha Cronbach formula with the help of SPSS 20.0 software and Stratified Alpha coefficient, and the reliability of the construct was measured by using the construct reliability formula. The reliability formula used in this study is as follows.

(1) Stratified Alpha Coefficient

$$\alpha_{strat} = 1 - \frac{\sum \sigma_i^2 (1 - \alpha_i)}{\sigma_x^2}$$

(2) Construct Reliability(CR)

$$CR = \frac{(\sum \text{Standarized loading})^2}{(\sum \text{Standarized loading})^2 + \sum \text{Measurement Error}}$$

(Hendryadi & Suryani, 2014)

The magnitude of the reliability index is at least 0.70 because the greater the reliability index the smaller the measurement error (Mardapi, 2012).

Findings and Discussion

The implementation of the try-out of the test is approved by using the exploratory factor analysis (EFA) with SPPS 20.0 and followed by the confirmatory factor analysis (CFA) with the help of Lisrel 8.54. through several stages of factor analysis, three times EFA and twice CFA. The steps of factor analysis to get the expected result are explained as follows.

Exploratory Factor Analysis 1

The result of the validation with exploratory factor analysis 1 on KMO value and Barlett's Test sig value is shown in Table 1.

Table 1. Exploratory factor analysis 1

Item	KMO	Sig *Barlett's Test*
25	0.851	0.000

Table 1 shows the value of KMO of 0.851 with a Barlett's test value of 0.000. These results show the KMO value> 0.5 and the Barlett's test < 0.05. Hence, it can be concluded that the sample size used in this factor analysis is sufficient so that the EFA analysis can proceed to the next step. Furthermore, the number of components or clusters formed from the 25 items of the statement can be seen from the total initial eigenvalues > 1.0 shown in Table 2.

Table 2. Eigenvalues 1

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	6.780	27.119	27.119
2	2.538	10.154	37.273
3	1.900	7.600	44.873
4	1.730	6.921	51.794
5	1.101	4.405	56.199

Table 2 shows that the total initial eigenvalues > 1.0. Thus it can be concluded that there are 5 components formed from 25 items in the instrument with the variance described as 56.199%. Furthermore, the number of factors in the instrument can be seen in the scree plot shown in Figure 1.

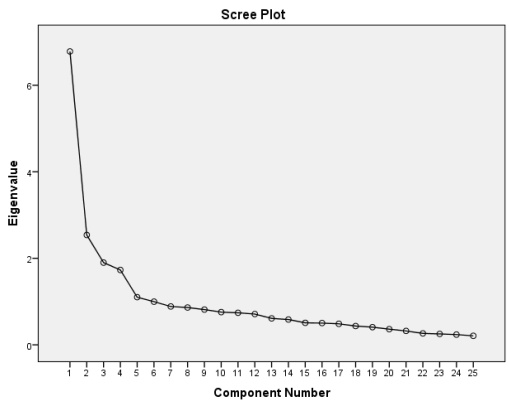

Figure 1. Scree plot EFA 1

Figure 1 shows the number of factors marked by a steep graph of eigenvalue value gain. Based on Figure 1 then, there is one dominant factor and the other four factors also contributing substantially to the component of the variance that can be explained, so that the instrument shows it measures at least five factors that are formed. Thus, it can be concluded that all of the items can be analyzed further through factor analysis with the extraction and rotation method using varimax and obtained the results as shown in Table 3.

Table 3. Rotated component matrix 1

	Component				
	1	2	3	4	5
A1	0.637				
A2	0.709				
A3					
A4	0.772				
A5	0.788				
A6	0.661				
A7	0.465				
A8	0.697				
B1		0.725			
B2		0.798			
B3		0.458			
B4		0.756			
B5		0.578			
B6		0.620			
B7					
B8					0.736
C1			0.442		
C2			0.752		
C3			0.820		
C4					
C5			0.634		
D1				0.766	
D2				0.823	
D3				0.803	
D4				0.590	

Table 3 shows the value of the loading factor does not meet the specified criterion, that is ≥ 0.4 and the difference with another factor > 0.1. There are three invalid items: A3 (item 3) which forms learning time factor, B7 (point 15) which forms socialization formation factor, and C4 (item 20) which forms examination material factor. This can happen because of the different interpretation between the researchers and respondents. Because the items are not good to use, then the invalid items are discarded then proceed with second exploratory analysis with 22 items.

Exploratory Factor Analysis 2

The second exploratory factor analysis was performed after the invalid items were discarded. Thus, there were 22 items left to be analyzed.

Table 4. Exploratory factor analysis 2

Item	KMO	Sig Barlett's Test
22	0.838	0.000

Table 4 shows the gain of KMO value of 0.838 with a Barlett's test value of 0.000. This result shows KMO > 0.5 and Barlett's test < 0.05. It can be concluded that the sample used in this research is adequate. Furthermore, the number of components or clusters formed from the 22 items can be seen from the total initial eigenvalues > 1.0.

Table 5. Eigenvalues 2

Component	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	6.216	28.254	28.254
2	2.446	11.119	39.372
3	1.718	7.811	47.183
4	1.592	7.235	54.419

Table 5 shows the total initial eigenvalues is > 1.0. It can be concluded that there are four clusters formed from 22 items on the sheet of an instrument with a cummulative percentage of 54.419% and it explains the variance. Furthermore, the scree plot also shows there are four dots that are above the value of 1 because the number of factors is marked by a steep graph of eigenvalue value gain, then there is one dominant factor and three other factors also contributing substantially to the cluster variance that can be ex-plained so that the instrument measured at least four factors and clarified on the scree plot as in Figure 2

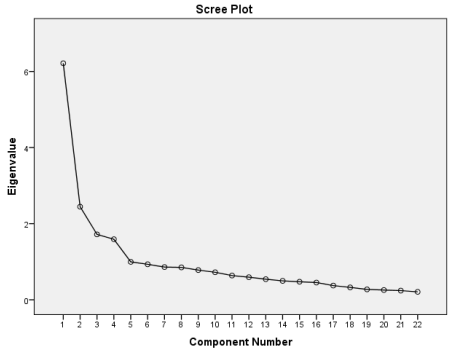

Figure 2. Scree plot EFA 2

Thus, it can be concluded that the whole items can be analyzed further through the factor analysis with the extraction and rotation method using varimax and it obtained the results as in Table 6.

Table 6. Rotated component matrix 2

	Component			
	1	2	3	4
A1	0.646			
A2	0.725			
A4	0.773			
A5	0.807			
A6	0.657			
A7	0.457			
A8	0.698			
B1		0.730		
B2		0.780		
B3				
B4		0.767		
B5		0.597		
B6		0.664		
B8				
C1				0.474
C2				0.801
C3				0.849
C5				0.615
D1			0.760	
D2			0.820	
D3			0.803	
D4			0.596	

Table 6 shows that the value of the loading factor does not meet the critereon specified that is ≥ 0.4 and difference with other factors > 0.1. There are two invalid items, namely B3 (point 11) and B8 (point 16), the items which form the socialization factor. The invalid items may be caused by the difference of interpretation between the researchers and the respondents or the items are unfavorable to use. Thus, the invalid items are discarded and then followed by the third exploratory factor analysis with 20 items.

Exploratory Factor Analysis 3

This third exploratory factor analysis was performed after invalidating the invalid items. Therefore, 20 items can be analyzed.

Table 7. Exploratory factor analysis 3

Item	KMO	Sig Barlett's Test
20	0.834	0.000

Based on Table 7, KMO value is 0.834 with Barlett's test significance value of 0.000.

These results show that the KMO value is > 0.5 and value of Barlett's test is < 0.05. It can be concluded that all items in the instrument can be analyzed further. Furthermore, the number of components or clusters formed by the 20 items of the statement can be seen from the total initial eigenvalues > 1.0.

Table 8. Eigenvalue 3

Comp	Initial Eigenvalues		
	Total	Variance (%)	Cumulative (%)
1	5.804	29.020	29.020
2	2.410	12.052	41.071
3	1.680	8.398	49.469
4	1.552	7.760	57.230

Table 8 shows the total initial eigenvalues is > 1.0. Thus, it can be concluded that there are four components formed by 20 items in the instrument with a percentage value of 57.230% variance that can be explained. A model that is a good fit will have less than 50% of the non-redundant residuals with absolute values that are greater than 0.05 (Yong & Pearce, 2013). Furthermore, the number of factors in the instrument can be seen through the scree plot shown in Figure 3.

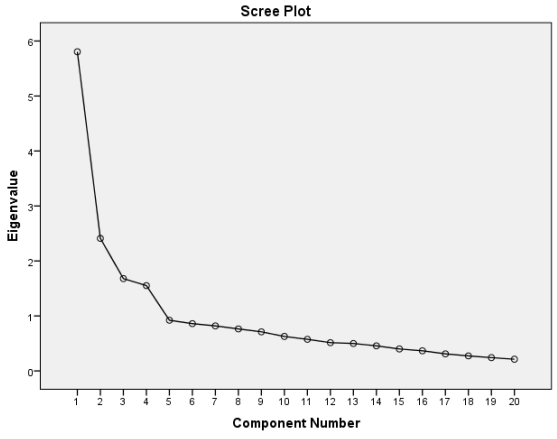

Figure 3. Scree plot EFA 3

Figure 3 shows that the number of factors is marked by a steep graph of eigenvalue gain. Based on the figure, there is one dominant factor and the other three factors also contributing substantially to the component of variance that can be explained and they begin to ramp up on a fifth factor. This indicates that the instrument shows at least four factors. The scree test consists of eigenvalues and factors (Cattell, 1978). The scree test is only reliable when the sample size is at least 200. In situation when the scree test is hard to interpret, it is necessary to rerun the analysis several times and manually set the number of factors to extract each time (Costello & Osborne, 2005).

Thus, it can be concluded that whole items can be analyzed further through factor analysis with the extraction and rotation method using varimax aiming to clarify the items included in the component. Yong and Pearce (2013) write that factors are rotated for better interpretation, since unrotated factors are ambiguous. Thus, the results obtained can be seen in Table 9.

Table 9. Rotated component matrix 3

	Component			
	1	2	3	4
A1	0.650			
A2	0.724			
A4	0.770			
A5	0.804			
A6	0.660			
A7	0.463			
A8	0.697			
B1		0.740		
B2		0.787		
B4		0.766		
B5		0.596		
B6		0.656		
C1				0.505
C2				0.810
C3				0.849
C5				0.622
D1			0.743	
D2			0.827	
D3			0.806	
D4			0.607	

Based on the EFA analysis, the four components formed by the 20 items are elaborated as follows: (1) The items related to group learning time are clustered in component 1; (2) the items related to socialization are clustered in component 2; (3) the items related to examination material are clustered in component 4; and (4) the items associated with the examination room are grouped in component 3. On the other hand, the clusters obtained by the exploratory factor analysis are then analyzed by using the Confirmatory Factor Analysis (CFA).

Confirmatory Factor Analysis

Before calculating the validity of the construct by using the CFA, the assumption of normal distribution is firstly tested. This normality test can be seen from the univariate normality test result that describes the distribution of one variable in the respondent, whereas the multivariate normality test result provides an overview of the shared distribution of all variables in the respondent. The calculations in this study employed Lisrel 8.54 program.

The results of the univariate normality analysis are shown in Table 10. Furthermore, the summary results of multivariate normality calculations are shown in Table 11.

The results of the univariate normality analysis showed that the data did not meet the normal univariate assumptions (p value skewness and kurtosis <0.05), in line with the results of the normal multivariate test which was not fulfilled (p value skewness and kurtosis <0.05). It can be concluded that the data used do not meet normal univariate or multivariate assumptions. Normal univariate distribution of each item is required, but multivariate distribution is more important because in general, data with no normal univariate distribution will result in a multivariate non-normal distribution (Hendryadi & Suryani, 2014).

Furthermore, due to the abnormal data, this research used an alternative estimation method that is Robust Maximum Likelihood (RML) by adding asymptotic covariance matrix which is useful for correcting the chi-square statistic value, commonly known as Satorra-Bentler Scaled Chi-Square. Maximum Likelihood attempts to analyze the maximum likelihood of sampling the observed correlation matrix (Tabachnick & Fidell, 2007). The Maximum Likelihood is more useful for confirmatory factor analysis (Yong & Pearce, 2013).

Table 10.Test of Univariate Normality for Continuous Variables

	Skewness		Kurtosis		Skewness and Kurtosis	
Variable	**Z-Score**	**P-Value**	**Z-Score**	**P-Value**	**Chi-Square**	**P-Value**
A1	1.917	0.055	-1.281	0.200	5.316	0.070
A2	-0.126	0.900	-1.711	0.087	2.945	0.229
A4	-3.319	0.001	-0.690	0.490	11.492	0.003
A5	0.608	0.543	-1.932	0.053	4.101	0.129
A6	-1.237	0.216	-2.507	0.012	7.817	0.020
A7	-1.785	0.074	-0.262	0.793	3.254	0.197
A8	0.732	0.469	-3.458	0.001	12.478	0.002
B1	-2.578	0.010	-1.498	0.134	8,890	0.012
B2	-3.566	0.000	0.749	0.454	13.276	0.001
B4	-3.673	0.000	1.550	0.121	15.892	0.000
B5	-4.506	0.000	2.549	0.011	26.808	0.000
B6	-2.866	0.004	1.552	0.121	10.626	0.005
C1	-3.338	0.001	2.086	0.037	15.493	0.000
C2	-5.979	0.000	2.360	0.018	41.311	0.000
C3	-5.837	0.000	2.415	0.016	39.907	0.000
C5	-2.525	0.012	1.634	0.102	9,047	0.011
D1	-3.361	0.001	2.243	0.025	16.322	0.000
D2	-2.063	0.039	0.388	0.698	4.407	0.110
D3	-1.836	0.066	0.238	0.812	3.428	0.180
D4	-4.165	0.000	-0.559	0.576	17.660	0.000

Table 11.Test of multivariate normality for continuous variables

Skewness			Kurtosis			Skewness and Kurtosis	
Value	**Z-Score**	**P-Value**	**Value**	**Z-Score**	**P-Value**	**Chi-Square**	**P-Value**
8.177	18.276	0.000	518.921	10.858	0.000	451.914	0.000

The CFA (Confirmatory Factor Analysis) was based on an exploratory analysis which resulted in four components and was supported by the existence of theory, and then was subsequently analyzed by the confirmatory analysis. The calculation of CFA was done twice with the help of Lisrel 8.54 program. The first result was Root Mean Square Residual (RMR) = 0.0346, Goodness of Fit Index (GFI) = 0.881 and Root Mean Square Error of Approximation (RMSEA) = 0.0421 and Satorra-Bentler Scaled Chi-Square = 281.932 with P-value 0.00268. The standardized solution model of the Package C execution of mathematics subjects is clearly presented in Figure 4, while when it is seen from t-value, the result is shown in Figure 5.

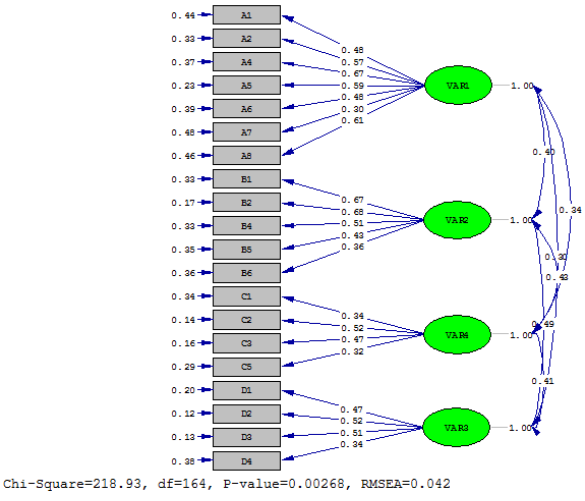

Figure 4. Standardized solution 1

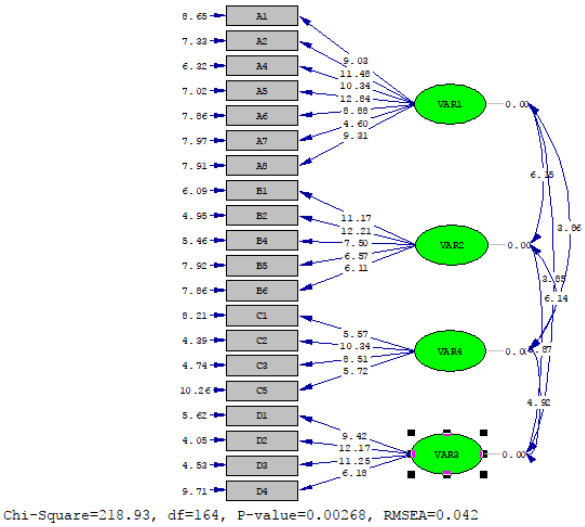

Figure 5. T -value 1

When viewed from the results and models obtained, P-value 0.00268 < 0.05, this indicated that the factor model used by all the tests was not good (the model was not fit). Therefore, to get the fit model, model respecification or model modification by looking at the modification indices to see the items that correlate each other was done. The results of the second-factor analysis calculation, after getting the correlated items, was gained through using modification indices as a reference. Thus, the standardized solution model 2 of the implementation of the Package C mathematics subject through Confirmatory Factor Analysis 2 is presented in Figure 6, while when it is seen from t-value, the result is shown in Figure 7.

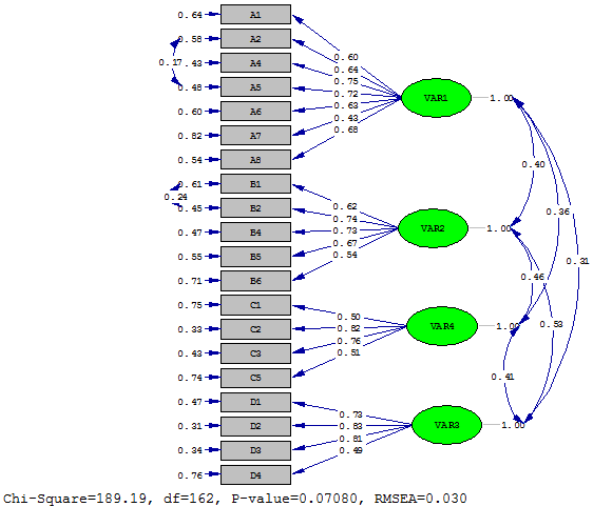

Figure 6.Standardized solution 2

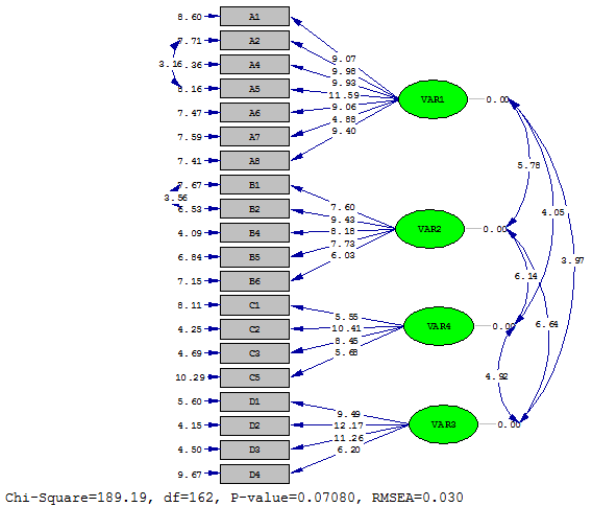

Figure 7. T -value 2

The results of this CFA 2 output show that the value of Goodness of Fit Index (GFI) = 0.894. and Root Mean Square Error of Approximation (RMSEA) = 0.030 < 0.080 (good fit) and Satorra-Bentler Scaled Chi-Square = 189.186 with P-value 0.0708 > 0.050 (good fit). Seen from the results and match models, the proposed model has a good match or the proposed model matches the data and the conceived items measure only the latent variables. The correlated items are due to identical statements. The result of the weighted coefficient significance of the 20 items rated on Standardized Loading Factors (SLF) shows that two items have less than 0.5, i.e. items A7 and D4. However, the overall value of t-value is > 1.96. Thus, there are two items with poor validity: items A7 and D4. Thus, based on the Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), it can be concluded that the instrument of the National Examination of Equivalency Education Package C is valid to measure the implementation and it is proven to be empirical.

Furthermore, the reliability of the instrument, in which the respondents take the role as the test participants, was calculated by using alpha Cronbach formula. Stratified Alpha and Construct Reliability (CR) were used to determine the reliability of the constructs. The component and total reliability coefficients were sought. The reliability coefficient results are shown in Table 12.

Based on Table 12, it can be concluded that the instrument with the test participants as the sample is stated to be reliable, and thus the instrument with 20 items can be valid and reliable if it is re-measured by using the same object because it has quite high reliability and feasibility value. It has the reliability coefficient value of at least > 0.7. It states that the used indicators already have adequate internal consistency reliability, meticulous in measuring and explaining the construct.

After that, some steps conducted during the research produced a final product, which was used as a questionnaire instrument developed from the Standard Operational Procedure (SOP) of the national examination. From the try-out of 25 items, only 20 items fulfilled the standard validity and reliability, so it was found using removal information. The result shows that it has validity value > 0.40 and reliability coefficient value > 0.70. Overall, the results show that the developed instrument is equal with the SOP of the national examination and it has been proven empirically that it is in a good category.

Conclusion and Suggestions

The research concludes that (1) based on the tested data with the test takers as the respondents, the instrument is valid, reliable, and qualified as the fit model; (2) the components in the instrument are the learning time, socialization, test materials, and examination room; and (3) the instrument has the validity value of > 0.40 and reliability value of > 0.70.

Based on the findings, some suggestions are proposed as follows: (1) the instruments developed in this model were only applied to the test takers as respondents. Thus, it is suggested to other researchers to develop it further, so that the evaluation instrument of national examination implementation of Package C at equality education will be better; and (2) the coverage of the objects in the evaluation instrument of the National Examination of Equivalency Education implementation of Package C is still too narrow, and therefore, other researchers need to add other components of the implementation so that the coverage can be more comprehensive.

Table 12. The reliability coefficient results of UNPK Evaluation

No	Component	*Cronbach's Alpha coefficiency*	*Stratified Alpha coefficiency*	CR	Remark
1	Learning time	0.831	-	0.886	Reliable
2	Socialization	0.809	-	0.865	Reliable
3	Examination material	0.734	-	0.826	Reliable
4	Examination room	0.797	-	0.878	Reliable
	Total	-	0.902	0.870	Reliable

References

Adow, I. M., Alio, A. A., & Thinguri, R. (2015). An assessment of the management of KCSE examination and its influence on irregularities among students: A case of secondary schools in Mandera County, Kenya. *Journal of Education and Practice*, *6*(28), 15–22.

Azwar, S. (2015). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. New York, NY: Plenum Press.

Child, D. (2006). *The essentials of factor analysis* (erd ed.). New York, NY: Continuum.

Christiansen, B., Howson, A. G., & Otte, M. (Eds.). (1986). *Perspectives on mathematics education*. Dordrecht: Springer Netherlands.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9. Retrieved from http://www.stat-soft.com/textbook/

Ghozali, I. (2005). *Structural equation modeling: Teori, konsep, dan aplikasi dengan program Lisrel 8.80*. Semarang: Badan Penerbit Universitas Diponegoro.

Government Regulation No. 19 of 2005, on National Education Standard (2005). Republic of Indonesia.

Hatfield, M. M., Edwards, N. T., Bitter, G. G., & Morrow, J. (2008). *Mathematics methods for elementary and middle school teachers*. Hoboken, NJ: John Wiley and Sons.

Hendryadi, & Suryani. (2014). *Structural equation modeling dengan Lisrel 8.80*. Jakarta: Kaukaba Dipantara.

Kahn, P., & Kyle, J. (Eds.). (2002). *Effective learning and teaching in mathematics and its applications*. London: Routledge.

Kartowagiran, B. (2008). Validasi dimensionalitas perangkat tes ujian akhir nasional SMP mata pelajaran matematika 2003-2006. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *12*(2), 177–195. https://doi.org/10.21831/pep.v12i2.1426

Kartowagiran, B. (2013). *Evaluasi dan pengembangan kurikulum*. A paper presented in Workshop Evaluasi Kurikulum STABN Raden Wijaya.

Law No. 20 of 2003 of Republic of Indonesia on National Education System (2003).

Mardapi, D. (2005). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekia.

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

Mardapi, D., & Kartowagiran, B. (2009). *Dampak ujian nasional*. A research report. Yogyakarta: Program Pascasarjana UNY.

Mathematical Sciences Education Board of National Research Council. (1993). *Measuring what counts: A conceptual guide for mathematics assessment*. Washington, DC: National Academy Press.

Mudjijanti, F. (2011). Pengaruh tes masuk berdasarkan nilai ujian nasional (UN) terhadap prestasi belajar siswa (Studi kasus di SMUK St. Bonaventura Madiun). *Widya Warta*, *35*(2), 19–31.

Raharjo, S. B. (2012). Evaluasi trend kualitas pendidikan di Indonesia. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*(2), 511–532. https://doi.org/10.21831/pep.v16i2.1129

Rossi, P. H., & Freeman, H. E. (1985). *Evaluation: A systematic approach*. Baverly Hills: Sage.

Setiadi, H. (2011). *Analisis kelemahan kompetensi siswa pada tingkat kabupaten/kota berdasarkan hasil UN rendah tahun 2011 di Kabupaten Alor*. A research report. Jakarta: Badan Penelitian dan Pengembangan Pusat Penilaian Pendidikan.

Sudjana, D. (2006). *Evaluasi program pendidikan luar sekolah untuk pendidikan nonformal dan pengembangan sumber daya manusia*. Bandung: Remaja Rosdakarya.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (4th ed.). Hillsdale: Erlbaum.

Weiss, C. H. (1972). *Evaluation research*. London: Prentice-Hall.

Worthen, B. R., & Sanders, J. R. (1981). *Educational evaluation: Theory and practice*. Worthington, OH: Charles A. Jones.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

REiD (Research and Evaluation in Education), 4(1), 2018, 70-78
Available online at: http://journal.uny.ac.id/index.php/reid

An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school

[*1]Mutiara Kusumawati; [2]Samsul Hadi
[1]Graduate School of Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[2]Faculty of Engineering, Universitas Negeri Yogyakarta
Karangmalang, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: mutia21@gmail.com

Submitted: 04 July 2018 | Revised: 31 October 2018 | Accepted: 06 November 2018

Abstract

The multiple-choice test is a common test format used in education. One of the purposes of this test is to evaluate the success of the learning process in a particular subject. Therefore, the efficiency of the evaluation depends on the quality of the test items used. This research was conducted in order to reveal the quality of the final mathematics examination items statistically. It was descriptive quantitative research employing two-parameter logistic (2pl) model of Item Response Theory (IRT). The data were obtained from the sample of 353 students established using the purposive sampling technique. This finding shows that 40% of the 35 items tested are very difficult, 60% are in the medium level, and there is no easy item. The most difficult material is the trigonometric calculation. The percentage of the item discrimination index is described as follows: 8.57% of the items are categorized as very low, 51.43% are categorized as low, 31.43% of the items have a medium item discrimination index value, 5.71% have a high item discrimination index value, and 2.86% of the items are categorized as very high. Moreover, the research found that all distractors functioned well. The highest information on ability $\theta = 0.4$ with information function value of 5.38 and SEM = 0.6. This test is suitable for students with the ability of $-1.42 < \theta < 2.65$.

Keywords: *difficulty index, discrimination index, efficiency distractor, mathematics examination*

Introduction

Evaluation refers to a systematic process to determine which instructional goals are achieved by students (Gronlund, 1982, pp. 5–6). The accomplishment of instructional objectives is done by a measurement. By measuring and evaluating, teachers can diagnose the strengths and weaknesses of their students and take action for their progress and improvement. If it is effective, measurement and evaluation can improve the learning situation. Without evaluation and measurement, it is impossible to know the needs and abilities of the students (Tshabalala & Ncube, 2014, p. 141).

Thus, the final examination of mathematics subject for class X is conducted in order to provide general information and illustration of student learning outcomes in the last semester. This is as a consideration of the teachers in particular and the schools in general that determines whether the students should keep on learning in the next grade or not. In addition, the results of these tests were used as an evaluation for educators in the implementation of the learning process. Therefore, the test items of mathematics final examination are one of the most important instruments in the learning process and must be well structured. A good final test item will give a good measurement result (Mardapi, 2012, p. 27). According to information from the drafting team of the Board of Muhammadiyah-School Principals Cooperation (or *Badan Kerjasama Kepala Sekolah Muhammadiyah*), gained through interviews during the survey,

ISSN 2460-6995

the materials of the mathematics subject test items for the 10th grade students used so far have not gone through a good stage in the preparation. Sulistiawan (2016, p. 10) finds that in the final examination of mathematics subject matter, there are 10% invalid items. This analysis was conducted to give some suggestions to the test developers of the mathematics final examination for the tenth grade students.

The test items are presented in the objective test form. The objective test is defined as a structured test that asks the participants to fill in one or two words, or to choose the correct answer from several options. An objective test consists of the problem/test item and list of alternative solutions. A list of alternative solutions can be in the form of words, numbers, symbols or phrases, and called key answers. Participants of the test are usually asked to read the problem/test items and alternative solutions list, and choose one right or best alternative (Gronlund, 1982, p. 135). The right option to each item is called an answer key, while the other options are called distractors. On the other hand, an essay test provides an opportunity for the test participants to organize, arrange, or answer freely from the questions given. For some instructional objectives, objective tests are considered to be more efficient in order to measure learners' skills at both low and high levels (Gronlund, 1982, pp. 5–6).

The selection of the appropriate test form is determined by the purpose of the test, the number of test participants, the range of test materials, and the characteristics of the subjects tested. The multiple-choice test and True or False test are particularly appropriate when the number of the test takers is large, the time for test correction is short, and the coverage of the tested material is numerous. The advantage of an objective multiple-choice test is that an answer sheet can be checked using the computer, so scoring objectivity can be assured (Saudi Commission for Health Specialties, 2011, p. 68).

Method

This research is a descriptive quantitative study conducted at two Muhammadiyah high schools in Yogyakarta, Indonesia. A population sampling technique in which the entire population was used as the sample (Sugiyono, 2001, p. 61) was used in this study. The participants of the study were 353 grade X students. The data of this study were the students' responses to the final examination consisting of 35 items of multiple-choice questions which have five options for each item. The data were analyzed quantitatively using Bilog.

The quality of the items on the 10th grade students' mathematics final examination was analyzed using modern Item Response Theory (IRT). This is a theory which employs the mathematical function to connect the opportunities of the correct answers to the students' ability. The IRT has a mathematical formula that connects the participants' characteristics and the item features in the model (Hambleton, Swaminathan, & Rogers, 1991, p. 12). The advantages of IRT include: the item statistics is not dependent on the group; the test scores obtained can illustrate individual capabilities; it does not require parallel tests to calculate the reliability coefficients; and it can provide the right measurement for each ability score.

There are three logistic models in the IRT, namely one-parameter logistic model (1pl), two-parameter logistic model (2pl), and three-parameter logistic model (3pl). These three models are suitable to respond to dichotomous forms (Hambleton et al., 1991, pp. 12–17). The three models are distinguished by the number of parameters which are used to describe the item characteristics of each logistic model or item parameters. The item parameters are item difficulty index (b), item discrimination index (a), and pseudo guessing (c). These three elements are so interrelated that it causes a function or response curve which is called the Item Characteristic Curve (ICC).

IRT can provide good results if the data used were in accordance with the selected logistic model. The selection of the logistic model is determined based on the p-value, which means that if the p-value is more than 0.05, then the item is said to fit the model (Retnawati, 2014, p. 25). The chosen logistic

model is determined from the logistic model that produces the most suitable items. The 2pl model produces the most suitable items so that this study employs the 2pl model. The 2pl model formula is as follows.

$$P_i(\theta) = \frac{e^{D.a_i\,(\theta - b_i)}}{1+e^{D.a_i\,(\theta - b_i)}} \text{ ..... i=1,2,3,...,n}$$

Notes:
$P_i(\theta)$: the chance that test participants can answer test item i correctly
a_i : item i discrimination index
b_i : item difficulty index, a point of ability where the possibility to answer correctly is 0.5.
θ : parameter of ability of test participants

The difficulty index is an opportunity to answer each item correctly at a certain level of ability. The percentage of the difficulty level used is elaborated as follows: an item of problem with a high difficulty level of 20%, 60% of items with medium difficulty level, and another 20% are items with low difficulty level (Arikunto, 1999, p. 210; Gajjar, Sharma, Kumar, & Rana, 2014, p. 19). The good index of difficulty levels is spread from -2.00 to +2.00 (Hambleton et al., 1991, p. 13). The closer the b, the easier the item is. The more the value of b approaching +2.00, the more difficult the item is. A good item is an item that is not too difficult or too easy. The overly easy question does not stimulate the students to increase the effort to solve the problem. Conversely, too difficult items will discourage the students from trying again because they are out of range (Daryanto, 2012, p. 197; Miller, Linn, & Gronlund, 2009, p. 21). In preparing the test item, the percentage of item difficulty level needs to be considered.

Discrimination Index (DI) is the effectiveness of an item measurement to distinguish learners with high ability from those with low ability. The discrimination index spreading from 0.01 to 0.34 is considered to be very low, 0.35 to 0.64 is low, 0.65 to 1.34 is moderate, 1.35 to 1.69 is high, and higher than 1.70 is very high (Baker, 2001, p. 34). The higher the DI, the more effective the item to distinguish learners with high ability from those with low ability.

The spread of alternative options is commonly used as the basis for the study of discrimination index. It is intended to find out whether the option is working or not**.** An option that is not the correct answer is called a distractor (Allen & Yen, 1979, p. 2). A distractor can be said to work well if it is at least selected by 5% of the test takers (Kolte, 2015, p. 321). If the distractor is selected by less than 5% of the respondents, then it is considered as a non-functioning distractor (NFD). The NFD must then be repaired or deleted, and replaced with another deceptive option (Haladyna & Downing, 1989, p. 55; Tarrant, Ware, & Mohammed, 2009, p. 3). The distractor efficiency is an indicator of whether the distractor on the item has been properly made or whether it has failed to perform its function as a distractor.

The Item Response Theory has several assumptions that need to be confirmed before modeling. These assumptions include: (1) unidimensional data, to show whether the model measures a single construct or not; and also (2) local independence, to show whether the response to each item is influenced by the response to another item (Hambleton et al., 1991, p. 19).

The unidimensionality assumption test of the data was conducted by employing SPSS application. The value of KMO shows 0.513 with the value of sig. = 0.000. Thus, the first assumption has been fulfilled. The local independence assumption is evident when the unidimensionality assumption of the participants' response data has been evident (Retnawati, 2014, p. 7).

Findings and Discussion

Findings

The number of multiple choice test items which are analyzed in this study is 35 items in 353 students. The average score achieved is 32.88 and the standard deviation is 13.81. The mean of discrimination index, difficulty level, and distractor efficiency are 0.71, 1.90, and 17.03 respectively. The standard deviations on each parameter are 0.40, 1.56, and 6.63, respectively (see Table 1).

Table 1. Mean and standard deviation of item parameters

Parameter	Mean	Standard Deviation (SD)
Discrimination index (DI)	0.71	0.40
Difficulty index (*p* value)	1.90	1.56
Distractor efficiency (DE)	17.03	6.63

The discrimination index in general is low; 51.43% (18 items) is low and 8.57% (3 items) is very low. Only 2.86% (1 item) has a very high discrimination index and 5.71% (2 items) have a high discrimination index value. Moreover, 31.43% (11 items) have a moderate discrimination value.

The distribution of the item difficulty level is generally acceptable. The acceptance threshold of problem difficulty level is between -2.00 to 2.00 (Hambleton et al., 1991, p. 13). A number of 21 items (60.00%) can be accepted with the threshold value (b) ranging from -0.43 to 1.94, while the other 14 items (40.00%) have the difficulty level of more than 2.00, meaning that these items are very difficult. The difficult item has a threshold value (b) ranging from 2.08 to 4.95.

There are 140 distractors in these tests, in which each item has 4 distractors. All distractors (100%) functioned well because each of them was chosen by more than 5% of the test participants. This means that there is no Non-Functional Distractor (NFD) (Table 2).

Discussion

A multiple choice test is one of the efficient evaluation tools. However, this efficiency is highly dependent on the quality of multiple choices that can be judged on the basis of the item analysis. The index of difficulty and discrimination levels is one of the steps to check whether multiple choice tests are well established or not. Another step used for further analysis is the functionality of the distractors.

Difficulty Index

It should be noted that the item difficulty level in a test is divided into the following precentages: a problem item with a high difficulty level of 20%, 60% items with medium difficulty level, and another 20% are items with low difficulty level (Arikunto, 1999, p. 210). However, the percentage of difficult, medium, and easy items is not balanced. A total of 21 items (60.00%) can be accepted with the threshold value (b) from -0.43 to 1.94. The other 13 items (37.14%) have a difficulty level of more than 2.00, meaning that these items are difficult. Difficult items have a threshold value (b) ranging from 2.08 to 4.95. In this case, there is no easy item. The percentage distribution of the difficulty level of the item is illustrated in Figure 1. Thus, this percentage does not reflect a good distribution of test item difficulty level. A good test item must have a balanced percentage of test items with high level of difficulty and low level of difficulty, at 20% each. Of 35 items, 14 items are categorized as having a high level of difficulty and there is no item with low level of difficulty.

Table 2. Difficulty index interpretation, discrimination index, and distractor efficiency

Parameter	Item statistic	Interpretation	Total item (%)
Difficulty index (*p* value)	<-2.00	Very easy	0 (0%)
	-2.00 to 2.00	Moderate	21 (60.00%)
	>2.00	Very difficult	14 (40.00%)
Discrimination index (DI)	>1.7	Very high	1 (2.86%)
	1.35 – 1.69	High	2 (5.71%)
	0.65 – 1.34	Moderate	8 (22.86%)
	0.35 – 0.64	Low	18 (51.43%)
	0.01 – 0.34	Very low	3 (8.57%)
Distractor Efficiency (DE)	>0.05	Well functioned	140 (100%)
	<0.05	Unwell functioned	0 (0.00%)

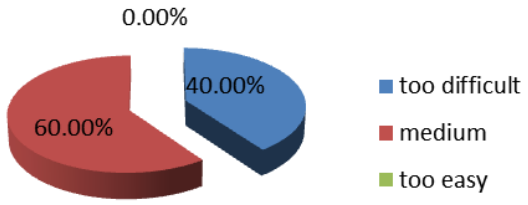

Figure 1. Distribution percentage of difficulty index

The test items should be sequenced from the easiest to the most difficult. Thus, the most difficult items should be placed at the last part of the test. However, when the subject matter or subject changes, the item's difficulty index begins with the easiest. Figure 2 shows the difficulty index of each item. The simplest item is of the threshold value of -0.78 and is placed in the earliest part of the test. Students' lack of success in taking the test can also be caused by the wrong order of items. The unfavorable placement of difficult items affects the students' result (Debeer & Janssen, 2013, p. 177). Therefore, it should be placed at the end of the test with regard to the materials being tested. This should be a concern for the test developers.

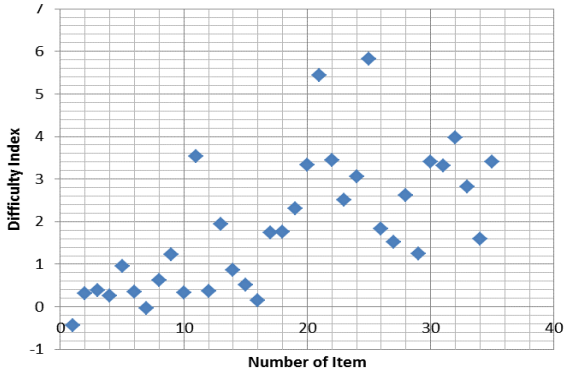

Figure 2. Scatter plot of difficulty index

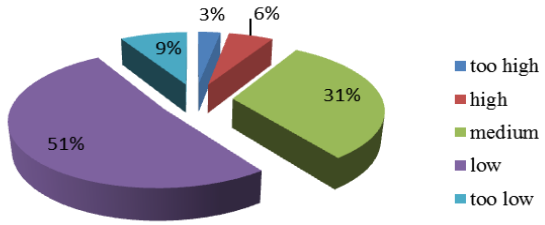

Figure 3. Distribution percentage of discrimination index

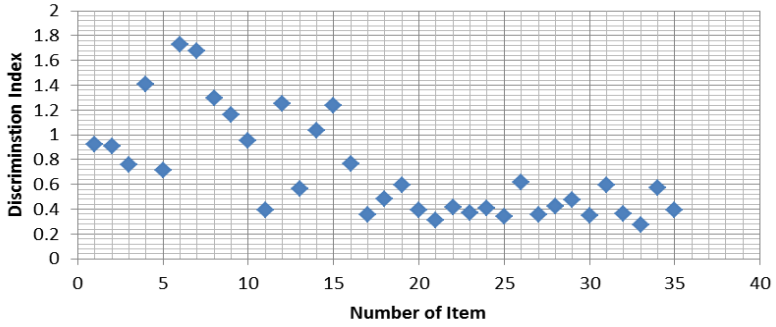

Figure 4. Scatter plot of discrimination index

Discrimination Index

The discrimination index on this test item has been distributed evenly although the majority of the discrimination index is low (see Figure 3). The discrimination index is important to know the difference between high and low ability groups. This instrument needs to be improved so that it has a better discrimination index. Figure 4 shows the discrimination index of each item.

Distractor Efficiency

Analyzing the distractor is done to determine the usefulness of each individual distractor on each item. In this study, 100% of the distractors (140 distractors) functioned well. This means that there is no non-functional distractor. If many students do not choose a particular distractor simultaneously, it is likely that these distractors do not make sense. Thus, the distractor does not effectively distract students. The non-functional distractor (NFD) will reduce the functionality of the distractor itself. The more NFD in the test, the easier the test items will be. Conversely, the fewer NFD in the test, the higher the items' level of difficulty will be. On the other hand, the functioning of the distractors themselves will get better (Allen & Yen, 1979, p. 2; Kolte, 2015, p. 321). The non-functionality of distractors is important in the preparation of a good multiple-choice test.

Information Function & Standard Error Measurement (SEM)

The value of the information function of test items denotes the strength or contribution of each item to revealing the measured latent trait. The information function with the two-parameter model depends on the level of difficulty and also the discrimination index of the item. The greater the value of information, the lower the value of SEM. Based on the results of the analysis, the highest information on the ability $\theta = 0.4$ with the information function value of 5.38 and SEM = 0.6. This is the maximum value of the information function. The two-point intersection between the information function and SEM is at $\theta = -1.42$ and $\theta = 2.65$. This shows that the test is suitable for students with the ability of $-1.42 < \theta < 2.65$.

There are many things which need to be considered in preparing objective form tests, including: (1) each question item must contain only one correct answer, (2) all distractors must be reasonable, (3) the length of the alternative answers should not give a clue to the correct answer, and (4) the correct answer should appear in each alternative position roughly in the same amount, but randomly (Gronlund, 1982, pp. 189–199). On the other hand, teachers must have the skills in preparing a test so that it is prepared to be of good quality. Educators should be aware that: (1) they should master the subject they teach, (2) they should have the skills to analyze test items, and (3) they should be able to help students make use of the information in the context of formulating educational policies. Thus, teachers' competence is not focused on the mastery of the material only. Their skills in analyzing the test results are also very important (Brookhart, 2011, p. 3). In addition, educators should also be able to arrange items that really matter in accordance with the subject matter to be tested.

In terms of the item's level of difficulty, the distribution of the items is in the levels of easy, medium, or difficult. This indicates the mastery of the material by the students. In terms of students' skill in the material being tested, there is no easy material. The material which has a medium level of difficulty includes determining the result of the subtraction operation on the function, determining the inverse of the linear function, determining the trigonometric ratio on the right triangle, determining the function of the composition consisting of two functions, solving the problem involving the addition operation of the function, determining the composition function of inverse function, determining the value of a composition function consisting of two functions, determining the function if the composition function is identified, determining the result of the mapping on the function, determining the inverse value of the composition function, determining the inverse of the fractional function, determining the function value if the composition function is identified,

determining the composition of the three functions, determining the inverse value of the fractional function, determining the diagonal length of the parallelogram, determining the angle of the triangle if two sides and one angle are known, determining the area of the hexagon if the radius of the outer circle is known, expressing the angle into degrees, determining the sides of a triangle if two angles and one side are identified, and determining the function if the composition is known.

On the other hand, the items which have a high-level of difficulty in the tenth grade Final Examination include determining trigonometric values in various quadrants and related angles, determining the pilot's visibility of the cruise ship if the plane's height and depth angle are identified, determining the cos angle if it is known to the three sides, determining the area of the triangle if two sides and an angle are known, analyzing the identity of trigonometry, determining the angle if both sides and extent are known, determining trigonometric values in various quadrants and related angles, determining the circumference of the octagon if the radius of the outer circle is found, determining the graphic equation, drawing the distance between the lower end of the staircase and the wall if the length of the ladder and the angle formed between the stairs and the floor is known, determining the inverse of the composition function, determining the area of the triangle if the length of the three sides is known, determining the trigonometric value in the various quadrants and correlation angles, and analyzing the identity of trigonometry.

In the tenth grade mathematics final examination test items, the easiest material is determining the result of the subtraction operation on the function. This is indicated by the smallest treshold value of -0.43. The difficult materials in the test include determining trigonometric values in various quadrants and related angles, determining the cos angle if the three sides are known, determining the angle if both side and the extent are known, and determining the trigonometric value in various quadrants and angle-related corners.

One of the difficult items found is determining the value from $\sec\frac{2\pi}{3} \times \tan\frac{5\pi}{6} \times \sin\frac{3\pi}{2}$ with the answer choice A. $-2\sqrt{3}$, B. $-\frac{2\sqrt{3}}{3}$, C. $\frac{2\sqrt{3}}{3}$, D. $\sqrt{3}$, and E. $2\sqrt{3}$. This problem requires several stages of completion. Each factor must be determined in advance. $\sec\frac{2\pi}{3}$ are -2, $\tan\frac{5\pi}{6}$ are $-\frac{\sqrt{3}}{3}$, and $\sin\frac{3\pi}{2}$ are -1 so that the result of these values is $-\frac{2\sqrt{3}}{3}$ (B). This answer option was only chosen by 78 students (22%). The highest number of students' answers was in choice C, i.e. 125 students (35%). Alternatives A and E were chosen by 43 students (12%), 63 students (18%) chose Alternative D and the other students did not choose any option.

This finding is consistent with the finding of the research conducted in Riau Province, that calculating trigonometric ratio with sinus, cosine, and tangent formulas is a difficult subject in high school maths (Aisyah, 2013, p. 153). These results provide an illustration that the basic competence of high school mathematics has not been achieved in terms of learning indicators. Therefore, the need for evaluation and improvement in the learning process of mathematics is urgent to improve the students' learning achievement. The results of this study can also be an input to improve the teachers' competence, especially in the teaching of these difficult materials.

The same findings are also expressed by Wongapiwatkul, Laosinchai, and Panijpan (2011, p. 54) that studying trigonometry is difficult for students and the difficulties are caused by many interconnected things. Students may first learn trigonometric function, or learn it because they have difficulty in reasoning in trigonometry. In addition, manipulating trigonometric calculations is not the same as manipulating the algebra operation.

Conclusion

The level of difficulty of the tenth grade mathematics final examination test items is in the medium category. Overall, the test items have a very good distractor efficiency. Of all given distractors, they were selected by over 5% of the test takers. The discrimination index in this test is not good because it is only at medium, low, and very low levels. In this test, it is known that the difficult materials in this test are: (1) determining the distance of an object using the concept of depression angle,

(2) determining the area of a triangle if two sides and one slant angle are known, (3) analyzing the trigonometric identity, (4) determining an angle if the length of the two sides of the triangle and its width are known, (5) determining the equation of a trigonometric graph if the graph is known, (6) determining the trigonometric values in various quadrants and related angles, (7) determining the circumference of an octagon if the diameter of the outer circle is known using the trigonometric formula, and (8) determining the area of a triangle if the length of the three sides is measured.

This finding is consistent with Aisyah's research finding that trigonometric material is difficult for learners (Aisyah, 2013, p. 153). She states that determining the equation for trigonometric charts if the graph images are known and determining trigonometric values in various quadrants and related angles are difficult material in mathematics. Educators must be concerned in this area to explore the material, and improve their teaching methods and strategies. Likewise, students should pay attention to the materials better. According to Keoviphone and Wibowo (2015, p. 8), the more systematic educators plan their learning materials, the more likely they will succeed.

References

Aisyah, U. (2013). *Pengembangan perangkat pembelajaran kompetensi sulit matematika SMA di Riau. Master Thesis*. Universitas Negeri Yogyakarta, Yogyakarta.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Montery, CA: Cole Publishing.

Arikunto, S. (1999). *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Daryanto, M. (2012). *Evaluasi pendidikan*. Jakarta: Rineka Cipta.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality Multiple Choice Questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, *39*(1), 17–20. https://doi.org/10.4103/0970-0218.126347

Gronlund, N. E. (1982). *Measurement and evaluation in teaching* (4th ed.). Cliffs, NY: Macmillan.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4

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

Keoviphone, C., & Wibowo, U. B. (2015). Factors discouraging students from schooling: A case study at Junior Secondary School in Laos. *REiD (Research and Evaluation in Education)*, *1*(1), 1–12. https://doi.org/10.21831/reid.v1i1.4894

Kolte, V. (2015). Item analysis of multiple choice questions in physiology examination. *Indian Journal of Basic and Applied Medical Research*, *4*(4), 320–326.

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

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in*

teaching (10th ed.). Upper Saddle River, NJ: Pearson.

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

Saudi Commission for Health Specialties. (2011). *Item writing manual for multiple-choice questions*.

Sugiyono. (2001). *Metode penelitian bisnis*. Bandung: Alfabeta.

Sulistiawan, C. H. (2016). Kualitas soal ujian sekolah matematika program IPA dan kontribusinya terhadap hasil ujian nasional. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(1), 1–10. https://doi.org/10.21831/pep.v20i1.7516

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, *9*(40), 1–8. https://doi.org/10.1186/1472-6920-9-40

Tshabalala, T., & Ncube, A. C. (2014). The effectiveness of measurement and evaluation in Zimbabwean primary schools: Teachers and heads' perceptions. *International Journal of Innovation and Applied Studies*, *8*(1), 141–148.

Wongapiwatkul, P., Laosinchai, P., & Panijpan, B. (2011). Enhancing conceptual understanding of trigonometry using earth geometry and the great circle. *Australian Senior Mathematics Journal*, *25*(1), 54–63.

REiD (Research and Evaluation in Education), 4(1), 2018, 79-93
Available online at: http://journal.uny.ac.id/index.php/reid

Continuing professional development (CPD) for junior high school mathematics teachers: An evaluation study

[*1]Pika Merliza; [2]Heri Retnawati
[1,2]Department of Mathematics Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
[*]Corresponding Author. E-mail: pikamerlizasoemali@gmail.com

Submitted: 06 March 2018 | Revised: 01 August 2018 | Accepted: 15 November 2018

Abstract

Responding to the importance of conducting evaluation on continuing professional development program for teachers, this study is aimed at describing the implementation and difficulty of Continuing Professional Development (CPD) of mathematics teachers of Junior High School (JHS) in Bandar Lampung, Indonesia. This research used a descriptive approach employing a quantitative-qualitative method with sequential explanatory strategy. The population of the research was 181 junior high school mathematics teachers who have already become civil servants. The samples were 63 teachers for quantitative research selected using stratified random sampling and proportional random sampling technique, while eight teachers for qualitative research were selected using purposive sampling technique. These eight teachers were selected because they were the only teachers handling the CPD program. The data were collected through a test, questionnaires, checklist sheet, study document, and interview. Data analysis was conducted using categorized performance trends, divided into five groups: Very Good/Difficult, Good/Difficult, Fair, Poor/Easy, and Very Poor/Very Easy. The data were analyzed using descriptive technique; the quantitative study analysis was performed by mean and standard deviation, whereas, the qualitative data analysis was obtained by data reduction, data display, and conclusion technique. The research results show that the majority of teachers' CPD implementation is very poor, meanwhile, the difficulty of the engagement of CPD is categorized as fairly difficult.

Keywords: *junior high school mathematics teachers, continuing professional development (CPD), teaching experience*

Introduction

As a developing country, Indonesia aspires to improve becoming a developed country which is independent, unified, sovereign, just, and prosperous (Preamble of the 1945 Constitution of the Republic of Indonesia). Various attempts are made to embody the ideals of the nation, one of which through the efforts of alleviation of poverty and unemployment, which became conspicuous case of developing countries. The high number of poverty and unemployment will impact badly on various aspects of life, such as increasing numbers of violence, theft, robbery, depression, political instability, and many others. According to Yacoub (2012, p. 178), if a community has job and earnings (instead of unemployment), and then the earnings are expected to meet the necessities of life, so it is stated that they are not poor. It can be inferred that by the low unemployment number, then the number of poverty is also low.

One of the factors contributing to the high number of unemployment is the low quality of human resources who are able to compete in both national and global scope. Nationally and globally competitive human resources provide opportunities to get a job to fulfill their life necessities and decrease the level of poverty of the nation. The low quality of the nation's human resources is a product of the poor quality of education. This is due

ISSN 2460-6995

to the fact that education has a major influence for various aspects of sustainable life development and the supporting factors contributing to the sustainability and peace, giving a direct influence to decrease poverty, and also promoting health, gender, and sustainable environment (UNESCO, 2014, p. 25). The importance of the influence of education to human life becomes one of the factors underlying UNESCO to continue stating the idea of lifelong learning which is started since 1972 (Tuijnman & Boström, 2002, p. 95).

The enhancement of abilities, skills, and attitudes determines the quality of human resources of a nation. Thus, it is the responsibility of every individual to become a lifelong learner, to learning developing themselves, to continue and enhance the competence and expertise along with the development of science and technology. This responsibility is valid to everyone in profesion, including teachers. Teachers are demanded to conduct professional development throughout their career related to the role and responsibility (Gray, 2005, p. 5).

Based on a research related to the implementation of CPD for teachers, Nuraeni and Retnawati (2016, pp. 137–138) reveal that the effectiveness of subject-matter teachers forum (MGMP) still belongs to the low category. Nuraeni and Retnawati (2016) suggest that the activities of the MGMP - that have been established in each regency and province allegedly - have not functioned optimally to facilitate the mathematics teachers to develop themselves.

Furthermore, in regard to the performance of post-certification teachers, the facts show that not all certified teachers in Indonesia have good competences and performances (World Bank in Jalal et al., 2009, p. 7). This is in line with a research conducted by Abubakar (2015, p. 116) about the impact of certification on Madrasah Aliyah teachers' competence in Kendari, South East Sulawesi, Indonesia, which states that teachers' certification has not had a positive impact on their competence improvement, either in their subject area or educational unit.

It is proven by the findings of a research of Kardiyem (2013, p. 17) that the overall performance of certified teachers in senior vocational schools in Grobogan Regency, East Java, Indonesia is in 'not good' category. Various obstacles are faced by the teachers, including low motivation achievement, limited time, lack of knowledge, and perceptions on the government regulations. In addition, the teachers' level of competence and skills before and after the certification is still the same. The teachers are less trying to improve their competence and tend to perform the same as before getting the certificate. As reported by Nuraeni and Retnawati (2016, p. 130) in their research on teacher performance in professional development in Wonosobo Regency, the teachers are still categorized as very poor in professional development, and certified teachers have less awareness in their professional development. Further, Fahmi, Maulana, and Yusuf (2011, p. 15) emphasize that teacher certification was expected to improve teachers' quality, however, in fact, it does not contribute positively to the improvement of the students' learning process.

This condition is contradictory to the Law No. 14 Year 2005 of Republic of Indonesia about Teachers and Lecturers, which states that in performing professional duties, teachers are obligated to improve and develop academic qualifications and competencies in a sustainable manner in line with the development of science and technology. All teachers must have professionalism in their profession; teachers must be mastering the competencies needed to achieve educational aim. Competence is an important component to support the performance of teachers in performing their duties and roles.

According to Hamilton-Ekeke (2013, p. 15), teacher competence is the ability of a teacher to help learners to reach higher levels of learning. Competence requires teachers to carry out professional responsibilities, hence the effectiveness of the implementation of teacher role as a learning agent depends on the teacher's level of competence. The teacher's level of competence is related to the professional and pedagogical knowledge. In Indonesia, teachers' professional competency standard consists of professional, pedagogic,

social, and personality competence (Regulation of the Minister of National Education No. 19 of 2005 on National Education Standard). The four competencies must be mastered by all of professional teachers in Indonesia, including mathematics teachers.

Mathematics is one of the important subjects that equips learners in facing a fully-competitive life. Alnoor and Yuanxiang (2000, p. 1) explain that mathematics is a necessary tool in the field of science and technology, since it aims not only to teach arithmetic, but also provides opportunities for learners to become scientists - exploring concepts related to everyday life. The purpose of mathematics education which requires logical, analytical, systematic, critical, and creative thinking as well as cooperative ability (Regulation of the Minister of Education and Culture No. 20 of 2016 on the competence standard of primary and secondary education graduates) are very useful in preparing highly competitive generation. It means that in order to reach a high quality mathematics learning, the competence of qualified teachers is required. It is essential for teachers to continue improving their competence to support their carriers. This is not only a demand for in-service mathematics teachers, but also for pre-service teachers.

According to the data of World Bank (2010, p. 18), 'Mathematics teachers have scored poorly on competency exams, raising concerns about the quality of their instruction'. Further results of teacher competency test indicate a significant impact on the learning practices seen from the learners' achievement (Ünal, Demir, & Kiliç, 2011, p. 3252). For example, data of National Examination (or *Ujian Nasional* - UN) result of junior high school in Bandar Lampung in the academic year of 2015/2016 released by National Education Standards Board (Badan Standar Nasional Pendidikan, 2015) show that the average score of mathematics is 53.99. This score is lower than other subjects such as Indonesian, English, and science which are 71.84, 64.28, and 63.66. This is assumed that mathematics' low average score is due to its non-routine patterns of questions with increasing challenges provided for the students which are not able to solve only through calculation, but also require higher order thinking skill, involvement of reasoning, analyzing, synthesizing, and evaluating. By this situation, the tasks and roles of mathematics teacher of junior high school are great in establishing the learning situations. Teachers are the determinant factor of students' learning experiences in the classroom, who prepare students to have higher order thinking skill (UNICEF, 2007, p. 93).

Furthermore, teachers are the central point to significantly improve their own competences. Teachers' knowledge and skills are the factors that influence the success of classroom learning. In the mid of rapid technological advances, as a professional, and due to the demand for high standard of education, teachers must continue learning (in-services learning) in order to improve their competence.

Continuing Professional Development (CPD) is a continuing learning for teachers who are the main vehicle in the effort to bring the desired changes related to the success of the learners (Ministry of National Education of Republic of Indonesia, 2010, p. 9). In the case of teaching, development which can be made is in-service training. The content of CPD for mathematics teacher is believed to revitalize the teacher's skills in designing the teaching and learning process, increasing enthusiasm on the instruction, and also help to maintain their scientific knowledge (Joubert, Back, De Geest, Hirst, & Sutherland, 2010, p. 1765).

The concept of CPD for teachers has been implemented in many various countries around the world. In Finland, the concept of CPD for teacher is based on the idea of 'long-life learning'. The government has adopted the concept of CPD since the education of prospective teachers at university level. The implementation of CPD agenda for teachers during 7 days in a year should be involved in inter-school teacher training. In addition, as another form of CPD implementation for teachers, the Finnish government requires all teachers to pursue higher education to attain at least a master degree as a minimum standard of teachers' education level (Layne, 2016, p. 9). In the UK, the types of CPD include:

(1) workshops held at school or outside the school, (2) certified courses, (3) courses held at the university, (4) teacher collaboration activities, (5) conferences, (6) guiding, training, or observing fellow teachers in teaching (*lesson study*), (7) joining committees, (8) teacher learning communities, and (9) self-learning (Opfer & Pedder, 2010, p. 241).

In Indonesia, CPD for teachers has been set forth in the Regulation of the Minister for the State Apparatus Empowerment and Bureaucracy Reform No. 16 of 2009, enforced in 2013, that teacher who has a certificate of educator is required to implement CPD with the calculated credit number. The credit number is needed by teachers to achieve higher degree which is automatically influential to their salary. CPD in Indonesia consists of: (1) subsection of self-development, (2) scientific publications, and (3) innovative works.

CPD activities provide benefits for teachers to improve their knowledge, skills, and competencies according to their professional standards. Adey, *et al.* (Wermke, 2011, p. 669) state that there are three important facts which become the basis of CPD for teachers: (1) CPD activities improve teachers' new understanding of learning and their belief; (2) CPD for teachers is influential to the classroom instructional practice in the form of feedback on CPD activities; (3) CPD for teachers is based on intuitive knowledge that influences the teachers' attitude intuition, one of which is influenced through training process. Thus, CPD can provide benefits for teachers' beliefs as well as classroom instructional practices.

In addition, Desimone (2009, pp. 183–184) states that the benefits of professional development for teachers are; (1) CPD can improve the teachers' knowledge and skills and/or change their attitudes and beliefs; (2) knowledge, skills, attitudes, and beliefs will be influential to improve the content knowledge that gives impact on pedagogical knowledge in the teaching practice; (3) CPD can cause changes in the way of teaching that has a positive impact towards the improvement of the learners' learning outcomes. However, based on the research results on CPD activities in some areas of Indonesia, the implementation of CPD is still not encouraging. Based on the results of Nuraeni and Retnawati (2016, pp. 137–138), the professional development of mathematics teachers of vocational senior high school in Wonosobo Regency is categorized in 'poor' category.

This situation is in line with the finding of Kartowagiran (2011, p. 463) stating that post-certification teachers' performance in professional development is still unsatisfying. Moreover, based on the results of a research conducted by Noorjannah (2015, p. 107), there is fraud conducted by teachers in the CPD, especially in the paper writing; 70% of the teachers employ writing services, action research, promotion, or conduct certain activities such as certification. According to the research findings by Aina, Bambang, Retni, Afreni, and Sadikin (2015, p. 31), the number of scientific papers published by teachers is low due to their difficulties in writing. The difficulties are related to the teaching hours, writing ethic and techniques, and unaccustomed experience of expressing ideas in a writing project.

According to Supriyanto (2015, p. 111), scientific writing in the CPD produced by teachers is basically not conducted periodically, and the policy of writing scientific paper for promotion is not responded positively. Furthermore, based on the research findings of Wibowo and Jailani (2014), the CPD of mathematics teachers of junior high school in Wonosobo based on its understanding, implementation, and difficulties are categorized in 'medium', 'low', and 'very low' categories. These categorizations may be caused by the difficulties faced by teachers in the implementation of CPD. Qablan, Mansour, Alshamrani, and Aldahmash (2015, p. 627) mention the obstacles in implementing CPD, including excessive workload, teaching hours, location of the activities, time of teaching, personal circumstances, funding, and the activities.

Furthermore, based on the data gained from interviews to mathematics teacher of junior high school in Bandar Lampung regarding the CPD activities, it is found that teachers do not focus to do their duty to teach mathematics in the class. It means that

teachers have tried to conduct CPD but they find various difficulties. Thus, this research aims to find out the implementation of CPD and the difficulties faced by mathematics teachers of junior high school in Bandar Lampung in conducting CPD.

Method

This research used descriptive approach employing quantitative-qualitative method (or mixed methods). The study was conducted at junior high schools in Bandar Lampung, Indonesia. Data collection was conducted from February to March 2017, through direct meeting of the researchers with respondents at their respective schools and mathematics subject-teachers forum (or *Musyawarah Guru Mata Pelajaran* - MGMP) for junior high school (JHS) teachers in Bandar Lampung.

The population in this research was 181 mathematics teachers of junior high schools in Bandar Lampung in the academic year of 2016/2017. Samples were identified using stratified random sampling procedure based on the teachers' teaching experiences, then subsequently selected through proportional random sampling technique. The samples were 63 mathematics teachers, whereas for qualitative research, eight respondents were chosen using purposive sampling technique.

Research Procedure

The research used mixed methods research design in the form of sequential explanatory design, in which the process of data collection were not done at the same time. The researchers collected and analyzed the quantitative data firstly, then the analyzed data were used as the basis for collecting and analyzing the qualitative data. The flow is presented in Figure 1.

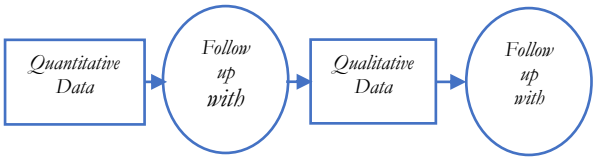

Figure 1. Research setting of sequential explanatory design (Creswell & Clark, 2011, p. 68)

The results of analysis from both studies were combined and compared, so that it was known which qualitative data were appropriate, expanded, or even aborted the results of quantitative data. Furthermore, the results of the both data were presented in a table to draw conclusions.

Data, Instruments, and Data Collection Techniques

The instruments which were employed in this study consist of checklist and document study which were used to measure the implementation of CPD for teachers, consisting of: (1) 27 items of CPD implementation on self-development sub-section; (2) 23 items of CPD implementation on scientific publication sub-section; and (3) 21 items of CPD implementation on innovative works sub-section. Document study was developed from the checklist to check the physical evidence related to the involvement of CPD activities in each aspect.

Furthermore, the instruments were in the form of 15 items of questionnaire about the difficulties aspect with a semantic differential scoring scale with the intervals of 1-7. The items were equipped with very easy to very difficult choices and were completed by 15 items of open-ended questions related to the efforts in overcoming the difficulties. In addition, an instrument in the form of an interview guidance was used to measure all aspects of CPD. The validity of the instruments was in the form of content validity: face and logical validity, and was done by two expert judgments. The reliability of the instrument difficulty was 0.964 and the SEM was 15.263, which means that it was in a very good category of reliability.

Data Analysis Technique

The quantitative data analysis was presented in the table based on the tendency of the respondents' answer on one of the criteria in each sub-variable. In categorizing the percentage of quantitative descriptive analysis, the researchers employed the categorization adapted from Widoyoko (2013, p. 238), which is presented in Table 1.

Table 1. Assessment categories

Formulas	Category
X >Mi + 1,8 Sbi	Very Good
Mi+ 0,6 SBi< X ≤ Mi + 1,8 SBi	Good
Mi - 0,6 SBi< X ≤ Mi + 0,6 SBi	Fair
Mi - 1,8 SBi< X ≤ Mi - 0,6 Sbi	Poor
X ≤ (Mi - 1,8 SBi)	Very Poor

Note:
Mi = ideal mean score
SBi = ideal standard deviation
X = score of the respondents or actual score

Qualitative Descriptive Analysis Technique

The qualitative data which were obtained through interview and document study were analyzed by interactive model, and then the result of the data analysis was processed through the following flow: data reduction, data display, conclusion drawing, and verification (Miles, Huberman, & Saldaña, 2014, pp. 12–13). The data analysis techniques are presented in Figure 2.

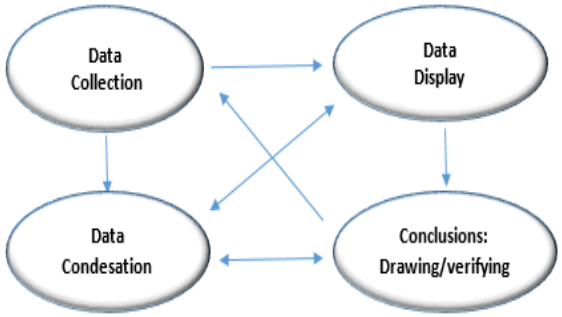

Figure 2. Interactive model analysis scheme

At the data reduction stage, simplifying the activities and selecting the key points were performed to create a core summary without changing the message. Furthermore, in the presentation stage, the data processed were briefly displayed on a table to make it easy to understand the unit related to the steps of the research. The last stage is the conclusion stage, in which the researchers collected the data in order to draw conclusion based on the tendency of the same or similar data categorization. This conclusion stage is temporary, so that the researchers need to verify the conclusion which was drawn with the data categorization and the quantitative data to make it more credible.

Findings and Discussion

Findings

Implementation of CPD to Mathematics Teachers of Junior High School

Based on the results of CPD checklist sheet of mathematics teachers which includes self-development, scientific publication, and innovative works aspects, the implementation of CPD activities has the actual mean score (X) of 48.42 (very poor category), the ideal score (Mi) of 162.5, and the ideal standard deviation of 54.17 with maximum score 325 and minimum score of 65, as presented in Table 2.

Table 2. Score of CPD implementation to mathematics teacher of junior high school

Category	Total	Percentage
Very Good	0	0%
Good	0	0%
Fairly Good	0	0%
Less Good	3	5%
Very Poor	60	95%
Total	**63**	**100%**

Table 2 shows that the performance of 95% of junior high school mathematics teachers in Bandar Lampung in implementing CPD activities is in 'very poor' category, meaning that the majority of the teachers are still less involved in the CPD activities, either in person or inside the mathematics teachers' learning community/institutions, both in formal and non-formal learning.

Furthermore, the overall results of the implementation of CPD activities are obtained based on the details of each aspect. In the aspects of self-development, the actual score (X) is 33.32 (very poor category), the ideal score (Mi) is 81, the ideal standard deviation is 18.00 with the maximum score of 135, and the minimum score of 27.

In addition, the result of CPD implementation of junior high school mathematics teachers in Bandar Lampung seen from scientific publication aspect, the actual score (X) is 6.93 (very poor category). The mean ideal score is 69, the ideal standard deviation is 15.33 with the maximum score of 115 and the minimum score of 23. Besides, in innovative

works aspect, the actual score (X) is 8.68 (very poor category), the ideal score (Mi) is 45, the ideal standard deviation is 10.00 with the maximum score of 75 and the minimum score of 15. The detail score of each aspect is presented in Table 3.

Difficulties of CPD Implementaion to JHS Mathematics Teachers

Based on the teachers' responses to the questionnaire, the actual score (X) is 505.39 (quite difficult category), the ideal score (Mi) is 452, the ideal standard deviation is 113.00 with maximum score of 113 and minimum score of 452. To make it clear, the difficulty criteria are presented in Table 4.

Based on Table 4, 70% of JHS mathematics teachers have difficulties in conducting CPD activities, while the 30% have no difficulties to be involved in CPD. It means that more than half respondents have difficulties in actively participating in the activities on all aspects. In more detail, 8% of the teachers are in difficult category, 62% of teachers are in the quite difficult category, and 30% of teachers have no difficulties to conduct CPD. The overall data of the difficulties category for the details of each CPD activity are presented in Table 5.

Based on Table 5, less than 30% of the teachers stated that they have no difficulties in engaging each aspect of CPD. In the aspect of self-development, most of the respondents (30%) are included in the fair category, whereas in the aspect of scientific publication, they mostly are in the difficult category (52%), and in the aspect of innovative works, they are in the very difficult category (40%). Meanwhile, the percentage comparison of the number of teachers who have difficulties in conducting CPD is presented in Figure 3.

Table 3. CPD implementation score of self-development, scientific publications, and innovative works

Category	Self-Development		Scientific Publications		Innovative Works	
	Total	(%)	Total	(%)	Total	(%)
Very Good	0	0%	0	0%	0	0%
Good	0	0%	0	0%	0	0%
Fair	0	0%	0	0%	0	0%
Less Good	2	3%	0	0%	0	0%
Very Poor	61	97%	63	100%	63	100%
Total	**63**	**100%**	**63**	**100%**	**63**	**100%**

Table 4. Score of CPD difficulties of mathematics teachers

Category	Total	Percentage
Very Difficult	0	0%
Difficult	5	8%
Fair	39	62%
Easy	19	30%
Very Easy	0	0%
Total	**63**	**100%**

Table 5. Score of difficulties of CPD implementation on self-development, scientific publishing and innovative works

Category	Self-Development		Scientific Publications		Innovative Works	
	Total	(%)	Total	(%)	Total	(%)
Very Difficult	8	13%	0	0%	25	40%
Difficult	8	13%	33	52%	21	33%
Fair	34	54%	27	43%	15	24%
Easy	13	20%	3	5%	2	3%
Very Easy	0	0%	0	0%	0	0%
Total	**63**	**100%**	**63**	**100%**	**63**	**100%**

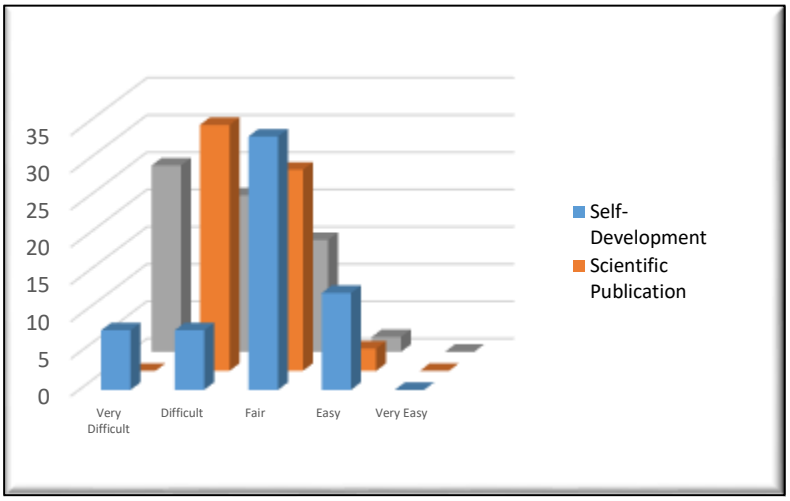

Figure 3. Diagram of comparison of questionnaire results of CPD implementation to mathematics teachers in Bandar Lampung

Based on Figure 3, the mean score of the difficulty of CPD implementation to the majority of mathematics teachers in Bandar Lampung on doing scientific publication is in difficult category. Besides, the difficulty faced by mathematics teachers in doing self-development is in fair category.

In each aspect of activities, JHS mathematics teachers in Bandar Lampung obtained information on self-development, namely (1) teachers' participation in the functional training and collective activities is related to the schedule and location of the activities and also the limitation of the participants; the difficulties to attend a master program are related to information, family permission, and funding; (2) difficulties in conducting action research (publications); it is found that some teachers have difficulties in identifying the problem, preparing good action research, elaborating research procedures, finding relevant studies, and finding expert guidance; (3) difficulties in making the classroom action research report; most of teachers have difficulties in preparing the stages of classroom action research report: writing the background, identifying the problem of the research, engaging with the objectives of the research, collecting relevant studies, getting expert guidance, determining the target achievement of the classroom action research, writing the findings and discussion, drawing conclusions, and finding motivation to write; (4) difficulties in creating popular scientific papers; some teachers complain about the difficulty of identifying ideas/issues, composing good sentences, finding appropriate references, and finding collaborative teacher colleagues, expert guidance, and also motivation; (5) difficulties in writing educational books (including textbooks, translation books, teacher manuals, enriching books, modules/ dictates) most of the teachers have difficulty in terms of finding relevant references, expert guidance, funding, and motivation, as well as in fulfilling their teaching responsibilities and credit report; (6) difficulties in journal writing and publishing book; most of difficulties emphasize that they do not know yet the benefits of publication so they do not have motivation to do it.

Furthermore, in the aspects of innovative work, the difficulties include: (1) difficulties in developing/creating/modifying media and teaching aids and developing visual aids with IT, which is also emphasized on the teaching hours, and the lack of IT skills and also motivation; and (2) difficulties in following the activities of creating instructional guidance/directions, questions, and also standard guidelines at the regency/municipality/province level where most teachers have difficulties with the opportunity to participate in the activities because only some teachers have opportunity to attend the activities.

Document Analysis and Interview Results

Document analysis using 54 research samples were conducted, while interviews

were conducted to eight respondents of the research samples. The document analysis was conducted by checking the physical evidence of teacher involvement in each aspect of self-development activities, scientific publications, and innovation such as scientific works, master degree certificate, certificate of participation in scientific activities (training, seminars, workshops, subject-matter teachers forum (or *Musyawarah Guru Mata Pelajaran* – MGMP), and/or courses), action research reports, the created/developed/modified mathematical instructional media, and other appropriate physical evidence in the last five years.

The results of the document analysis indicated that in the self-development, some teachers have been improving their education qualifications to master level. Meanwhile, in terms of the certificate of activity, most of teachers have not been involved in MGMP activities in the past two years. Regarding to activity certificates, some respondents are not so much worried about the certificates of their involvement in the activities; they are more concerned with the knowledge in the activities.

In addition, according to the document studies in scientific publications, it is revealed that most of the respondents have followed the activities of writing classroom action research reports, but no document of scientific articles and books was found. Furthermore, in the aspect of innovative work, it was found that most of teachers have created/modified simple instructional media needed for the sub-materials of mathematics learning. Moreover, some respondents followed the activities of creating guidance/directions and mathematics questions sheet at regency/municipality level.

Based on the results of interviews to eight respondents, it is known that the respondents have followed various forms of CPD activities. In fact, most of teachers have been actively involved in regular MGMP activities every month, although some teachers are still less actively involved in the forum. The schedule of the CPD activities which is at the same time of their teaching schedule is one of the difficulties oftenly complained by the teachers.

Another activity followed by the teachers is the national instruction of teaching teachers which is held by the Center of Development and Empowerment of Mathematics Teachers and Personnels (or *Pusat Pengembangan dan Pemberdayaan Pendidik dan Tenaga Kependidikan Matematika* – P4TK Matematika). It is held based on the result of Teacher's Competence Test (or *Uji Kompetensi Guru* – UKG) (PPPPTK Matematika, 2016) which end up with some teachers appointed to be national instructors. In addition, some teachers have been actively searching for information on teacher scholarship as well as improving the quality of master degree education qualification with their own funds.

In the aspect of scientific publications, most teachers conduct writing activities based on the action research results which become the prerequisite for obtaining credit numbers promoting their higher level of career. Most of the respondents have understood the action research stages and how they are reported, although they do not know whether their reports are correct or not. In addition, teachers complain about the importance of expert guidance to guide or check their writing. Then, related to the writing activities, some teachers are known to have oftenly written simple works such as arranging action research proposal, but the proposal has not yet completed due to some constraints, namely their teaching schedule, the absence of a guiding expert, the absence of positive support from their working environment, and the motivation to write. Further, related to action research activities, some teachers state the importance of collaborative activities, both in terms of the implementation and reporting.

Furthermore, in creating innovative works, the difficulties that the teachers face in developing teaching media are the lack of motivation and IT skill, their teaching schedule, and and the absence of colleagues to collaborate in carrying out the activities. Meanwhile, the difficulty in following the activity of creating the mathematics instructional guidance and question sheets is that there is no offer and opportunity to follow it. The interview result is presented in Table 6.

Table 6. Data analysis of interview results

Aspect	Reduction and Presentation Results	Conclusion (Interpretation)
CPD Implementation	- Most respondents conduct CPD only in monthly MGMP activity. - Some respondents have been, are being, and will continue to master program either by trying to find scholarships or with personal funding. - Some respondents often try to write scientific paper related to action research but it has not yet finished. - There is not found yet a teacher who made publication papers in the form of popular scientific writings and books related to mathematics learning. - There are respondents who are active to conduct CPD after the announcement of UKG results.	The implementation of CPD is still very less. Most respondents rely on the activities organized monthly by the MGMP. Furthermore, in terms of scientific publication, most teachers have not started writing scientific papers, except the action research result. Meanwhile, in the innovative works, teachers have started to create simple innovative works. The implementation of teacher's CPD in the aspect of self-development is better than other aspects.
CPD Difficulties	- Some respondents are constrained to participate in self-development activities related to the teaching schedule, information, and limitations of the participants. - Some respondents have difficulty related to the duration of using computers in writing and creating IT-based teaching media. - Most of the respondents have difficulties to write scientific papers related to action research because of family responsibilities, as well as the absence of expert guidance and feedback on the writing results. - Some respondents need fellow-teachers to collaborate with on the creation of scientific publications and innovative works.	The difficulties that become the teachers' obstacles related to the implementation of CPD are as follows: (1) in the aspects of self-development: teachers' teaching schedule, the lack of information on the CPD activities, and the establishment of limited participants; (2) in the aspects of scientific publications and innovative works: problems related to the lack of time, motivation, expert guidance, and feedback/response of the results of writing/works.

Discussion

Based on the research findings, it is indicated that the implementation of CPD for mathematics teachers in Bandar Lampung is categorized very less in the aspects of self-development, scientific publications, and innovative works. These findings are supported by the results of document study and interview which make it clear that most teachers are more active in self-development activities than in scientific publications and innovative works. The findings are in accordance with the finding of Wibowo and Jailani (2014, p. 209) that only a few mathematics teachers in have engaged in CPD. Further, based on the details of CPD implementation in each aspect, Kasmayadi (2016, p. ii) states that many mathematics teachers have been involved in CPD activities in the self-development aspect, but they are still poor in the scientific publications and innovative works aspects. Thus, it must be realized that as it is needed by a mathematics teacher, it is important to continuously improve self-competence through self-development activities, scientific publications, and innovative works, which affect the mathematics learning process and the achievement of the learners (Badri, Alnuaimi, Mohaidat, Yang, & Al Rashedi, 2016, p. 1; Powell, Terrell, Furey, & Scott-Evans, 2003, p. 389; Ünal et al., 2011, p. 3252).

Another thing to consider about the implementation of scientific publications and innovative works as a part of aspects in CPD is that those aspects require a special assessment team which will follow-up or give feedback to in assessing and reviewing their works, especially related to the results of the action research. Teachers need feedback and review in professional development activities,

especially in action research, (Chval, Abell, Pareja, Musikul, & Ritzka, 2008, p. i; Kaur, Bhardwaj, & Wong, 2017, p. 172) because both feedback and review are important factors to encourage development (New Trier Township High School, 2012, p. 4).

In scientific publication aspect, teachers' CPD activities are dominated by creating classroom action research report because submitting the report is a requirement for gaining credit numbers for the promotion of their level as a civil servant. Meanwhile, other activities such as being a speaker in a seminar/conference are still less conducted. Only some teachers become speakers as in a national instruction program provided by P4TK Mathematics. In addition, there is no teachers found conducting the activities of writing and publishing popular articles, textbooks, modules/ dictates, and translation books.

Furthermore, based on the difficulties faced by mathematics teachers of junior high school in Bandar Lampung, it is known that teachers are categorized in fairly difficult category in the aspects of self-development and scientific publications, and they are in the difficult category in the aspect of innovative work. Based on the difficulties of the CPD implementation in all aspects, it is clear that in self-development aspect, i.e. participation in functional training and collective activities or joining domestic and abroad scholarship programs for teachers, they do not face difficulties related to the support from school principal, especially on functional training and collective activities. This fact is supported by the information obtained from the interviews that all of the interview respondents state that the principals fully support them to conduct CPD, especially in being involved in the monthly MGMP activity. Even, the schools give 0 hour (day off) of teaching on the regular schedule of the MGMP activity. Furthermore, the difficulties to pursue the teachers' higher educational degree to master program are related to their teaching schedule, minimum family support, and funding.

Moreover, the difficulties faced by the teachers face in scientific publications aspect are (1) the difficulty of becoming speakers in scientific forums; some teachers' difficulties are related to the schedule and location of the agenda, and the lack of ideas and motivation; (2) difficulties in conducting action research; some teachers find it is fairly difficult to identify the problem, prepare the action research, arrange the research procedures, find relevant studies, and they find it is very difficult to have experts' guidance; (4) difficulties in writing popular scientific papers; some teachers complain about the difficulty of identifying ideas/issues, composing good sentences, and finding appropriate references, collaborative teacher colleagues, expert guidance, as well as motivation; (5) difficulties in writing educational books (which include textbooks, translation books, enriching books, teacher manuals, modules/dictates); most teachers have difficulty in terms of finding relevant references, expert guidance, fund, motivation, and fulfilling their teaching responsibilities as well as credit report at the same time; and (6) difficulties in writing journal article and publishing book; most of the difficulties emphasize that they do not know the benefits of publication yet so they do not have motivation to do it.

In scientific publication activities, the most common reasons for the difficulty are motivation, expert guidance, and collaboration. Based on the interview results, respondents state that some teachers often join training on scientific works writing, but they are still lack of writing practice. Some of the interview respondents emphasized that they need colleague collaboration and expert guidance to conduct action research and write the report. This result is supported by the important findings that some teachers have not been very proficient or do not know the writing procedures yet, such as how to write popular scientific articles and good textbooks. Further, the respondents hope that there will be such training to improve their skills. This situation is in line with the findings of a research conducted by Kasmayadi (2016, p. 176) which suggests that a certain training, such as functional training, is needed to facilitate CPD and other related topics.

Furthermore, in term of innovative works aspect, the difficulties faced by the teachers include: (1) difficulties in developing/creating/modifying learning media and

teaching aids with IT, due to their full teaching hours so that they have no opportunities to explore the probability of developing media and kits with or without IT. Some teachers have actually tried to create simple media commonly used in learning mathematics; (2) difficulties in attending the regency/municipality/province/national activity on creating mathematics learning guidance and question sheets; the difficulties are related to the opportunity to participate in such activities because only some teachers can participate in the activities. Furthermore, in relation to the innovative works, most teachers are 'users', they prefer to use the mostly used mathematics materials that have already been available in their working environment. Thus, the main factor is because the limited time and fund (Wibowo & Jailani, 2014, p. 209).

Teachers' responses in the interview are similar to the finding of a research conducted by Kasmayadi (2016, pp. 175–176) which insists that some teachers state that they have difficulties to implement the CPD. In the self-development aspect, the reasons are; there is no offer to join the training/course activities, they do not know the way of developing their selves, the have limited time due to their teaching duties or other additional tasks, they do not know the form of the activities (colloquium/panel discussion), and they do not receive the information. In the aspect of scientific publication, the reasons are: there is no offer to become a speaker, they do not have the material and ideas of what to write, they have limited time, and they are not confident to write. Meanwhile, in the innovative work aspect, the reasons of the difficulty of the CPD implementation are their lack of computer skills, limited time, and less motivation as well as idea. Related to the participation in the activities of questions/standards/guidelines preparation, the problem is because the absence of offer to join those activities.

Conclusion and Suggestions

Conclusion

Based on the results from CPD checklist sheet, the implementation of CPD activities of junior high school mathematics teachers in Bandar Lampung is in very poor category. The majority of teachers are still less involved in the CPD activities either in person or in the forum of subject-matter teachers (MGMP). Furthermore, the aspect of self-development is categorized fair, very poor, the scientific publication aspect is categorized very poor, and the innovative works aspect is also in very poor category. The difficulties of CPD implementation of JHS mathematics teachers is categorized fair. Furthermore, the difficulties in self-development aspect is categorized fair, in scientific publications aspect is categorized difficult, and in innovative works aspect is categorized very difficult.

The difficulties faced in self-development include: (1) teachers' minimum participation in functional training and collective activities caused by the schedule and location of the activities, the limitation of the participants who can join the activities; and difficulty of pursuing master degree education program because of the lack of information, family permission, and funding; (2) difficulty in conducting action research (publications), especially in identifying the problem, preparing good action research, arranging research procedures, and finding relevant studies and expert guidance. Most teachers have difficulties in creating popular scientific papers, writing journal article, and publishing books. The difficulties face in innovative works aspect include: (1) developing learning media and teaching aids because of teachers' lack of IT skills and motivation; and (2) joining the activities because of limited opportunity to participate in the activities.

Based on document analysis and interview, the implementation of CPD is still very less. Most respondents rely on the activities organized by MGMP monthly. Furthermore, in scientific publication, most teachers have not started writing scientific papers except writing action research report. Meanwhile, in the innovative works, teachers have started to create simple innovative works. The difficulties that become obstacles faced by teachers related to the implementation of CPD include their teaching schedule, the lack of information, and the limitation of participants. In the aspects of scientific publications and innova-

tive works, the obstacles are related to time, motivation, expert guidance, and feedback from the results of the writing/works.

Suggestions

Based on the discussion and conclusion of the research, suggestions for the CPD implementation are proposed: (1) for JHS mathematics teachers, it is a must to improve their ability related to classroom action research, paper writing or scientific works, and innovative works. Those activities can give impact on the better learning outcomes of the learners; (2) for school principals, they need to provide support and motivation for teachers to actively engage in various types of CPD activities constantly; (3) for the government, funds are needed to facilitate teachers in their engagement in CPD, especially in joining self-development training; (4) universities, as educator producers, should add organized coaching programs and equip students with the writing and researching abilities.

References

Abubakar, A. (2015). Dampak sertifikasi guru terhadap kualitas pendidikan pada Madrasah Aliyah di Kota Kendari. *Al-Qalam*, *21*(1), 117–128. https://doi.org/10.31969/alq.v21i1.204

Aina, M., Bambang, H., Retni, S. B., Afreni, H., & Sadikin, A. (2015). Pelatihan penulisan karya tulis ilmiah bagi guru-guru SMA 8 Kota Jambi. *Jurnal Pengabdian Pada Masyarakat*, *30*(3), 29–32.

Alnoor, A. G., & Yuanxiang, G. (2000). *Assessment mathematics teacher's competencies*. Wuhan: Central China Normal University.

Badan Standar Nasional Pendidikan. (2015). *Aplikasi PAMER UN 2015/2016*. Jakarta: BSNP (Badan Standar Nasional Pendidikan).

Badri, M., Alnuaimi, A., Mohaidat, J., Yang, G., & Al Rashedi, A. (2016). Perception of teachers' professional development needs, impacts, and barriers: The Abu Dhabi case. *SAGE Open*, *6*(3), 1–15. https://doi.org/10.1177/2158244016662901

Chval, K., Abell, S., Pareja, E., Musikul, K., & Ritzka, G. (2008). Science and mathematics teachers' experiences, needs, and expectations regarding professional development. *Eurasia Journal of Mathematics, Science & Technology Education*, *4*(1), 32–43.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE Publications.

Desimone, L. M. (2009). Improving mpact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. https://doi.org/10.3102/0013189X08331140

Fahmi, M., Maulana, A., & Yusuf, A. A. (2011). *Teacher certification in Indonesia: A confusion of means and ends*. *Working Papers in Economics and Development Studies (WoPEDS)*. Bandung: Center for Economics and Development Studies (CEDS), Padjadjaran University.

Gray, S. L. (2005). *An enquiry into continuing professional development for teachers*. Cambridge: Esmee Fairbain Foundation.

Hamilton-Ekeke, J.-T. (2013). Conceptual framework of teachers' competence in relation to students' academic achievement. *International Journal of Networks and Systems*, *2*(3), 15–20.

Jalal, F., Samani, M., Chang, M. C., Stevenson, R., Ragatz, A. B., & Negara, S. D. (2009). *Teacher certification in Indonesia: A strategy for teacher quality improvement*. Jakarta: Departemen Pendidikan Nasional Republik Indonesia.

Joubert, M., Back, J., De Geest, E., Hirst, C., & Sutherland, R. (2010). Professional development for teachers of mathematics: Opportunities and change. In *Proceedings of CERME 6* (pp. 1761–1770). Lyon, France: INRP.

Kardiyem. (2013). Analisis kinerja guru pascasertifikasi (Studi empiris pada guru akuntansi SMK se-Kabupaten Grobogan). *Journal of Economic Education*, *2*(1), 18–23.

Kartowagiran, B. (2011). Kinerja guru profesional (guru pasca sertifikasi). *Cakrawala Pendidikan*, *30*(3), 463–473. https://doi.org/10.21831/cp.v3i3.4208

Kasmayadi, W. (2016). *Model asesmen pengembangan keprofesian berkelanjutan guru sekolah menengah atas*. Doctoral dissertation. Universitas Negeri Yogyakarta, Yogyakarta.

Kaur, B., Bhardwaj, D., & Wong, L. F. (2017). Teaching for metacognition project: Construction of knowledge by mathematics teachers working and learning collaboratively in multitier communities of practice. In B. Kaur, O. N. Kwon, & Y. H. Leong (Eds.), *Professional development of mathematics teachers: An Asian perspective* (pp. 169–187). Boston, MA: Springer.

Law No. 14 Year 2005 of Republic of Indonesia about Teachers and Lecturers (2005).

Layne, H. (2016). Teacher education and teacher's professional development in Finland: Myths and realities. In *International Conference on Teacher Education and Professional Development* (pp. 8–12). Yogyakarta: LPPMP Yogyakarta State University.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: Sage.

Ministry of National Education of Republic of Indonesia. (2010). *Pembinaan dan pengembangan profesi guru buku I: Pedoman pengelolaan keprofesian berkelanjutan (PKB) dan angka kreditnya*. Jakarta: Direktorat Jenderal Peningkatan Mutu dan Tenaga Kependidikan.

New Trier Township High School. (2012). *Characteristic of professional practice at New Trier High School*. Northfield, IL: New Trier Township High School District 203.

Noorjannah, L. (2015). Pengembangan profesionalisme guru melalui penulisan karya tulis ilmiah bagi guru profesional di SMA Negeri 1 Kauman Kabupaten Tulungagung. *Jurnal Humanity*, *10*(1), 97–114.

Nuraeni, Z., & Retnawati, H. (2016). The post-certification performance of mathematics teachers. *The Online Journal of New Horizons in Education*, *6*(2), 130–142.

Opfer, V. D., & Pedder, D. (2010). Benefits, status and effectiveness of Continuous Professional Development for teachers in England. *The Curriculum Journal*, *21*(4), 413–431. https://doi.org/10.1080/09585176.2010.529651

Powell, E., Terrell, I., Furey, S., & Scott-Evans, A. (2003). Teachers' perceptions of the impact of CPD: An institutional case study ed. *Journal of In-Service Education*, *29*(3), 389–404. https://doi.org/10.1080/13674580300200282

PPPPTK Matematika. (2016). *Hasil UKG 2015*. Unpublished. PPPPTK Matematika, Yogyakarta.

Qablan, A., Mansour, N., Alshamrani, S., & Aldahmash, A. (2015). Ensuring effective impact of continuing professional development: Saudi science teachers' perspective. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(3), 619–631. https://doi.org/10.12973/eurasia.2015.1352a

Regulation of the Minister for the State Apparatus Empowerment and Bureaucracy Reform No. 16 of 2009 on Teachers' Functional Position and Their Credit Points (2009). Republic of Indonesia.

Regulation of the Minister of Education and Culture No. 20 of 2016 on the competence standard of primary and secondary education graduates (2016). Republic of Indonesia.

Regulation of the Minister of National Education No. 19 of 2005, on National Education Standard (2005). Republic of Indonesia.

Supriyanto, A. (2015). Harapan, kenyataan dan strategi peningkatan kemampuan guru dalam penulisan karya tulis ilmiah. In *Prosiding Seminar Nasional Pengembangan Keprofesian Menuju Guru Profesional* (pp. 109–114). Malang: Universitas Negeri Malang.

Tuijnman, A., & Boström, A.-K. (2002). Changing notions of lifelong education and lifelong learning. *International Review of Education*, *48*(1/2), 93–110.

Ünal, H., Demir, I., & Kiliç, S. (2011). Teachers' professional development and students' mathematics performance: Findings from TIMSS 2007. *Procedia - Social and Behavioral Sciences*, *15*, 3252–3257. https://doi.org/10.1016/j.sbspro.2011.04.280

UNESCO. (2014). *Education strategy 2014-2021*. Paris: UNESCO.

UNICEF. (2007). *A human rights-based approach to education for all*. New York, NY: United Nations Educational, Scientific, and Cultural Organization.

Wermke, W. (2011). Continuing professional development in context: Teachers' continuing professional development culture in Germany and Sweden. *Professional Development in Education*, *37*(5), 665–683. https://doi.org/10.1080/19415257.2010.533573

Wibowo, E., & Jailani, J. (2014). Analisis kesulitan guru matematika SMP dalam pengembangan profesi di Kabupaten Wonosobo. *Jurnal Riset Pendidikan Matematika*, *1*(2), 202–215. https://doi.org/10.21831/jrpm.v1i2.2676

Widoyoko, E. P. (2013). *Evaluasi program pembelajaran: Panduan praktis bagi pendidik dan calon pendidik*. Yogyakarta: Pustaka Pelajar.

World Bank. (2010). Transforming Indonesia's teaching force. Washington, DC: Human Development East Asia and Pacific Region, World Bank.

Yacoub, Y. (2012). Pengaruh tingkat pengangguran terhadap tingkat kemiskinan Kabupaten/Kota di Provinsi Kalimantan Barat. *Jurnal EKSOS*, *8*, 176–185.

SUBMISSION GUIDELINES

- The manuscript submitted is a result of an empirical research or scientific assessment of an actual issue in the area of educational measurement, evaluation, and assessment in a broad sense, which has not been published elsewhere and is not being sent to other journals.
- Only articles written in English will be considered. Any consistent spelling and punctuation styles may be used. Please use single quotation marks, except where 'a quotation is "within" a quotation'. Long quotations of 40 words or more should be indented without quotation marks.
- A typical manuscript is approximately 4,000-7,000 words (or 8-15 pages using the journal template) including the abstract, tables, figures, references, and captions. Manuscripts that greatly exceed this will be critically reviewed with respect to length. (A4; margins: top 3, left 3, right 2, bottom 2; double columns [Except in Abstract: single column]; single-spaced; font: Garamond, 12).
- Manuscripts should be compiled in the following order: (1) title; (2) abstract; (3) keywords; (4) main text: introduction, method, findings and discussion, conclusion and implications, recommendations, or suggestions (if any); (5) acknowledgements for the Funding and grant-awarding bodies (if any); (6) references; and (7) appendices (as appropriate).
- (If any) The funding or grant-awarding bodies are acknowledged in a separate paragraph. *For single agency grants:* "This work was supported by the [Name of Funding Agency] under Grant [number xxxx]."
- The title of the manuscript should clearly represent the content of the article.
- Authors' identities under the title should be omitted, and replaced by the following item:

  Anonymous

  (Author's identity is omitted due to review process)
- An abstract that does not exceed 250 words is required for any submitted manuscript. It is written narratively containing the aim(s), method, and the result(s) of the research.
- Each manuscript should have 3 to 6 keywords written under the abstract.
- All tables and figures are adjusted to the paper length and are numbered and referred to the text.
- The citation and references are referred to American Psychological Association (APA) (Sixth Edition) style.
- APA Style format for references can be checked in http://www.citationmachine.net/apa/cite-a-website
- The author is strongly preferred to use Reference Manager application.
- The manuscript must be in *.doc or *.rtf , and sent to **REiD's Management** via online submission by creating account in the Open Journal System (OJS) [click **REGISTER** if you have not had any account yet; or click **LOG IN** if you have already had an account].
- Authors' biography is written in the form of narration, including author's full name, place and date of birth, educational qualification/information started from bachelor degree (S1) until the latest educational degree, the affiliation in which the author is currently working, phone number, and email address.
- All Author(s)' names and identity(es) must be completely embedded in the form filled in by the corresponding author: email; affiliation; and each author's short biography (in the column of 'Bio Statement'). [if the manuscript is written by two or more authors, please click 'Add Author' in the 3rd step of 'ENTER METADATA' in the submission process and then enter each author's data.]
- All correspondences, information, and decisions for the submitted manuscripts are conducted through the email/s used for the submission.
- Word template is available for this journal. Please visit the journal's homepage at https://journal.uny.ac.id/index.php/reid
- If you have submission queries, please contact *reid.ppsuny@uny.ac.id* or *reid.ppsuny@gmail.com*