

**REiD (RESEARCH AND EVALUATION IN EDUCATION)**  
**Vol. 4, No. 2, December 2018**

“My lecturer’s expressionless face kills me!” An evaluation of learning process of German language class in Indonesia  
--Primardiana Hermilia Wijayati; Rofi’ah; Ahmad Fauzi Mohd Ayub

Technology-enhanced pre-instructional peer-assessment: Exploring students’ perceptions in a Statistical Methods course  
--Yosep Dwi Kristanto

Developing assessment model for bandel attitudes based on the teachings of Ki Hadjar Dewantara  
--Restituta Estin Ami Wardani; Supriyoko; Yuli Prihatni

Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics  
--Syukrul Hamdi; Iin Aulia Suganda; Nila Hayati

Performance assessment and the factors inhibiting the performance of Buddhist education teachers in the teaching duties  
--Hesti Sadtyadi

Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT  
--Edi Istiyono; Wipsar Sunu Brams Dwandaru; Revnika Faizah

An evaluation of internship program by using Kirkpatrick evaluation model  
--Lathifa Rosiana Dewi; Badrun Kartowagiran

Comparing the methods of vertical equating for the math learning achievement tests for junior high school students  
--Chairun Nisa; Heri Retnawati

Indexed in:



Research and Evaluation  
in Education

Vol. 4, No. 2, December 2018

**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

**Editor in Chief** : Djemari Mardapi  
**Editors** : Badrun Kartowagiran  
Edi Istiyono  
Samsul Hadi  
Elizabeth Hartnell-Young  
John Hope  
Suzanne Rice  
Nur Hidayanto Pancoro Setyo Putro  
Alita Arifiana Anisa  
Lye Che Yee  
Socheath Mam  
Suhaini M. Saleh

**Journal Coordinator of Graduate School of Universitas Negeri Yogyakarta**

Ashadi

**Setting**

Rohmat Purwoko  
Ririn Susetyaningsih  
Syarief Fajaruddin

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

THE EDITORS ARE NOT RESPONSIBLE FOR THE CONTENT OF AND  
THE EFFECTS THAT MIGHT BE CAUSED BY THE MANUSCRIPTS.

RESPONSIBILITY IS UNDER THE AUTHORS'.

**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

Copyright © 2018, REiD (Research and Evaluation in Education)

## Foreword

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

REiD (Research and Evaluation in Education) contains and spreads out the results of research which is not limited to the area of common education, but also comprises the results of research in education in a broader coverage, such as language education, cultural education, physics education, mathematics education, and teacher 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, December 2018

Editor in Chief

## TABLE OF CONTENT

<i>Primardiana Hermilia Wijayati</i> <i>Rofi'ah</i> <i>Abmad Fauzi Mohd Ayub</i>	“My lecturer’s expressionless face kills me!” An evaluation of learning process of German language class in Indonesia	94-104
<i>Yosep Dwi Kristanto</i>	Technology-enhanced pre-instructional peer assessment: Exploring students’ perceptions in a Statistical Methods course	105-116
<i>Restituta Estin Ami Wardani</i> <i>Supriyoko</i> <i>Yuli Prihatni</i>	Developing assessment model for bandel attitudes based on the teachings of Ki Hadjar Dewantara	117-125
<i>Syukrul Hamdi</i> <i>Iin Aulia Suganda</i> <i>Nila Hayati</i>	Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics	126-135
<i>Hesti Sadtyadi</i>	Performance assessment and the factors inhibiting the performance of Buddhist education teachers in the teaching duties	136-143
<i>Edi Istiyono</i> <i>Wipsar Sunu Brams Dwandaru</i> <i>Revnika Faizab</i>	Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT	144-154
<i>Lathifa Rosiana Dewi</i> <i>Badrin Kartowagiran</i>	An evaluation of internship program by using Kirkpatrick evaluation model	155-163
<i>Chairun Nisa</i> <i>Heri Retnavati</i>	Comparing the methods of vertical equating for the math learning achievement tests for junior high school students	164-174



## “My lecturer’s expressionless face *kills* me!” An evaluation of learning process of German language class in Indonesia

\*<sup>1</sup>Primardiana Hermilia Wijayati; <sup>2</sup>Rofi’ah; <sup>3</sup>Ahmad Fauzi Mohd Ayub

<sup>1,2</sup>German Department, Faculty of Letters, Universitas Negeri Malang  
Jl. Semarang No.5, Sumbersari, Kota Malang, Jawa Timur 65145, Indonesia

<sup>3</sup>Faculty of Educational Studies, Universiti Putra Malaysia  
43400 Serdang, Selangor, Malaysia

\*Corresponding Author. E-mail: [primardiana.hermilia.fs@um.ac.id](mailto:primardiana.hermilia.fs@um.ac.id)

*Submitted: 19 December 2018 | Revised: 20 December 2018 | Accepted: 21 December 2018*

### Abstract

This qualitative study aimed at evaluating the circumstances in plenary class that provoke learners’ speaking anxiety. To meet the objectives, this study investigated students of German as a Foreign Language (GFL) course who were experiencing speaking anxiety symptoms in the plenary class. The research was a narrative qualitative study, and the data were collected through observation and interview. The result of this study reveals that learners’ speaking anxiety occurred in particular circumstances of the plenary class, such as unfamiliar topic, still class, students’ unpreparedness for spontaneous speaking, expressionless face of the lecturer, and students’ fear of native speaker lecturers.

**Keywords:** *speaking anxiety, German as a foreign language (GFL), foreign language anxiety (FLA), sozialform*

### Introduction

German is one of the foreign languages learned by several learners in Indonesia, including in Universitas Negeri Malang. Unlike English, German is not an international language that had been learned earlier in the elementary school or even kindergarten. It is also not as familiar as English. In German Department of Universitas Negeri Malang, the students have various backgrounds. Some students have prior knowledge about German from their high schools, some others do not have any knowledge at all and they started to learn German in the university. Students’ German language knowledge is standardized through Gemeinsamer Europäischer Referenzrahmen (GER) (2004) or Modern Language Division (2001). According to GER, German skills are divided into three levels i.e. basic level that consists of A1 (breakthrough) and

A2 (Waystage), independent level that consists of B1 (Threshold) and B2 (Vantage), and competent level that consists of C1 (Effective Operational Efficiency) and also C2 (Mastery) (Glaboniat, Müller, Rusch, Schmitz, & Wertenschlag, 2005). In Universitas Negeri Malang, German was taught with a different level in each semester. Students learned German I (A1) in the first semester, German II (A2) in the second semester, German III (A2-B1) in the third semester, German IV (B1) in the fourth semester, and then German B2 (Deutsch auf B2 Niveau) in the fifth semester. The students had to pass the lower level first in order to reach the higher level (Department of German Letters, 2016).

In the class of Deutsch auf B2 Niveau, the students did not learn the whole level. They only learn the beginning or the basis of B2 level. However, students’ German language skill at this level should be good, since

the students already got B1 and B2, which are independent levels. The point is students at this level should be able to communicate in German fluently. However, in fact, according to the preliminary research, the students were quite passive and looked anxious to speak especially when they had to speak in the plenary class or in front of their classmates and teacher. It indicated that they suffered classroom speaking anxiety.

Speaking is one of the language skills that should be acquired by students in a foreign language class. Speaking as a productive activity is very important for them to communicate with each other not only for academic, but also interpersonal context (Lightbown & Spada, 2006). Speaking is often considered the most difficult language skill because students need to go through a complicated process in order to speak correctly and understandably (McLaren, Madrid, & Bueno, 2006).

Speaking includes a combination of some cognitive and psychological aspects. In order to achieve successful speaking, students need to have sufficient language knowledge and good psychological (mental) state. The cognitive aspect consists of bottom-up and top-down processes (Bashir, Azeem, Ashiq, & Dogar, 2011; Saville-Troike, 2006). The bottom-up process involves language knowledge such as vocabulary, pronunciation and grammatical patterns. Meanwhile, the top-down process involves content knowledge about a topic and cultural knowledge of the spoken language. Furthermore, the psychological aspect or mental state also affects students' speaking skill. One of the psychological aspects that affect speaking skill is anxiety (Ansari, 2015; Muhaisen & Al-Haq, 2012).

Speaking anxiety in a language class is manifested in some ways. Some researches show that speaking anxiety increases students' monitor use (Dulay, Burt, & Krashen, 1982; El-Sakka, 2016). Students cannot speak fluently because they are self-conscious. This situation worsens their speaking ability (Shabani et al., 2013; Von Wörde, 2003). They cannot achieve their maximum achievement in speaking. Some researches show the cause of students' speaking anxiety, such as lack of fluency, poor knowledge of vocabulary, unfamil-

iar topic, and negative feedback (Awan, Azher, Anwar, & Naz, 2010; Barahmeh, 2013; Nazir, Bashir, & Raja, 2014). This phenomenon can be seen in almost every language class, including German language class.

Some researches on speaking anxiety in German language class have revealed some familiar findings. Students can suffer fear by speaking in a German language class. The main causes are, for example, fear of negative feedback, low language proficiency, and shyness (Fischer & Modena, 2005). That fear by speaking leads to speaking anxiety. This anxiety affects students' language ability and worsened their linguistic mastery because they cannot think clearly under those circumstances (Sevinç & Backus, 2017). These findings are found in German as a second language class. Because speaking anxiety has a huge effect, it is important to investigate speaking anxiety and its form. This research's context is different from previous researches, namely German as a foreign language (GFL) in Indonesia.

It is familiar to see in German as a foreign language class in Indonesia: the lecturer asks a question to the students and they respond it as if it is in a choir. However, when lecturer asks a student to raise a hand and to speak in front of the class, the student keeps quiet as if the class becomes a 'graveyard'. This is because most of the students are passive and anxious to speak in front of the class (Cansrina, 2015). Such description is a kind of culture in German as a foreign language classes in Indonesia.

Based on the preliminary research, the students' passiveness became a serious problem in the classroom. It also gave negative effects toward their performance. During the class, a few students who spoke actively were always the same persons. Thus, the lecturer had to ask or even force the other students to speak. Otherwise, they would only speak with their classmates when they had to interact with each other in pair work or group work.

The students' passiveness and fear of speaking, as mentioned before, show that there were speaking anxiety symptoms among them. This situation normally happened in the plenary class. Plenary class is an interactive form or Sozialform, which is a term defined

as a didactic methodology that arranges the interaction pattern between students and teachers and among students which consists of plenary class (Frontalunterricht/pleno), individual work (Einzelarbeit), pair work (Partnerarbeit), and group work (Gruppenarbeit) (Kiper, Meyer, & Topsch, 2002). Thus, this study focused only on the learners' speaking anxiety in the plenary class.

Classroom speaking anxiety is a kind of unpleasant feeling suffered by foreign language learners as they are asked to speak in the classroom. Speaking anxiety is defined as a feeling of fear, nervous, and lack of self-confidence during speaking which are associated with visual signs (Horwitz, 2001; Horwitz, Horwitz, & Cope, 1986; Tseng, 2012; Wilson, 2006; Zhiping & Paramasivam, 2013). Basic (2011) also states that anxiety is a sort of fear manifested by visual signs. Speaking anxiety is a part of Foreign Language Anxiety (FLA) experienced by foreign language learners (Bashir et al., 2011; Horwitz, 2001; Horwitz et al., 1986). Thus, anxiety in the speaking skill is a problem experienced by most of the students in foreign language classes (Arnaiz & Guillén, 2012; Basic, 2011; Horwitz, 2001; Horwitz et al., 1986; Marwan, 2007; Tseng, 2012; Wilson, 2006; Zhiping & Paramasivam, 2013). It is caused by the complexity of speaking skill (Basic, 2011). It becomes a reason why the researchers attempted to conduct deeper studies about speaking anxiety with various focuses and results.

Nowadays, there is a number of speaking anxiety studies in English as a Foreign Language (EFL) classes as well as in German as a Foreign Language (GFL) classes. Tseng (2012) explains that there are some factors that can cause speaking anxiety in English classes, such as parents' and teachers' demands for students to get good grades at school in English, lack of confidence in students' ability to learn English, fear of making mistakes and of getting subsequent punishment or ostracism, i.e. fear of having embarrassing feeling for not being perfect, condition in childhood to believe that English is extremely difficult, and fear of foreigners and their behavior. It all shows that English triggers anxiety because of its role as an inter-

national language. However, the cause of speaking anxiety in another foreign language, such as German, should be different.

Zhiping and Paramasivam (2013) attempted to look for the cause of speaking anxiety in an international class in Malaysia where the students are from Nigeria, Iran, and Algeria. Their findings revealed that there are particular factors that provoke speaking anxiety, (e.g. fear of being in public and shyness, fear of negative evaluation, and fear of speaking inaccurately). In addition, students' speaking anxiety level is various. It depends on the student and also their culture. Therefore, the cause of speaking anxiety among students was much related to cultural difference since they came from different countries.

The researches about speaking anxiety in German as a Foreign Language (GFL) class had been done by Fischer and Modena (2005) and Cansrina (2015) who investigated speaking anxiety in Modena University in Italy. The results indicate that motivation has a great deal to the success of students' speaking skill. Students with high motivation as well as self-confidence in learning German can speak German well. Meanwhile, students who suffer speaking anxiety and are afraid to get negative evaluation have low speaking skill. Gnjudić (2016), in his study, has found that anxiety and fear are the biggest obstacles to learning German for Croatian students. When students have a high anxiety level, they can poorly concentrate in producing and expressing their idea through speaking (Fischer & Modena, 2005; Inozemtseva, 2017).

A local study by Cansrina (2015) divided the causes of German students' speaking anxiety in German Literature Padjajaran University based on three aspects, i.e. personal, social-didactic, and cultural aspects. The cause of speaking anxiety based on personal aspect is too much thinking about grammar and fear of negative evaluation. Based on social-didactic aspect, students will feel anxious to speak when the topics are unfamiliar. Meanwhile, seen from the cultural aspect, students' feeling of fear was provoked by Indonesian teachers' behavior since elementary school, i.e. the students have to keep silent in the class.



A bit different to the studies by Fischer and Modena (2005), Cansrina (2015), Gnjidić (2016), and Inozemtseva (2017), this study investigated speaking anxiety in particular circumstances of interaction forms in the classroom. Actually, the interaction form of the classroom consists of plenary class, individual-, pair-, and group work, but this study focused on speaking anxiety that occurs only in the plenary class. According to the authors' teaching experiences and the preliminary research, it can be assumed that the students were more passive and anxious in the plenary class rather than in the individual-, pair-, or group work activities. That is why this study focused only on speaking anxiety in the plenary class.

## Method

This qualitative research aimed to evaluate the learning process of German language class in Indonesia by trying to reveal which circumstances of the plenary class that provoked speaking anxiety of German learners in Universitas Negeri Malang. The respondents in this study were students of Universitas Negeri Malang who had ZiDs or Zertifikat Indonesische Deutschstudierende (certificate of German skill for Indonesians students), and who attended Deutsch auf B2 Niveau (German level B2) class as well as Deutsche Literatur (German Literature) class. Such students were selected because they had sufficient input of German. Ideally, they should be able to communicate in German fluently.

Data of the study were collected through observation and interview. The observation was conducted in Deutsch auf B2 Niveau class and Deutsche Literatur class to observe and to notice the symptoms of learners' speaking anxiety during the plenary class. Deutsch auf B2 Niveau was taught by two Indonesian lecturers, while Deutsche Literatur was taught by a German lecturer. In addition, the interview was conducted to support and confirm the data. This study used participant observation, i.e. passive participation. The researchers were not directly involved in the classroom activity, because they rolled as camera persons who recorded and observed the learning process. The researchers came to the

class as researchers who observed and recorded the whole activities of the class by using a video recorder. Through the videos, the data were analyzed using an observation sheet. There were several indicators on the sheet to find students who showed speaking anxiety symptoms.

After conducting the observation, there are eight students who were indicated suffering from speaking anxiety were interviewed. The researchers met the students one by one and interviewed them personally to dig deeper data about their speaking anxiety. There were several questions in the interview sheet, but the questions could develop according to the information from the interviewee. It means that the interview was arranged to expose the interviewees' personal view (Creswell, 2013; Sugiyono, 2012).

In qualitative research, data analysis is a continuous process that needs continuous reflection along the study. The technique used in this study was adapted from Spradley that consist of three kinds of analyses, i.e. domain analysis, taxonomy analysis, and componential analysis (Spradley, 1980; Wijayati, 1995). The data in this study were analyzed by those three techniques as mentioned before.

In domain analysis, there is a term called cultural domain. It is a category of cultural meaning that includes small categories. Domain, as the cultural category, consists of three basic elements, i.e. cover term, included term, and semantic relationship (see Figure 1). The cover term is a term for a cultural domain category, included term is a term for smaller cultural domain category, and the semantic relationship is a term that relates the cover term and the included term (Spradley, 1980, p. 89; Wijayati, 1995, p. 32).

The results of the domain analysis were analyzed through taxonomy analysis. It was almost the same as domain analysis that consists of categories arranged by semantic relationship. The difference was taxonomy analysis focused on the relationship that appears at the cultural domain. Then the data were analyzed through componential analysis. A componential analysis was systematic research for the meaning components related to the structural category. The componential analysis

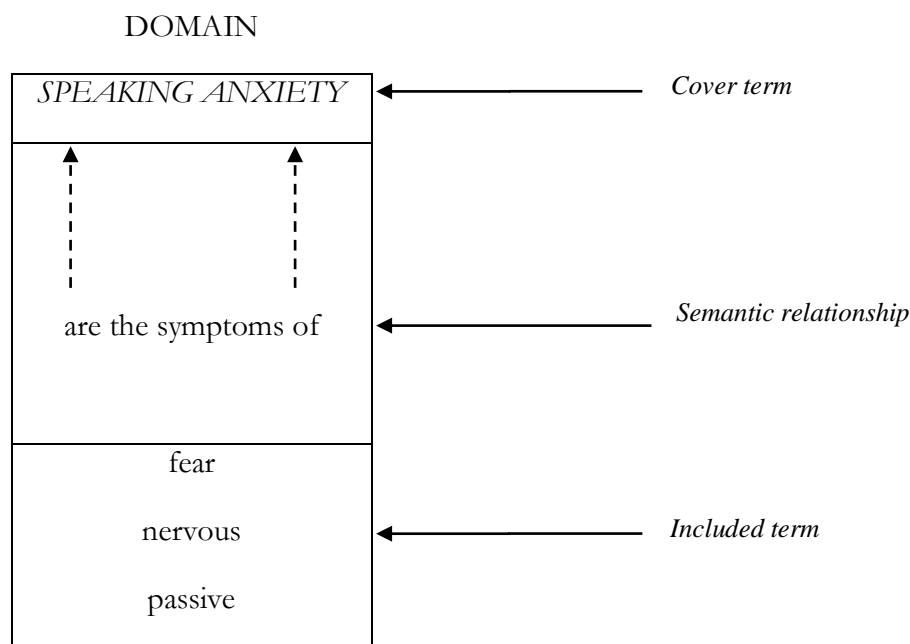


Figure 1. Example of domain analysis

was used to find the meaning which were shown by the research object toward the cultural category (Spradley, 1980; Wijayati, 1995, p. 43).

### Findings and Discussion

Plenary is one of the interaction forms that arrange the interaction pattern between teacher and students in the classroom, in which the teacher stands in front of the class, while the students sit towards the teacher. In the plenary class, the teacher is the master of the class who conducts the learning process. The teacher controls the communication and the learning process of the class (Nuhn, 2000). This kind of interaction form is easy to conduct because it does not need much preparation. Moreover, the teacher could know or even evaluate students' progress in learning a foreign language directly. However, there are particular circumstances of plenary class that provoke learners' speaking anxiety. Based on the study, learners' speaking anxiety occurs during plenary class on particular circumstances, e.g. when the topic is unfamiliar, when the students are unprepared for spontaneous talks, when nobody answers and the class is so still, when the lecturer is expressionless, and when the students are taught by a native speaker.

### Unfamiliar Topic

In the plenary class, the students would like to speak when the topic of the lesson was interesting and familiar to them because they found that it was great and easy. On the contrary, when the topic of the lesson was unfamiliar, they were passive because it was boring and difficult for them to speak. It happened in all classes no matter if they were taught by a German or Indonesian lecturer. Data (1)-(13) are the data that show that issue.

- (1) When I talk a kind of topic that I used to talk in my daily life, I think my vocabulary is relatively okay, so, it doesn't matter. But when the topic is rather difficult, I feel nervous, *Ma'am*. (MDCW)

Datum (1) shows that the student was feeling nervous when he had to talk about a difficult topic. Nervousness is one of the speaking anxiety symptoms that occur during speaking (Horwitz, 2001; Horwitz et al., 1986; Lightbown & Spada, 2006; Spielberger, 1983; Tseng, 2012; Wilson, 2006; Zhiping & Paramasivam, 2013). Such a case also happened to the following student, as recorded in Datum (2).

- (2) In *ZidS* class, I got a topic to speak, namely online shop. At that time I really had no idea about that and I said, 'I'm sorry, I have no idea about that, because I never shop online yet... I had no experience about that, so I just can't tell about that topic.' (FNS)

The afore-mentioned description shows that topic played an important role in the classroom interaction. Interesting, easy, and familiar topics could help the students to communicate easily, while difficult and unfamiliar topics provoked learners' speaking anxiety. The students suffered speaking anxiety when the topic was not in their interest. Difficult topics that require higher vocabulary skill provoked the students' fear to speak as well. In addition, unfamiliar topics such as German culture or something that the students never experience by themselves triggered their speaking anxiety. Those research findings support what Cansrina (2015) explains that students have no fear to speak when the topic of the class is interesting.

#### Unprepared Students for Spontaneous Talk

The next circumstance that provoked learners' speaking anxiety in the plenary class was when the students were unprepared to talk spontaneously. Based on the observation, the students looked so shy and were smiling when they were being called and being asked to speak. Some of them showed their tension and nervousness. The other students showed repetitive gestures such as scratching their hand and head or touching their face several times, which mean that they were nervous. It can be seen that they suffered speaking anxiety as Basic (2011) says that anxiety is a sort of fear manifested by visual signs.

Based on the interview results, the students were feeling tension when they were suddenly asked to speak. They were afraid because they have no preparation before. It made them speechless, as seen in Datum (3).

- (3) When I'm suddenly asked to speak, my mind was blank, and I don't even know what to speak ... First, it is because I don't prepare it well, sometimes the

grammar sounds odd either. It is a kind of a mix between tension and confusion. (DNA)

Datum (3) shows that the student experienced mental block or losing an idea because of the sudden call to speak. Mental block indicated that the students suffered anxiety. Horwitz et al. (1986) call it communication apprehension. Communication apprehension is a part of FLA that causes speaking disruption such as stutter, mind blank or losing idea and words, and high intonation or on the contrary. In this study, the students experienced mental block because they were shocked as they were suddenly called to speak. In a long-established habit in Indonesian classes, most of the teachers used to call the students in sequence (based on position or alphabetically). Thus, the students could prepare what to speak while they were waiting for their turns. That is why they were shocked and nervous when they were unprepared for spontaneous talks.

Other data which show that students were feeling the tension and afraid to speak are presented in Data (4) and (5).

- (4) When I'm not ready, it is disturbing to be asked to speak. Also it is much better not to ask or *freivillig*\*. But when I already prepared, it is okay to be asked to speak. (HI)  
\**freivillig*: free willing (to speak)

- (5) I feel more afraid when I am asked to speak, because when I don't understand yet, I am not ready, so what do I have to speak? (FWEP)

Data (4) and (5) show that the students were afraid to talk if they are not ready and do not understand the material yet. They were afraid to make mistakes, whether it is in the content or the grammar. Fischer and Modena (2005) and also Zhiping and Paramasivam (2013) state that the cause of speaking anxiety is the fear of making mistakes. In Cansrina (2015) research, the fear of making mistakes as one of the factors that provoke speaking anxiety was not significant, while, in this stu-

dy, the fear of making mistakes was a big reason that provoked most of the learners' speaking anxiety.

In addition, the students were afraid of negative evaluation, so they were afraid to make mistake when they answered the lecturer's question. It also supports the finding by Fischer and Modena (2005), Zhiping and Paramasivam (2013), and Cansrina (2015) who state that students are afraid of negative evaluation, especially from the lecturer. That is why they suffer speaking anxiety.

#### Quiet Class

Based on the observation in Detusche Literatur class, the lecturer asked for the students' opinions about a part of a novel that they have read. Nobody answered. The quietest the class, the worrier were the students. It could be seen in Datum (6).

- (6) I don't really understand what that German lecturer wants. I mean, what he wants us to do. Even if I know, but why do my classmates keep silent? The lecturer asked us to do this, but why they say nothing? So I keep silent too. (WDB)

As seen in Datum (6), the students kept silent when there was nobody that has the courage to answer first. They were afraid to reveal their ideas orally (Basic, 2011; Horwitz, 2001; Horwitz et al., 1986; Spielberger, 1983; Tseng, 2012; Wilson, 2006; Zhiping & Paramasivam, 2013). They also have no interest to speak which indicates that they suffer speaking anxiety (Horwitz et al., 1986).

Based on the observation, the students answered the lecturer's question in a choir. Cansrina (2015) says that they did that because if they were making mistakes, at least they were not alone. They did it together, so they felt safe. However, if no one had the courage to answer, it was better to keep silent. It seemed if somebody talked alone in front of the class and he/she made mistakes, then he/she would become the 'defendant'. It made her/him embarrassed. That circumstance triggered students' shyness, fear, and tension to speak.

Such circumstances created a negative atmosphere in the classroom. The negative atmosphere gave a negative impact to the students. Thus, the negative atmosphere contributed to learners' speaking anxiety. It means that the class needs a positive atmosphere as stated by Cansrina (2015) that a positive atmosphere of the class contributes to learners' speaking activity.

#### Expressionless Face of the Lecturer

The next circumstance that provoked learners' speaking anxiety in the classroom was the expressionless face of the lecturer. According to the interview, expressionless face of the lecturers triggered students' tension as presented in Data (7) and (8).

- (7) It is even more frightening if the listener's\* face was expressionless. If they are nodding, it means 'o, everything is alright' (laugh), but when they show their flat expression, o my, it kills me! What should I do then? (FWEP)  
\*lecturer
- (8) Lecturers' expression decides whether I can speak or not. If they ask me to speak with smiling face, I feel, well, still nervous, but not much. But when they are expressionless I'm afraid to speak in front of the class. (FNS)

According to Data (7) and (8), it could be seen that expressionless face of the lecture provoked the learners' speaking anxiety. They were afraid to interpret the lecturer's expression, so they were feeling nervous and afraid to speak. When the lecturer's face was expressionless, the students were frightened by him/her so that they were afraid to speak freely and they more focused on language accuracy.

The learners' fear caused by the expressionless face of the lecturer appeared because of the learners' own perception. This result did not appear in other relevant speaking anxiety researches. The students were not sure with their own answers. So they guessed the lecturers' expression to know whether their answers were true or false. Thus, they were afraid of making mistakes and afraid of get-

ting a negative evaluation from the lecturers (Cansrina, 2015; Fischer & Modena, 2005; Zhiping & Paramasivam, 2013). Logically, if they were not afraid of negative evaluation, they would not be afraid of making mistakes. It means that such perception came from the students themselves.

#### Native Speaker Lecturer

In the plenary class of the study, there was a distinction between the class that was conducted by Indonesian and German lecturers. Based on the observation, the students suffered speaking anxiety in particular circumstances, but they were still active enough when they were taught by Indonesian lecturers. Meanwhile, in the class that was conducted by a native speaker, the students were extremely passive. The students were passive and did not want to speak before the lecturer directly asked a student to speak. When the lecturer asked the class, nobody would answer. When the lecturer repeated the question, the students whispered to their classmates and discussed it with them in a whisper. In case they did not understand the question, instead of asking the lecturer directly, they asked their classmates. Some students even avoided eye contact which is one of the speaking anxiety's symptoms (Cansrina, 2015).

According to the interview outcomes, it was because the students had difficulty to understand what the native speaker said. His dialect and accent were a bit different. When Indonesian lecturers spoke, the students could understand their accent, because they had the same mother tongue as the students.

(9) If the language used in the class is full German but the lecturer is Indonesian, they still could express it and their accent is still like Indonesians. But in *Deutsche Literatur* class that is conducted by a native speaker, it is so confusing, because we have to speak full German and his pronunciation sounds 'extremely German'. Sometimes I do not like to attend the class (laughing). (MD)

(10) Indonesian lecturers may understand when we made grammar mistakes. But

native speaker, I'm afraid if they don't understand what we said. I'm afraid so. (TN)

According to Data (9) and (10), the students found that Indonesian lecturers could understand them well. Their accent was easy to understand. The students assumed that Indonesian lecturers knew their difficulties in grammar because they had the same mother tongue. In addition, if the students did not understand particular words, Indonesian lecturers could explain it in Indonesian language. That is why the students were feeling glad and safe when they were taught by Indonesian lecturers.

On the contrary, students were nervous when they were taught by a native speaker because they thought that a native speaker could not understand their difficulties and their culture as well as Indonesian lecturers which are evidenced by Datum (11).

(11) When I talked with a native speaker I feel so nervous, because, *eee*, every time we speak slowly and stuttered, he shows different expressions (confused), but Indonesian lecturers, just like our lecturers, know our behavior well. (TNTR)

In addition, the students thought that a native speaker was the owner of the language they learned. That is why they were being forced by themselves to make the native speaker understood what they said. They thought that the native speaker would notice every grammar mistake they made more than Indonesian (lecturer). For that reason, they had to focus on grammar accuracy that made them more nervous. It could be seen from Data (12 and (13).

(12) Mostly I feel nervous when I talk with German native speaker (German lecturer) because German is his mother tongue. I'm very afraid, whether my grammar is true or false. (WDB)

(13) When there are Germans, I mean, outside of the university, actually I really want to speak with them. But, I'm

afraid if I make mistake during speaking. They are foreigners who don't understand us well. I'm afraid if I make mistakes and they find it odd or something like that. (FNS)

According to Data (12) and (13), it is found that the class that was conducted by a native speaker provoked difficulty for students. The students felt more nervous and were afraid to interact if the lecturer was a German native speaker because they were afraid to make grammar mistakes. In addition, the data show that the students had a feeling of fear of foreigner. According to the observation, the students avoided the chairs near with the German lecturer. Some chairs in front of the German lecturer were empty at the beginning of the class and the chairs were only for them who came too late as if it was a punishment for them. It shows that the students avoided taking a seat near the native speaker because they did not feel ease and they were afraid of a foreigner.

All of the data mentioned reveal that the students suffered speaking anxiety in the plenary class, when the lecturer was a native speaker. This finding supports the finding of Tseng (2012) that the cause of the learners' speaking anxiety is a fear of foreigner and their behavior. In this study, the students were quite passive and afraid when they were taught by a native speaker, but they did not afraid of his/her behavior. According to the interview, the students found that the German lecturer was nice and humble. However, the students found that the pronunciation and the accent of the German lecturer were quite different and sounded so difficult to understand, unlike the Indonesian lecturers' pronunciation that was easy to understand. The students were also afraid to make grammar mistakes and if the German lecturer did not understand them and their culture as well.

When the students spoke German in front of a German lecturer, the fear of making mistakes intensified because the students assumed that the German lecturer was the owner of the language (German) who would easily notice when the students were making mistakes. That is why they focused on language

accuracy. Like what Cansrina (2015) says, that learners' speaking anxiety occurs because they think too much about grammar. That circumstance provoked students not to focus on meaning, but to focus on their fear of making mistakes.

In addition, based on the observation in the Deutsche Literature class, the German lecturer's teaching methods were not quite interesting to the students. They only read the stories in the books. The lecturer asked them the content or the main idea of the stories and their opinion about them. When nobody answered the lecturer's question, the lecturer explained it by himself. On the other day, the German lecturer showed a German poem, explained the difficult vocabulary, and then asked the students to interpret it. Every student kept silent. Then the lecturer explained and interpreted the poem by himself, again. Such methods were boring and too difficult for the students. That is why they had no desire to speak in the classroom.

From all those data, it could be concluded that the students suffered speaking anxiety when the lecturer was a native speaker. They were afraid of making mistakes and getting a negative evaluation from the owner of the language. Besides, they were also afraid of foreigner. In addition, their speaking anxiety increased when the native speaker lecturer's teaching methods were not interesting and too difficult for them.

## Conclusion and Suggestions

Plenary is an interactive form that was often used in the classroom since the teacher could control the communication and the learning process. It was also easy to do (for the teacher/lecturer) and the teacher could know or even evaluate the students' progress in learning a foreign language directly. On the other hand, there were particular circumstances of plenary class that provoked learners' speaking anxiety, such as an unfamiliar topic, unprepared students for spontaneous talks, a still class and nobody who has the courage to talk, the expressionless face of the lecturers, and students' fear of native speaker lecturers.

To decrease the learners' speaking anxiety, the lecturers need to use particular strategies. The topic spoken in the class should be familiar and interesting so that the learners have the interest to speak. To avoid a silent class, the lecturers should have an asking strategy such as asking with an easy question form, reformulating the question, and giving some examples to the learners. In the end, the lecturers have to appreciate and help the learners by giving good attention with a calm and smiling face to avoid learners' nervousness. It is actually fine to be taught by a native speaker, it would be even more useful, but the native speaker lecturer has to find strategies that could decrease the learners' fear of foreigner, e.g. come closer to the learners' lives in learning context, be humble, and use interesting methods such as games that could enhance the learners' motivation and interest, so that the atmosphere of the class will be fun.

## References

- Ansari, M. S. (2015). Speaking anxiety in ESL/EFL classrooms: A holistic approach and practical study. *International Journal of Educational Investigations*, 2(4), 38–46.
- Arnaiz, P., & Guillén, F. (2012). Foreign language anxiety in a Spanish university setting: Interpersonal differences. *Revista de Psicodidáctica*, 17(1), 5–26.
- Awan, R.-N., Azher, M., Anwar, M. N., & Naz, A. (2010). An investigation of foreign language classroom anxiety and its relationship with students' achievement. *Journal of College Teaching & Learning*, 7(11), 33–40.
- Barahmeh, M. (2013). Measuring speaking anxiety among speech communication course students at the Arab American University of Jenin (AAUJ). *European Social Sciences Research Journal*, 1(3), 229–248.
- Bashir, M., Azeem, M., Ashiq, & Dogar, H. (2011). Factor effecting students' English speaking skills. *British Journal of Arts and Social Sciences*, 2(1), 34–50.
- Basic, L. (2011). *Speaking anxiety: An obstacle to second language learning?* Gävle: University of Gävle.
- Cansrina, G. (2015). Ursachen von sprech- angst im DaF-Unterricht - Ergebnisse einer untersuchung von Indonesischen studentInnen an der Universitas Padjadjaran. *Jurnal Ilmiah Babasa, Sastra, Dan Budaya Jerman*, 2, 168–186.
- Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, CA: SAGE Publications.
- Department of German Letters. (2016). *Katalog jurusan Sastra Jerman*. Malang: Fakultas Sastra Universitas Negeri Malang.
- Dulay, H. C., Burt, M. K., & Krashen, S. (1982). *Language two*. New York, NY: Oxford University Press.
- El-Sakka, S. M. F. (2016). Self-regulated strategy instruction for developing speaking proficiency and reducing speaking anxiety of Egyptian university students. *English Language Teaching*, 9(12), 22–33. <https://doi.org/10.5539/elt.v9n12p22>
- Fischer, S., & Modena. (2005). Sprech- motivation und sprech angst im DaF- Unterricht. *German as A Foreign Language GFL*, 3, 31–45.
- Gemeinsamer Europäischer Referenzrahmen (GER). (2004). *Gemeinsamer europäischer referenzrahmen für sprachen: Kurzinformationen*. Langenscheidt: Landesverlag, Linz.
- Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch*. Langenscheidt: Klett.
- Gnjidić, V. (2016). *L2 English and L3 German vocabulary learning strategies*. Zagreb.
- Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, 21, 112–126. <https://doi.org/10.1017/S0267190501000071>
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom

- anxiety. *The Modern Language Journal*, 70(2), 125–132. <https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- Inozemtseva, N. (2017). *Sprechangst internationaler studierender in der fremdsprache Deutsch*. Essen: Universität Duisburg-Essen Fakultät für Geisteswissenschaften Institut für Deutsch als Zweit- und Fremdsprache.
- Kiper, H., Meyer, H., & Topsch, W. (2002). *Einführung in die Schulpädagogik*. Oldenburg: Cornelsen.
- Lightbown, P. M., & Spada, N. (2006). *How languages are learned* (3rd ed.). Oxford: Oxford University.
- Marwan, A. (2007). Investigating students' foreign language anxiety. *Malaysian Journal of ELT Research*, 3(1), 37–55.
- McLaren, N., Madrid, D., & Bueno, A. (2006). *TEFL in secondary education*. Granada: Universidad de Granada.
- Modern Language Division. (2001). *Common European framework of reference for language: Learning, teaching, assessment*. Strasbourg: Cambridge University Press.
- Muhaisen, M. S., & Al-Haq, F. A.-A. (2012). An investigation of the relationship between anxiety and foreign language learning among 2nd secondary students in Second Amman Directorate of Education. *International Journal of Humanities and Social Science*, 2(6), 226–240.
- Nazir, M., Bashir, S., & Raja, Z. B. (2014). A study of second language speaking-anxiety among ESL intermediate Pakistani learners. *International Journal of English and Education*, 3(3), 216–229.
- Nuhn, H.-E. (2000). Die sozialformen des unterrichts. *Pädagogik (Weinheim)*, 52(2), 10–13.
- Saville-Troike, M. (2006). *Introducing second language acquisition*. Cambridge: Cambridge University Press.
- Sevinç, Y., & Backus, A. (2017). Anxiety, language use and linguistic competence in an immigrant context: A vicious circle? *International Journal of Bilingual Education and Bilingualism*, 1–19. <https://doi.org/10.1080/13670050.2017.1306021>
- Shabani, D. B., Carr, J. E., Pabico, R. S., Sala, A. P., Lam, W. Y., & Oberg, T. L. (2013). The effects of functional analysis test sessions on subsequent rates of problem behavior in the natural environment. *Behavioral Interventions*, 28(1), 40–47. <https://doi.org/10.1002/bin.1352>
- Spielberger, C. D. (1983). *Manual for the state-trait anxiety inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Spradley, J. P. (1980). *Participant observation* (1st ed.). New York, NY: Holt, Rinehart and Winston.
- Sugiyono. (2012). *Metode penelitian pendidikan: Penelitian kuantitatif, kualitatif, dan R&D*. Bandung: Alfabeta.
- Tseng, S.-F. (2012). The factors cause language anxiety for ESL/EFL learners in learning speaking. *WHAMPOA - An Interdisciplinary Journal*, 63, 75–90.
- Von Würde, R. (2003). Students' perspectives on foreign language anxiety. *Inquiry*, 8(1), 1–15.
- Wijayati, H. W. (1995). Analisis data penelitian etnografi. *Forum Penelitian Kependidikan*, 7(1), 32–47.
- Wilson, J. T. S. (2006). *Anxiety in learning English as a foreign language: Its association with student variables, with overall proficiency, and with performance on an oral test*. (Doctoral Thesis). Universidad de Granada, Granada, Spain.
- Zhiping, D., & Paramasivam, S. (2013). Anxiety of speaking English in class among international students in a Malaysian university. *International Journal of Education and Research*, 1(11), 1–16.



## Technology-enhanced pre-instructional peer assessment: Exploring students' perceptions in a Statistical Methods course

Yosep Dwi Kristanto

Department of Mathematics Education, Universitas Sanata Dharma  
Paingan, Maguwoharjo, Depok, Sleman, Yogyakarta 55282, Indonesia  
E-mail: [yosepdwikristanto@usd.ac.id](mailto:yosepdwikristanto@usd.ac.id)

*Submitted: 25 August 2018 | Revised: 20 November 2018 | Accepted: 22 November 2018*

### Abstract

There has been strong interest among higher education institution in implementing technology-enhanced peer assessment as a tool for enhancing students' learning. However, little is known on how to use the peer assessment system in pre-instructional activities. This study aims to explore how technology-enhanced peer assessment can be embedded into pre-instructional activities to enhance students' learning. Therefore, the present study was an explorative descriptive study that used the qualitative approach to attain the research aim. This study used a questionnaire, students' reflections, and interview in collecting student's perceptions toward the interventions. The results suggest that the technology-enhanced pre-instructional peer assessment helps students to prepare the new content acquisition and become a source of students' motivation in improving their learning performance for the following main body of the lesson. A set of practical suggestions is also proposed for designing and implementing technology-enhanced pre-instructional peer assessment.

**Keywords:** *peer assessment, pre-instructional activities, perceptions, Statistical Methods, higher education*

### Introduction

There has been strong interest among higher education institutions in implementing peer assessment as a tool for enhancing students' learning. Indeed, the growth of computer technology has a significant role in improving peer assessment applications in various educational settings (Yang & Tsai, 2010). It is also the case in mathematics learning. Mathematics education researchers have shown substantial evidence of technology-enhanced peer assessment's benefits on the students' learning (Chen & Tsai, 2009; Peter, 2012; Willey & Gardner, 2010). Specifically, Tanner and Jones (1994) posit that peer assessment helps the students to perform reflection through reviewing the works of others and recalling their own works.

Reflection process through which the students recall their existing mental context is fundamental components in learning (Lee &

Hutchison, 1998; van Woerkom, 2010; Wain, 2017). Therefore, this process of reflection meets the purpose of pre-instructional activities. In the pre-instructional activities, it is expected that students can link their prior knowledge with the new content to be learned (Dick, Carey, & Carey, 2015). For this rationale, it is acceptable to stimulate reflection process by conducting peer assessment in pre-instructional activities. However, little has been shown in the literature that peer assessment is used in pre-instructional activities, though Scott (2017) has utilized the simulated peer assessment in improving numerical problem-solving skills as a prerequisite for learning Biology. The questions and the solutions used in Scott's study were not genuine students' works but were constructed by the researcher. Therefore, the present study tries to shed a light on how to embed technology-enhanced peer assessment into pre-instructional activities to enhance students' learning. This paper

investigates students' perceptions in an attempt to portray students' learning.

### Technology-Enhanced Peer Assessment

In understanding peer assessment, this study refers to the definition proposed by Topping (1998). He defined peer assessment as a process in which student measures the learning achievement of his/her peers. In the process, students have two different roles, namely assessors and assessees. As assessors, they evaluate and, in many cases, provide feedback to the works of their fellow students. In assessees role, they receive marking and feedback for their works and may act upon it.

Recent studies found that peer assessment has positive impacts on the students' learning. Several studies demonstrate that peer assessment can benefit the students in the assessment task, i.e. the quality of assessment they provided (Ashton & Davies, 2015; Gielen & De Wever, 2015; Jones & Alcock, 2014; Patchan, Schunn, & Clark, 2018). Furthermore, peer assessment also has effects on the students' acquisition of knowledge and skills in the core domain. In their study, Hwang, Hung, and Chen (2014) show that peer assessment effectively promotes the students' learning achievement and problem-solving skills. In particular, gaining learning achievement was also shown in Statistics class (Sun, Harris, Walther, & Baiocchi, 2015). One possible rationale of such benefits of peer assessment in the students' learning is the exposure to the works of their peers. When the students view their peers' works, they compare and contrast the works with their alternative solutions. This process of comparing and contrasting has the potential to facilitate students learning (Alfieri, Nokes-Malach, & Schunn, 2013; Reinholz, 2016).

Even though peer assessment has a number of advantages in facilitating learning, it also has several issues. The major concern in peer assessment is its validity as well as reliability (Cho, Schunn, & Wilson, 2006). Topping (1998) found disagreement on the degree of validity and reliability of peer assessment on his review, some studies report high validity and reliability (Haaga, 1993; Stefani,

1994; Strang, 2013), and the others report otherwise (Cheng & Warren, 1999; Mowl & Pain, 1995). However, the issues regarding validity and reliability can be reduced by providing the students with assessment rubrics (Hafner & Hafner, 2003; Jonsson & Svingby, 2007) since it makes expectations and criteria explicit.

Another issue regarding the peer assessment system is about administrative workload (Hanrahan & Isaacs, 2001). When implementing peer assessment in their class, instructors at least should manage the students' submission, assessment, and grading evaluation. Fortunately, these functions can be administered by using technology (Kwok & Ma, 1999). Technology can be used to record and assemble the results of scoring and commentary efficiently. In addition, technology also enables the teacher to provide immediate feedback based on the automated score calculation.

In the spirit of making the most of peer assessment's benefits and addressing its problems, peer feedback can be employed to accompany the peer assessment process. In peer feedback, the students discuss each other regarding performance and standards (Liu & Carless, 2006). They comment or annotate the draft or final assignments of their peers to give advice for the improvement of the assignments. When feedback comes with grading, it can be used to explain and justify the grade. It is also used to pose thought-provoking questions. The presence of the thought-provoking questions can foster the assessees' reflection on their assignments.

### Pre-instructional Activity: Theory and Practice

From the instructional design perspective, Gagné, Briggs, and Wager (1992) posit that an instruction should be designed systematically to affect the students' development. Thus, instructional activities should be designed to facilitate the students' learning. One major component of the activities is pre-instructional activities. The activities are done prior to beginning formal instruction and it is significantly important to motivate the students, inform them the learning objectives, and stimulate recall of prerequisite skills. This study

will not theoretically discuss all of the pre-instructional activities in depth. Instead, it will briefly present the examples of pre-instructional activities that appear in literature.

Pre-instructional activities can be done in different strategies. It also applies to mathematics learning. Loch, Jordan, Lowe, and Mestel (2014), in the Calculus of Variations and Advanced Calculus class, use screencasts to facilitate students in revising the prerequisite knowledge regarding the calculus techniques. Further, some scholars (Jungić, Kaur, Mulholland, & Xin, 2015; Love, Hodge, Corritore, & Ernst, 2015) use peer instruction as a pre-instructional strategy. The lesson introduction also can be done by simply telling the students of the prerequisites or testing them on entry skills (Conner, 2015).

## Method

This study was an explorative descriptive research employing a qualitative approach in exploring how technology-enhanced peer assessment can be embedded into pre-instructional activities to enhance students' learning. The following sections give details of the research's setting, data collection, and also data analysis.

### Research Setting

The research was conducted at a private university in Yogyakarta, Indonesia to investigate students' perceptions of the peer assessment system in Statistical Methods class. The class was conducted in a multimedia laboratory in which students have a computer to assist them in learning statistics. The author was the instructor of the class. The class utilized Exelsa, Moodle-based learning management system developed by the university, for the course administration purpose. In Exelsa, the students can access learning materials, post to a forum, and discuss with their peers about a certain topic, submit their assignments, assess and give feedback to their peers' works. The class was conducted biweekly with 24 meetings of instruction, one meeting of the mid-term exam, and one meeting of the final exam. Each meeting consisted of 100-minute learning activities.

In three out of twenty-four meetings, the class was begun with peer-assessment activity. Therefore, students must submit their assignments before the class started. The assignments used in peer assessment were on the topics of one-way and two-way ANOVA. The assignments were done individually and required Microsoft Excel and SPSS Statistics in processing and analyzing real data given in the problems. The more details of the assignments will be described in the Findings section.

The peer assessment system used in this study was a workshop module (Dooley, 2009) provided by the LMS. The peer assessment takes place during the pre-instructional activities. The peer assessment system has five phases, i.e. setup, submission, assessment, grading evaluation, and reflection phases. In the setup phase, the instructor should set the introduction, provide submission instructions, and create an assessment form. After all of the components are set up, the instructor can activate the submission phase. In this phase, students can submit and edit their assignment. Optionally, they also can give a note on their assignments. However, students can only submit and edit their assignments before the class started.

Right after the class started, the instructor activated the assessment phase. In this phase, each student was assigned randomly to review assignments by their two peers. Thus, each student has two assessors. In reviewing their peers' assignments, students used a rubric to obtain a more objective assessment. The grading strategy used in the peer assessment system is the number of errors through which students grade each criterion by answering yes or no questions and optionally provide comments on the criterion. After all of the assignments were reviewed, the instructor can switch on the grading evaluation phase in which submission and assessment grade of each student were calculated automatically. In the end, students can directly see their score and feedback provided by their peers and reflect on it. The last mentioned is a reflection phase. The peer assessment process can be seen in Figure 1.

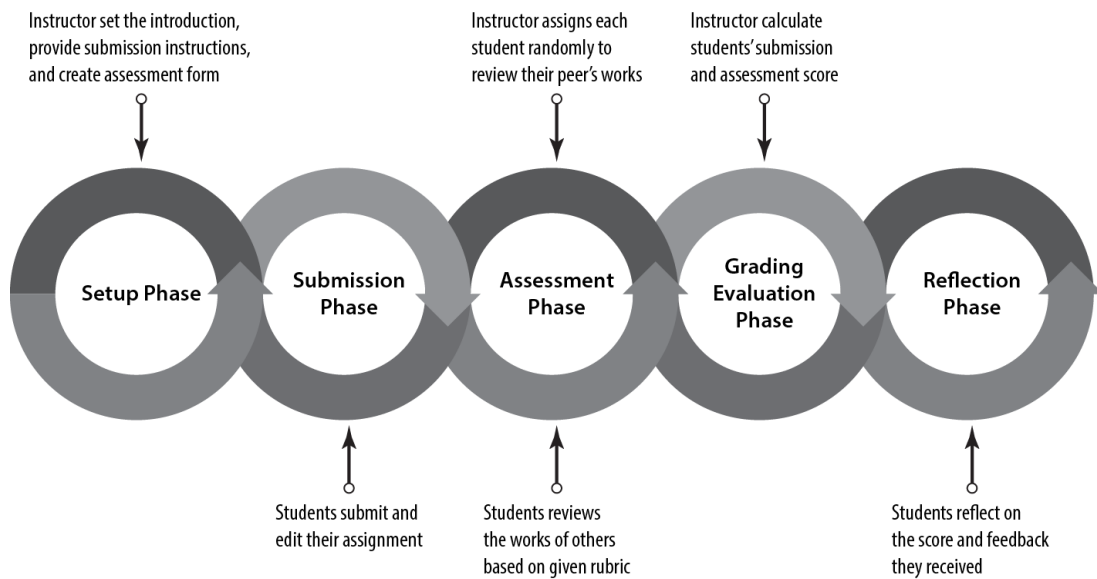


Figure 1. Peer assessment phases

### Data Collection

The data collection process in this study was conducted between May and June 2018 and has been carried out in three phases. In the first phase, the researcher asked the students to write the reflection about their learning experiences in the course. The researcher prompted the students to use Gibbs reflective model (Gibbs, 1988). One learning experience that should be reflected by students was their experience in peer assessment activity. This phase of the data collection process was administered by the LMS.

In the second phase, a questionnaire adapted from Brindley and Scofield (1998) was used to examine students' perceptions on the peer assessment. The questionnaire consists of three sections. The first section asked the students' personal data while the second section asked students' perceptions of peer assessment. The last section invited students to assess how useful the peer assessment process was. The second phase was done in the week right before the final exam and administered by Google form.

The third phase was conducted by interviewing three students on their general opinion about the learning process. The three students were purposively chosen to represent students' achievement. These students were interviewed simultaneously so they feel comfortable since the interviewer was their lectur-

er. The interview was recorded with the approval of these students to prevent data loss.

In addition, logs of three peer assessment activities in the LMS was also generated and downloaded. This logs file records the students' activities in the peer assessment system. Once downloaded, the logs data were then sorted in Microsoft Excel to know the duration of assessment task fulfillment done by each student. Moreover, the data also were used to find total time-frame of the assessment phase in each meeting.

### Data Analysis

Data from the questionnaire and data logs were analyzed using descriptive statistics. Students' response from each item of the questionnaire was described as a proportion or mean value whereas data logs were described as a mean value for each meeting. Data from students' reflection and interview were examined and categorized by the researcher. The categories are derived from Wen and Tsai (2006, pp. 33–34) study, i.e. positive attitude, online attitude, understanding-and-action, and negative attitude. The data were labeled with the corresponding codes and analyzed via the Atlas.ti package program (for more information about conducting qualitative data analysis with Atlas.ti see, Friese, 2014).

## Research Participants

In total, 34 students were enrolled in the author-taught course under study. Student gender demographics consisted of eight male and 26 female students. Most students were in their junior year with only five students from senior year. All of the students were prospective mathematics teachers.

## Findings and Discussion

### Findings

#### *Pre-instructional Activities Profiles*

The three meetings utilized technology-enhanced peer assessment in the pre-instructional activities. At the beginning of each session, the instructor informed the students about the learning objectives that should be achieved and linked the objectives with the previous assignments. The students were then asked to assess their peers' works through LMS. During the peer assessment process, the instructor moved about the classroom, observed students' progress on the assessment task, provided guidance if necessary, and answered questions if they arose. After the peer assessment process was complete, the instructor gave the students the opportunity to reflect on the score and feedback they received. The latter activity was the end of the pre-instructional activities.

The description of the assignments to be submitted before each meeting started is as follows. First meeting required students to submit an assignment on the topic of one-way ANOVA. The assignment asked students to investigate if there is a difference in the mean of football players' height in each position, i.e. forward, midfielder, defender, and goalkeeper. In the assignment, the instructor provided real data obtained from various sources. In this meeting, two students did not submit their assignment and there were also two stu-

dents who submitted their assignment but did not attend the class.

In the second meeting, the students should have submitted a one-way ANOVA problem from the accompanying textbook (Bluman, 2012, p. 632). The problem asked them to determine the effective method in lowering blood pressure by examining the mean of individuals' blood pressure from three samples categorized by the methods they follow. The peer assessment process used in this meeting was slightly different from the previous meeting. In the assessment phase, the students had to assess an example submission provided by the instructor as an assessing practice before they assessed their peers' works. Three students did not submit their assignment in this meeting.

In the third meeting, students should have submitted their assignment for the 'Car Crash Test Measurements' problem from the accompanying textbook (Triola, 2012, p. 643). In this problem, the students were instructed to test for an interaction effect, an effect from car type and car size. One student did not submit their assignment in this meeting and there were also three students who submitted their assignment but did not attend the class.

The mean of assessment tasks carried out by all students in each meeting was calculated and reported in Table 1. On average, the period starting from the assessment phase begins until the assessment phase closes were 43.50 minutes. The table reveals that there has been a sharp decrease in the mean of first and second assessment tasks period carried out by the students in each meeting. In particular, the decreasing trend also applied in the second meeting when the students first reviewed an assessment example. In this meeting, the students reviewed example assessment in nearly a half of an hour (26.83 mins), the first peer's works in almost a quarter of an hour (12.69 mins), and the second peer's works in just over six minutes (6.72 mins).

Table 1. Mean of assessment phase time-frame in minutes

	Example assessment	Assessment 1	Assessment 2	Total
Meeting 1	–	20.03	5.83	39.97
Meeting 2	26.83	12.69	6.72	59.80
Meeting 3	–	10.77	4.83	30.72
	M = 26.83	M = 14.50	M = 5.79	M = 43.50

*Students' Perception*

To investigate the students' perceptions of peer assessment, this study employed both quantitative and qualitative data. The quantitative data were obtained from the questionnaire, while the qualitative data were obtained from the students' reflections, the questionnaire, and interview.

From the questionnaire result, it is reported that most of the students (86.21%) in this study had previous experience on peer assessment. It is also found that approximately three out of four students perceived the necessity of assessing their peers. Further, it is only 27.59% of the students who fully understood the expectation imposed on them when reviewing their peers' works, whereas the rest only have a moderate understanding. In other words, all students understood what others expect on them in assessment tasks.

Four items of the questionnaire were rating-scale questions and used to explore the students' perceived easiness, fairness, pressure, and benefit of peer assessment. A mean

report of the students' responses to the items is shown in Table 2. The students gave a high rating on fairness and responsibility of their marking (M = 4.07) and benefits of peer assessment they receive (M = 4.21). With regard to the grading task, they tend to posit that they have difficulties in assessing their peers' works (M = 3.24). However, they were under moderate pressure when they are doing the assessment task (M = 3.03). The sources of the pressure are various, more than half comes from their role (62.07%), almost a third comes from their experiences (31.03%), and the rest comes from their peers (6.90%).

The students' written reflection and interview are used to examine the students' perceptions as well. The perceptions were grouped into four defined categories and presented in Table 3. The main theme of the students' statements was the helpfulness of peer assessment in enhancing their learning. Regarding this theme, students stated that peer assessment helps them to enable reflective process, viz., reflecting on their mistakes shown by peers as well as reflecting and reviewing their

Table 2. Students' perceptions scale on peer assessment

Question	Mean
How difficult was assessing your peers' work?	3.24
How fair and responsible were you in assessing your peers' work?	4.07
How much pressure did the experience put you under?	3.03
How beneficial was the peer-assessment to you?	4.21

Table 3. Categories of students' perceptions

Category	Code (frequency)
Positive attitude	Helping learning (42)
	Providing Feedback (5)
	Enabling interaction (3)
	Sustainability (3)
	Helping instructor (2)
	Engaging (1)
	Motivating (1)
Online attitude	Anonymity (1)
	Efficiency (1)
	Transparency (1)
Understanding-and-action	Grading strategy (8)
	Action for improvement (7)
	Assessment criteria (2)
Negative attitude	Credibility (15)
	No feedback (6)
	Underestimating self-ability (3)

own works to be compared and contrasted to peers' works. Second, the students perceived the peer assessment process as a tool for knowledge building since they should review their knowledge when assessing others. They added that assessing their peers encouraged them to discuss to their friends if they are indecisive about their assessment. This discussion led them to construct new knowledge to provide marking and feedback on the assessment task. Third, the students thought that peer assessment process develops their evaluative judgment making skills regarding their own works or others when they provide feedback to peers. Finally, the process of reviewing peers' works gives critical understanding and develops higher-level learning skills, such as analyzing and evaluating. The quotations from five students that reflect the benefits of peer assessment with regard to its usefulness in enhancing their learning are given below:

*In my opinion, the peer assessment is useful. (It is) because it encourages me to review my own works if there is a mismatch between my own works and peers. So, (I) learned twice at once regarding the works. (S<sub>6</sub>)*

*... because I don't know (it is right or wrong) ... I ask for help to my friend and found that my insight was improved. (S<sub>15</sub>)*

*This (peer) assessment was good to provide feedbacks to peers' works as well as to be responsible with my marking. (S<sub>31</sub>)*

*(Peer assessment) help us to think critically in assessing friends' works. (S<sub>12</sub>)*

*... we also must evaluate the answer of our friends which indirectly makes us reviewing the topics so that we can know/analyze where the friends' mistakes are. (S<sub>29</sub>)*

Assessment credibility is another major theme of students' perceptions on peer assessment. On one hand, the students agreed that peer assessment gives the instructor other perspectives to provide more accurate grading and timely feedback. On the other hand, the students also questioned their peers' ability in assessing their works. It is possible that their peer assessors made an inaccurate assessment if the assessors' own works were inaccurate since the assessors often referred to it when

undertaking an assessment task. Underrating self-ability also becomes a source of credibility issues. When the students feel incompetence on the subject-specific tasks, they are afraid of not being able to provide appropriate judgments. Reliability is students' next concern on peer assessment. They found that their assessors give different grades on the same item. Hence, they questioned peers' understanding of rubric criteria given by the instructor. The following are the students' statements related to the credibility of peer assessment.

*Peer assessment is very useful as if the instructor makes an error on assessment, it can be remedied by peers' grading. (S<sub>8</sub>)*

*... However, the peer assessment doesn't work optimally when the assessor lacks understanding on what being assessed. (Moreover) the accuracy of each student's assessment is different from one another. (S<sub>34</sub>)*

*... Maybe the assessors' opinions are different from each other, since there are two friends that get different scores although their answers are more or less the same. (S<sub>19</sub>)*

The students thought that feedback is an important component in peer assessment. Corrective feedback provided by peers was helpful for the students to know the errors on their works whereas suggestive feedback useful to make improvements later on. The importance of feedback was also reflected in students' responses when they did not receive feedback. They believe that assessors' task was not only give marking but also provide constructive comments. Some of the students' comments regarding the importance of feedback are as follows.

*The one who said 'no' also comment. It is a constructive thing for us (to know) our mistakes that (the location of) the mistakes are in here, in here, and in here ... There is a friend (that not only) said 'correct' but also give a comment, (you) should write like this and like this. So, that's the positive. It's like constructing (the understanding of) us. (S<sub>15</sub>)*

*Sometimes there is a friend who said that our answer was not correct, but does not give a single comment. That's it. So, we do not know where it goes wrong. (S<sub>24</sub>)*

Other peer assessment aspects did not escape the students' attention. With regard to the number of errors grading strategy, they perceived that it provided not many options in marking peers' works. Instead of answering yes or no in each criterion, they prefer to use scale-rating strategy. However, they thought that the peer assessment process can facilitate students' discussion as well as students-instructor interaction. Other benefits of technology-enhanced peer assessment were also unfolded. Students stated that such assessment model was transparent and efficient as well as engaging and motivating.

## Discussion

The aim of this study was to explore how technology enhanced peer assessment can be embedded into pre-instructional activities to enhance the students' learning. This paper interprets the students' perceptions in an effort to investigate students' learning experiences. In general, the research results show that technology-enhanced peer assessment holds significant promise to be an effective pre-instructional strategy. The learning benefits provided by peer assessment meet the purpose of the pre-instructional strategy.

The findings of the present study show that the process of assessing and commenting on the works of others facilitate the students' learning. This finding is in line with the result of prior studies in peer assessment investigation (Hanrahan & Isaacs, 2001; Sun et al., 2015). One possible explanation of this finding can be derived from comparative thinking perspective (Alfieri et al., 2013; Silver, 2010). When the student reviews peers' works, they compare and contrast it with their own works. If they doubt their own works, they ask for help to others or the instructor. This process of comparing and contrasting helps them to rehearse their own understanding that is useful for preparing them to gain new knowledge related to it.

The findings also suggest that peer assessment stimulates reflective thinking that drives action for improvement. Similar to the results of other studies (Davies & Berrow, 1998; Liu, Lin, Chiu, & Yuan, 2001), the peer assessment process leads the students to think

critically and reflect the quality of their own works compared to the others'. This evaluative process helps the students to devise a plan in improving their learning products later on. As a feedback receiver, the students also take advantages of the feedback to enhance their learning. In other words, peer assessment can become a source of students' motivation in improving their learning performance in the commencing main body of a lesson (Jenkins, 2005).

The study also shows the importance of feedback in students' learning. As a salient element of peer assessment, peer feedback facilitates students in taking an active role in their learning (Liu & Carless, 2006). When the students provide corrective feedback on the peers' works, they develop an objective attitude in conducting their assessment task (Nicol & Macfarlane-Dick, 2006). Through providing suggestive feedback, the students think critically on the drawbacks of their peers' works even when the works are correct (Chi, 1996). As a feedback receiver, the students use peers' comment to improve their works. Moreover, peer's comments are potential to spark cognitive conflict when the comments contradict the student's prior knowledge. From the socio-cognitive perspective, cognitive conflict is fundamental in facilitating students' learning when it is successfully resolved (Nastasi & Clements, 1992).

However, the results of this study also reveal the resistance of peer assessment. Many students in this study have negative attitudes toward the fairness of peer grading. The similar result also can be found in the literature (Cheng & Warren, 1999; Davies, 2000; Liu & Carless, 2006). The negative perceptions come from the students' skepticism about the expertise of their fellow students. Even when a rubric was provided, the students thought that some of their peers were not really fair in giving marking. Another issue arose from grading strategy used in the assessment task. The correct and not-correct dichotomy into which students should categorize their peers' work is considered to be inflexible (Sheatsley, 1983). The students want more flexible grading strategy in order to be more confident in assessing their peers.



Last but not least, the study has several limitations to be considered. The first limitation of the current study relates to its exploratory design in investigating the students' learning experience. Future studies with a larger sample and a longer period are needed to verify the evidence found in this study. Second, this study only focuses on implementing peer assessment. Comparative studies are needed to compare the effectiveness of peer assessment and other strategies, such as advance organizers and overviews, to be used in pre-instructional activities. Finally, design-based studies could contribute to future literature in giving peer assessment design that optimizes the learning transition from lesson introduction to the main body of the lesson.

### Conclusion and Suggestions

The contribution of this study is to show the potential of technology-enhanced peer assessment to be used as pre-instructional activities. The results of the current study, in general, suggest that the technology-enhanced pre-instructional peer assessment helps the students to prepare the new content acquisition for the following lesson. It is also found that peer feedback has a significant role in the peer assessment process in facilitating students' learning.

Based on the findings in the present study, the author proposed a set of suggestions for designing and implementing technology-enhanced pre-instructional peer assessment. First, a training should be provided to students so that they can provide and manage feedback as well as take action upon it effectively. Second, discussions between students and the instructor about assessment criteria are needed in order to improve students' understanding about what to be assessed by their fellow students' works. If necessary, the instructor also can invite students to develop the assessment criteria. Third, the instructor should monitor students' attitude toward grading strategy. This monitoring process aims to know the suitability of the grading strategy to students, tasks, and learning context. Finally, the instructor should use the assignment features (e.g., its content and con-

text) used in peer assessment as a link to the commencing main body of the lesson.

### Acknowledgment

The researcher would like to thank the students who participated in this study and LPPM of Universitas Sanata Dharma that supported this study. In addition, the researcher expresses gratitude to Russasmita Sri Padmi who kindly agreed to edit this manuscript.

### References

- Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, 48(2), 87–113. <https://doi.org/10.1080/00461520.2013.775712>
- Ashton, S., & Davies, R. S. (2015). Using scaffolded rubrics to improve peer assessment in a MOOC writing course. *Distance Education*, 36(3), 312–334. <https://doi.org/10.1080/01587919.2015.1081733>
- Bluman, A. G. (2012). *Elementary statistics: A step by step approach* (8th ed.). New York, NY: McGraw-Hill.
- Brindley, C., & Scofield, S. (1998). Peer assessment in undergraduate programmes. *Teaching in Higher Education*, 3(1), 79–90. <https://doi.org/10.1080/1356215980030106>
- Chen, Y., & Tsai, C. (2009). An educational research course facilitated by online peer assessment. *Innovations in Education and Teaching International*, 46(1), 105–117. <https://doi.org/10.1080/14703290802646297>
- Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education*, 24(3), 301–314. <https://doi.org/10.1080/0260293990240304>
- Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology*, 10(7), 33–49. <https://doi.org/>

- 10.1002/(SICI)1099-0720(199611)10:7  
<33::AID-ACP436>3.0.CO;2-E
- Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891–901. <https://doi.org/10.1037/0022-0663.98.4.891.supp>
- Conner, K. (2015). Investigating diagnostic preassessments. *Mathematics Teacher, 108*(7), 536–542.
- Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Training International, 37*(4), 346–355. <https://doi.org/10.1080/135580000750052955>
- Davies, R., & Berrow, T. (1998). An evaluation of the use of computer supported peer review for developing higher-level skills. *Computers & Education, 30*(1), 111–115.
- Dick, W., Carey, L., & Carey, J. O. (2015). *Systematic design of instruction* (8th ed.). Boston, MA: Pearson.
- Dooley, J. F. (2009). Peer assessments using the moodle workshop tool. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education* (Vol. 41, pp. 344–344). New York, NY: ACM. <https://doi.org/10.1145/1562877.1562985>
- Friese, S. (2014). *Qualitative data analysis with ATLAS.ti* (2nd ed.). London: SAGE.
- Gagné, R. M., Briggs, L. J., & Wager, W. W. (1992). *Principles of instructional design* (4th ed.). Fort Worth, TX: Harcourt Brace College.
- Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. London: FEU.
- Gielen, M., & De Wever, B. (2015). Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning, 31*(5), 435–449. <https://doi.org/10.1111/jcal.12096>
- Haaga, D. A. F. (1993). Peer review of term papers in graduate psychology courses. *Teaching of Psychology, 20*(1), 28–32. [https://doi.org/10.1207/s15328023top2001\\_5](https://doi.org/10.1207/s15328023top2001_5)
- Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. <https://doi.org/10.1080/0950069022000038268>
- Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53–70. <https://doi.org/10.1080/07294360123776>
- Hwang, G.-J., Hung, C.-Ming, & Chen, N.-S. (2014). Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. *Educational Technology Research and Development, 62*(2), 129–145.
- Jenkins, M. (2005). Unfulfilled Promise: Formative assessment using computer-aided assessment. *Learning and Teaching in Higher Education, 1*(1), 67–80.
- Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. <https://doi.org/10.1080/03075079.2013.821974>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. <https://doi.org/10.1016/J.EDUREV.2007.05.002>
- Jungić, V., Kaur, H., Mulholland, J., & Xin, C. (2015). On flipping the classroom in large first year calculus courses. *International Journal of Mathematical Education in Science and Technology, 46*(4), 508–520. <https://doi.org/10.1080/0020739X.2014.990529>
- Kwok, R. C. W., & Ma, J. (1999). Use of a group support system for collaborative assessment. *Computers & Education, 32*(2), 109–125.

- Lee, A. Y., & Hutchison, L. (1998). Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied*, 4(3), 187–210. <https://doi.org/10.1037/1076-898X.4.3.187>
- Liu, E. Z.-F., Lin, S. S. J., Chiu, C.-H., & Yuan, S.-M. (2001). Web-based peer review: The learner as both adapter and reviewer. *IEEE Transactions on Education*, 44(3), 246–251. <https://doi.org/10.1109/13.940995>
- Liu, N.-F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, 11(3), 279–290. <https://doi.org/10.1080/13562510600680582>
- Loch, B., Jordan, C. R., Lowe, T. W., & Mestel, B. D. (2014). Do screencasts help to revise prerequisite mathematics? An investigation of student performance and perception. *International Journal of Mathematical Education in Science and Technology*, 45(2), 256–268. <https://doi.org/10.1080/0020739X.2013.822581>
- Love, B., Hodge, A., Corritore, C., & Ernst, D. C. (2015). Inquiry-based learning and the flipped classroom model. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, 25(8), 745–762. <https://doi.org/10.1080/10511970.2015.1046005>
- Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing: A case study from geography. *Innovations in Education and Training International*, 32(4), 324–335. <https://doi.org/10.1080/1355800950320404>
- Nastasi, B. K., & Clements, D. H. (1992). Social-cognitive behaviors and higher-order thinking in educational computer environments. *Learning and Instruction*, 2(3), 215–238. [https://doi.org/10.1016/0959-4752\(92\)90010-J](https://doi.org/10.1016/0959-4752(92)90010-J)
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
- Patchan, M. M., Schunn, C. D., & Clark, R. J. (2018). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in Higher Education*, 43(12), 2263–2278. <https://doi.org/10.1080/03075079.2017.1320374>
- Peter, E. E. (2012). Critical thinking: Essence for teaching mathematics and mathematics problem solving skills. *African Journal of Mathematics and Computer Science Research*, 5(3), 39–43. <https://doi.org/10.5897/AJMCSR11.161>
- Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, 41(2), 301–315. <https://doi.org/10.1080/02602938.2015.1008982>
- Scott, F. J. (2017). A simulated peer-assessment approach to improve students' performance in numerical problem-solving questions in high school biology. *Journal of Biological Education*, 51(2), 107–122. <https://doi.org/10.1080/00219266.2016.1177571>
- Sheatsley, P. B. (1983). Questionnaire construction and item writing. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 195–230). New York, NY: Academic Press.
- Silver, H. F. (2010). *Compare & contrast: Teaching comparative thinking to strengthen student learning (A strategic teacher PLC guide)*. Alexandria, VA: Association for Supervision & Curriculum Development.
- Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, 19(1), 69–75. <https://doi.org/10.1080/03075079412331382153>
- Strang, K. D. (2013). Determining the consistency of student grading in a Hybrid Business course using a LMS and statistical software. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, 8(2), 58–76. <https://doi.org/10.4018/jwltt.2013040103>

- Sun, D. L., Harris, N., Walther, G., & Baiocchi, M. (2015). Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class. *PLoS ONE*, 10(12), e0143177. <https://doi.org/10.1371/journal.pone.0143177>
- Tanner, H., & Jones, S. (1994). Using peer and self-assessment to develop modeling skills with students aged 11 to 16: A socio-constructive view. *Educational Studies in Mathematics*, 27(4), 413–431. <https://doi.org/10.1007/BF01273381>
- Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3), 249–276. <https://doi.org/10.3102/00346543068003249>
- Triola, M. F. (2012). *Elementary statistics technology update* (11th ed.). Boston, MA: Addison-Wisley.
- van Woerkom, M. (2010). Critical reflection as a rationalistic ideal. *Adult Education Quarterly*, 60(4), 339–356. <https://doi.org/10.1177/0741713609358446>
- Wain, A. (2017). Learning through reflection. *British Journal of Midwifery*, 25(10), 662–666. <https://doi.org/10.12968/bjom.2017.25.10.662>
- Wen, M. L., & Tsai, C.-C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education*, 51(1), 27–44. <https://doi.org/10.1007/s10734-004-6375-8>
- Willey, K., & Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education*, 35(4), 429–443. <https://doi.org/10.1080/03043797.2010.490577>
- Yang, Y.-F., & Tsai, C.-C. (2010). Conceptions of and approaches to learning through online peer assessment. *Learning and Instruction*, 20(1), 72–83.

## Developing assessment model for *bandel* attitudes based on the teachings of Ki Hadjar Dewantara

<sup>1</sup>Restituta Estin Ami Wardani; <sup>2</sup>Supriyoko; <sup>\*3</sup>Yuli Prihatni

<sup>1</sup>SMP Negeri 1 Kalasan

Jl. Jogja-Solo Km.14, Tirtomartani, Kalasan, Sleman, Yogyakarta 55571, Indonesia

<sup>2,3</sup>Universitas Sarjanawiyata Tamansiswa

Tuntungan, Jl. Batikan UH III/1043, Tahunan, Umbulharjo, Yogyakarta 55167, Indonesia

\*Corresponding Author. E-mail: [yuliku7781@gmail.com](mailto:yuliku7781@gmail.com)

Submitted: 09 November 2018 | Revised: 14 November 2018 | Accepted: 22 November 2018

### Abstract

The study was aimed at (1) identifying indicators for *bandel* attitudes in the teachings of Ki Hadjar Dewantara, and (2) finding out the results of the implementation of the developed *bandel* attitude assessment instrument. The study was developmental research in the affective domain using Mardapi's ten developmental steps. Subjects were selected by cluster random sampling of 392 junior secondary school students, 57 for the limited-scale try-out and 335 for the wider-scale try-out. The data analysis techniques included those for Aiken content validity, concurrent validity, and Cronbach Alpha reliability. Data for the instrument implementation were analyzed using descriptive statistics. Findings show that (1) there are six indicators for the *bandel* instrument to be developed in a self-assessment questionnaire format of 24 items consisting of 12 common statements and 12 factual statements; all items are valid and reliable; (2) Students' score in the implementation of the *bandel* assessment instrument is categorized into the very high level.

**Keywords:** *affective assessment, bandel attitude, Ki Hadjar Dewantara*

### Introduction

The true concept of education has been proposed by Ki Hadjar Dewantara (KHD). As Indonesia's Father of Education, KHD maintains that education is an effort to advance the growth of good conducts (inner powers, characters), thinking (intellect), and also body (Dewantara, 2013, pp. 14–15). This can be understood that education is aimed at forming humans who have good conducts, think intellectually, and have a healthy body.

This concept is in conformity with the functions of national education. The national education functions to develop the ability of and form the characters and civilization of the nation in the frame of intellectualizing the life of the nation, developing the potentials of the students to become the persons who believe in and worship God the Omni-One; behave nobly; are healthy, skillful, creative, and inde-

pendent; and become citizens who are democratic and responsible (Law of Republic of Indonesia No. 20 of 2003 on national education system, 2003).

These two educational concepts are too sufficient to develop excellent students. This excellence is not only reflected in the cognitive thinking abilities psychomotor skills but is also shown in the characters of the students. It is therefore important that character education is realized for the development of a great generation as it has been stated by Agboola and Tsai (2012, p. 163) that character education is a discipline to deliberately optimize students' ethical behaviors.

The reality shows that such education functions have not been achieved in as much as education in Indonesia places emphases on the cognitive domain. Students' learning outcomes are also dominated by cognitive aspects. Assessment in the affective aspects re-

lated to feelings and sensibilities has not been done maximally. The learning-teaching processes, therefore, must pay more attention to affective aspects. Olatunji (2013) states that affective learning is related to the learners' attitudes, thoughts, and behaviors in the future. This learning mode is closely related to students' feelings when learning.

Galo (2014) shows the importance of the instrument in the assessment of the affective domain (Setiawan, 2017). The government has taken various steps in the efforts to develop affective evaluation. Two efforts have been revising Curriculum 2006 to become Curriculum 2013 and launching the enforcement of character education (EEC) in 2016.

Learning and evaluation processes are two essential components of the implementation of the Curriculum 2013. Quality learning is one that is able to achieve the basic competencies prescribed by the curriculum. Quality evaluation is able to measure, assess, and evaluate the achievement of the curricular basic competencies. According to Kumaidi (2017) in Setiawan (2017, p. 3), supporting quality learning needs quality assessment.

Ministry of Education and Culture of Republic of Indonesia (2016, pp. 1–2) states that the results of monitoring and evaluation of the implementation of Curriculum 2013 in 2014 found that one of the teachers' difficulties in the junior secondary level was related to evaluation. Approximately, 60% of the respondents reported that they were not able to plan, develop, administer, analyze, report, and even use well the evaluation. The main difficulties were related to formulating indicators, writing the test items, and conducting affective evaluation in various techniques.

Considering these facts, it is important that an instrument package is developed for evaluating students' attitudes. The development of the instrument is focused for the junior secondary school students in relation to 'obstinate' attitude, having strong persistence, perseverance, and unyielding to success. The problems to be addressed are: (1) what are the indicators for developing an instrument to measure obstinance? (2) what is the students' abstinence like as measured by the developed assessment model?

Assessment or evaluation, according to the Regulation of the Minister of Education and Culture of Republic of Indonesia No. 53 of 2015, is the process of gathering data/information about students' learning achievement in aspects of attitudes, knowledge, and skills. Assessment explains an individual's characteristics by accessing the individual's attitudes and mental processes that can be done by observation, interviews, rating scales, checklists, projective techniques, and tests (Aiken, 2003, p. 54).

According to the Ministry of Education, affective evaluation is done to find out the development of the spiritual and social attitudes of the learner (Ministry of Education and Culture of Republic of Indonesia, 2016). Affective evaluation is done to obtain the achievement of the students' spiritual and social values on the levels of receiving, responding, valuing, characterizing, and implementing.

Ministry of Education and Culture of Republic of Indonesia (2017, pp. 8–9) has simplified the 18 character values into five main character values as follows: (1) religiosity, (2) nationalism, (3) autonomy, (4) solidarity, and (5) integrity. Each main value is categorized into several sub-values. For example, autonomy is sub-categorized into work ethos, toughness, perseverance, professionalism, creativity, truth, and long-life education. At present, the whole autonomy sub-categories are important to be planted and enforced in order that students will have persistent struggles to attain education and reach ambitions.

Ki Hadjar Dewantara, a phenomenal avant-garde figure with his mental and intellectual sharpness, has given the quantum leap pillars of educational and cultural concepts. These intellectual investment inheritances become, among others, thoughts of national education and concepts of cultures that last the test of time (Susanto & Retnaningsih, 2018, p. 81). One of his inheritances is the saying '*ngandel-kendel-bandel-kandel*', meaning that a free person who is struggling for independence should be *ngandel* (self-confident), *kendel* (risk-taking, brave), *bandel* (obstinate, not giving up when falling), and *kandel* (immune against negative criticisms) (Soenarno, 2012, p. 35).

The sub-value toughness, perseverance, and hard-working in the EEC are in line with one of Ki Hadjar Dewantara's teachings, that is *bandel*, obstinate. Etymologically, the word *bandel* is originated from Javanese meaning 'strong'. In Indonesian, the word is defined as 'able to bear pain, not easily weep'. The word *bandel* is identical with powerful, unyielding, and resourceful. In the Great Dictionary of the Indonesian Language, the word '*tanggub*' means (1) 'not easily defeated', 'dependable'; (2) 'very strong in self-position'; and (3) 'brave and bearing' (from pain, etc.) (Department of National Education, 2010, p. 1138).

Retno and Haryanto (2016, p. 27) found six indicators for being resourceful, namely (1) spirit of unyielding and not giving up, (2) serious in doing a task to achieve objectives/ambition, (3) discipline, (4) diligent, (5) not afraid of failing, and (6) optimistic. Therefore, the attitude of being tough is realized in working hard, persevering, and not afraid of failing as expressed in the decree of the Ministry of Education and Culture (2017, p. 9) about the enforcement of the affective skills (EEC).

According to Dewantara (2013), the word *bandel* means being obstinate and patient. *Bandel* means not giving up when falling (Soenarno, 2012, p. 35). In *Kamus Besar Bahasa Indonesia*, the word '*tahan uji*' means (1) having the evidence for being strong; (2) willing to be tested (Department of National Education, 2010), while the word '*tawakal*' means (1) giving in to God's wishes; and (2) fully trusting God (in suffering, etc.) (Department of National Education, 2010, p. 1150).

Ki Hadjar Dewantara sees moral education is of utmost importance. Moral education is all the parents do to support the advancement of their child's life, in the sense of improving the growths of all potentials, mentally and physically, of their children (Soenarno, 2014, p. 15). By having good behaviors, every person will be able to stand as an independent person, who can instruct and control his self.

The development of this assessment instrument for measuring attitudes can be used by the teacher and students in the class. The teacher will be able to carry out his jobs easily and correctly. In addition, the students will be able to do self-evaluation honestly and easily.

The study is aimed at: first, obtaining accurate indicators as a basis for developing the *bandel* assessment model following Ki Hadjar Dewantara's teachings; and second, finding out the results of the implementation of the *bandel* attitude as measured by the developed assessment instrument.

## Method

The study is developmental research, a research to develop a product and evaluate the effectiveness of the product (Sugiyono, 2010, p. 407). The model of the development is one suggested by Mardapi (2008, pp. 109–120), consisting of (1) determining the instrument specification (2) writing the items, (3) determining the scale of the instrument, (4) deciding on the scoring system, (5) reviewing the instrument, (6) conducting try-outs, (7) analysing the items, (8) packaging the instrument, (9) administering the test, and (10) analyzing the results of the test.

The design for the try-out was constructed through theoretical reviews on education, the *bandel* obstinate attitude as one of KHD's teachings, and assessment according to the Regulation of the Minister of Education and Culture of Republic of Indonesia No. 23 of 2016. Initial observation was also done on the assessment instrument so far used by the teacher. Based on these reviews and observation, an initial instrument draft was constructed.

The initial instrument draft consisted of formulations of operational definitions, indicators, questionnaire items, and measurement scales. The initial draft was subjected to consultation with the advisors. The next step was the validation of the contents by experts and practitioners by using the Aiken approach. This was conducted by giving out the initial draft to the experts for quantitative evaluation. The aim of the validation was to know whether or not the instrument had decent validity measure so that it could be used for the next steps.

The next step was conducting a limited-scale try-out (readability) involving 57 students. The results of the limited-scale try-out, as empirical validation I, was used as a basis for the instrument revision. The revised in-

strument was then administered to a sample of 335 students from seven junior secondary schools in the district area of Kalasan from the total of 6,200 students as empirical validation II. The sampling was cluster random sampling which was done by using the Krecjie and Morgan table as the basis.

The construct of the *bandel* instrument was developed from analyses of Ki Hadjar Dewantara's theories and subjected to the expert judgment. The data analyses technique of the content validity by the experts was that of item validity indexing suggested by Aiken (Kumaidi, 2014, p. 4; Setiawan, 2017, p. 36). Estimation for the non-test instrument was conducted using the Cronbach's Alpha formula  $>0.700$  (Nunnally Jr., 1981, p. 245). Finally, the instrument was subjected to a concurrent validity analysis. The results of the test administration were analyzed descriptively using Excel and SPSS 17.0 on the computer.

## Findings and Discussion

### Findings

#### *Results of the Instrument Development*

##### Initial Draft

The *bandel* assessment instrument has been constructed by using relevant theories of effective assessment from Ki Hadjar Dewantara's teachings as the basis. A focus group discussion (FGD) was conducted to obtain a picture of the existing affective assessment instrument that is so far used by teachers. The results of the FGD was used in the writing of the instrument items.

Subsequently, the developmental steps for the instrument development were carried out as outlined by Mardapi (2008). Step 1 through Step 5 were carried out, started with the construction of the instrument specification up to the review of the instrument. The table of the specification was developed from the theories and concepts of the term *bandel* based on Ki Hadjar Dewantara's teachings. The results were subjected to the experts' judgment to produce six factors, namely: (1) hard-working (2) enthusiasm, (3) patience, (4) diligence, (5) unyielding, and (6) perseverance. These six indicators were then developed into

item indicators of the *bandel* assessment model consisting of 12 items of statements and 12 items of facts. After being subjected to initial reviews, a revision was made generally on the sharpening of terms for the indicators, replacing inappropriate vocabulary words, and fixing ambiguous statements.

##### Content Validity

The item statements from the initial draft were subjected to consultation to four experts. The four experts were one of the Tamansiswa knowledge, one of educational psychology, one of educational evaluation, and one of instrument assessment for validation in terms of the match between the items and the indicators. Two practitioners were also asked to validate the first draft; these were a guidance-counseling teacher and an Indonesian teacher. The Aiken approach was used. The fit between the 24 item statements and six instrument indicators was represented by the Aiken indexes. All of the Aiken indexes are above 0.750 as seen in Table 1.

Table 1. Aiken indexes for the fit between statement items and instrument indicators of *bandel* attitudes

No.	Indicator	Item	Aiken Index
1.	Hard-working	V1.p	0.944
		V1.n	0.833
		F1.p	0.833
		F1.n	0.944
2.	Enthusiasm	V2.p	0.833
		V2.n	0.944
		F2.p	1.000
		F2.n	0.889
3.	Patience	V3.p	1.000
		V3.n	0.944
		F3.p	1.000
		F3.n	0.889
4.	Diligence	V4.p	1.000
		V4.n	0.889
		F4.p	1.000
		F4.n	0.778
5.	Unyielding	V5.p	0.889
		V5.n	0.833
		F5.p	1.000
		F5.n	0.778
6.	Perseverance	V6.p	0.944
		V6.n	1.000
		F6.p	0.778
		F6.n	0.778



From the Aiken Indexes in Table 1, it can be stated that all the items are in good category.

Limited-Scale Try-out (Empirical Validation I)

After it was known that all the items were at the good category, the readability was conducted. The result of the limited-scale validation is also called empirical validity I. The limited-scale validation was done involving 57 students of Grades VII, VIII, and IX of junior high schools in the Kalasan district. The results of the try-out can be seen in Table 2.

Table 2. Results of the readability test

No.	Criteria	Understanding		Ease	
		Total	%	Total	%
1.	Good	45	78.95	48	84.21
2.	Medium	7	12.28	6	10.53
3.	Poor	5	8.77	3	5.26
Total		57	100	57	100

According to Table 2, most students, 45 students (78.95%), are able to understand the instrument items up to above 75%. The ease aspect of reading the instrument was responded by 48 students (84.21%). This shows that the instrument is good to be used although it undergoes revision in word choice and terms as suggested by students.

Table 3. Results of revision

Item No.	Before	After
2	Student continues to practice until he can really he can do the test correctly.	Student continues to practice until he really can do the test correctly.
4	Student only studies when there will be an exam.	Student will study when there is an exam.
9	There is a tendency for students to play with the cellphone rather than to study.	Students prefers playing with the cellphone to studying.
16	I don't like to study lesson material which is very difficult.	I don't want to study when the lesson material is difficult.
23	I don't want to do home assignment that is hard and difficult.	I only do easy home assignment.

Finally, the setting of the instrument was conducted for the wider-scale try-out.

Wider-scale Try-out (Empirical Validation II)

The wider-scale try-out was conducted in seven junior secondary schools in Kalasan involving 335 students. The result is called empirical validation II. The results show that 24 items were valid, consisting of 12 common statements and 12 factual statements. For the reliability estimation, the Cronbach's alpha was used and it was found that the reliability value of the *bandel* instrument was 0.850. This means that the instrument is reliable since its reliability coefficient is > 0.70. The results of the reliability checks can be seen in Table 4.

Table 4. Results of the estimation of the instrument reliability

Cronbach's Alpha	N of Items
.850	24

Concurrent Validity

Concurrent validity shows how each subject fits in groups which are conceptually different in terms of the treatment or decision that will be taken. In other words, the test of concurrent validity is to know whether or not there is consistency between attitudes and behaviors (Haryanto, 1994, p. 46). The results of the test for concurrent validity can be seen in Table 5.

Table 5. Matches between common statements and factual statements

Indicator	Item	Item No.	r
Hard-working	1. 4	13. 16	0.179*
Enthusiasm	3. 10	15. 22	0.124**
Patience	7. 12	19. 24	0.143**
Diligence	5. 9	17. 21	0.129*
Unyielding	6. 8	18. 20	0.360**
Perseverance	2. 11	14. 23	0.578**

Results of the Bandel Instrument Implementation

Implementation of the use of the *bandel* assessment instrument was conducted on 335 junior high school students from different areas in the Kalasan district. Because of time limitation, and considering that Grade IX students were preparing for the practice exam,

school exam, and national exam during April-May 2018, the same subjects were involved twice. This means that students who participated in empirical validation II were simultaneously subjects of the implementation phase.

In other words, the 335 students who took part in the second validation were subjected to the instrument. The results of the assessment were analyzed descriptively using the SPSS 17.0 software program on the computer. Descriptive analysis was also done to each indicator. The results are presented in Table 6.

Table 6. Results of the descriptive analyses of the assessment implementation

Implementation		
N	Valid	335
Mean		80.5552
Median		81.0000
Std. Deviation		6.32662
Minimum		54.00
Maximum		93.00

In Table 6, the mean score of the obstinate attitudes of the junior secondary school students in the district of Kalasan is 80.555. The minimum score is 54.00 and the maximum score is 93.00. The median is 61.00 and the standard deviation is 6.327. Intervals

are plotted for ideal categories using the determined formula. Five levels are found from the calculation, which are categorized as very high (VH), high (H), medium (M), low (L), and very low (VL). These results are represented in Table 7.

Table 7. Ideal categorization

Interval	Category	Absolute Freq.	Relative Freq.
78.00 up to 96.00	Very High (VH)	230	68.65%
66.00 up to 78.00	High (H)	99	29.55%
54.00 up to 66.00	Medium (M)	6	1.79%
42.00 up to 54.00	Low (L)	-	-
42.00 up to 24.00	Very Low (VL)	-	-
Total		335	100%

Table 7 shows that the highest frequency of the results of the *bandel* assessment is of the very (VH) category with 68.65%. Subsequently, the high (H) category has 55% and medium (M) 1.79%. No student has the *bandel* competencies at the low (L) and very low (VL) categories. These results are clearly presented in the format of a diagram in Figure 1.

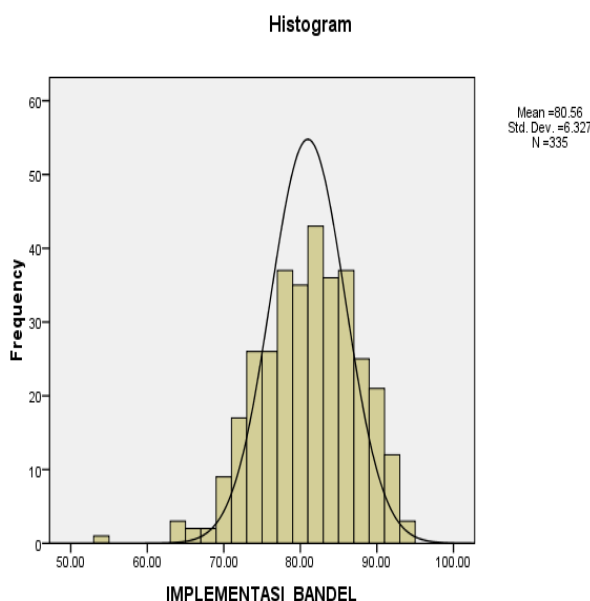


Figure 1. Results of the descriptive analyses of the implementation of the *bandel* assessment

## Discussion

The development of the assessment instrument for the students' *bandel* attitudes is based on the teachings of Ki Hadjar Dewantara. From the theoretical conceptual reviews, six indicators are found for the *bandel* attitudes; namely hard-working, enthusiasm, patience, diligence, unyielding, and perseverance. These six indicators are developed into the specification table of the model assessment. Self-assessment statements are written out of the specification table to be responded honestly by the students.

The resulting assessment model is a questionnaire with 24 items in the forms of 12 common statements (attitudes) and 12 factual statements (behaviors), each consisting of six positive statements and six negative statements. A modified Likert scale is used with response options scored from 1 to 4.

The assessment of the *bandel* attitudes has an expressive function. This means that the common items have a correlation with the factual items, all reflect the attitudes and behaviors of the subject students on *bandel* characteristics.

The first draft of the instrument is subjected to consultations to education experts and Tamansiswa experts. Inputs and suggestions from the experts are used to revise the draft. The result is the construction of an initial instrument assessment for the *bandel* indicators.

The initial items are then subjected to expert judgment for content validity to four experts in educational evaluation and educational psychology and two practitioners (one guidance-counseling teacher and one Indonesian teacher). The results show that all items are at the good category, meaning that are in fit with the indicators, each with an Aiken index of  $> 0.750$ . Nevertheless, minor revisions are made on some of the statements as suggested by the experts and practitioners.

The instrument having been revised, the try-outs are conducted. The first is a limited-scale try-out (readability checks) to 57 grades VII, VIII, and IX students of junior high schools in the Kalasan district taken by random sampling. This is empirical validation

I focusing on readability with two aspects of understanding and ease. The understanding check is to see how far the statements are understood by students, e.g. whether or not they are ambiguous in meaning. Meanwhile, the ease aspect is to see how far the vocabulary words are known and understood by students.

The results show that, out of the 57 students, 45 (78.95%) are able to understand the items more than 75%. The ease aspect is responded by 48 students (84.21%). These results show that the instrument can be understood by the students so that it is feasible to be used for the wider-scale try-out. A minor revision was done, however, in word choices and terms, in accordance with students' feedbacks.

The wider-scale try-out is conducted to 335 students of grades VII, VIII, and IX of the junior secondary schools in the Kalasan district. To the results of this wider try-out, item validity, and reliability are computed. The results of the validity test show that 24 items are valid, consisting of 12 common items and 12 factual items. It can be stated that all the items are valid. They are then subjected to the reliability test. The reliability check produces the score of 0.850 to mean that the instrument is reliable.

The next step is to conduct content validation to see whether or not the instrument items represent the instrument indicators being measured. It is found that all the items do represent the indicators. Subsequently, a concurrent validity check is conducted to see that there is consistency between the attitudes and the behaviors. The results show that there is a correlation in the scores between the common statements and the factual statements, indicating that there is consistency between the attitudes and behaviors.

All validation tests have been done and the results show that the *bandel* assessment instrument is valid and reliable. The instrument has fulfilled the requirements of being a standardized instrument. The last step is done in the form of setting the instrument to become the final version of the instrument, ready to be administered.

The implementation of the *bandel* measurement using the developed product gives

the following results. The mean score is 80.555 which is above 78.00. This can be interpreted that the students' score of the *bandel* attitudes in the academic year of 2017/2018 is in the very high category (VH). The same result is found for the six *bandel* indicators, namely hard-working, enthusiasm, patience, diligence, unyielding, and perseverance, also giving scores of the very high category.

## Conclusion and Suggestions

### Conclusion

Based on the concept of *bandel* attitude in the teachings of Ki Hadjar Dewantara, six indicators can be identified to develop the *bandel* assessment instrument; they are hard-working, enthusiasm, patience, diligence, unyielding, and perseverance. The instrument is developed in the format of a self-assessment questionnaire consisting of 24 statement items (12 common statements and 12 factual statement).

The findings show that the developed instrument is good. Also, the concurrent validation shows that there is consistency between students' attitudes and behaviors. A standardized instrument has been developed to measure the *bandel* attitudes of junior secondary school students which has the characteristics of the very good category.

### Suggestions

It can be suggested to the related parties, especially junior secondary school teachers, to make use of this developed instrument to assess their students' levels of *bandel* attitudes. It is also suggested that teachers understand, have high enthusiasm, and work hard to develop an evaluation instrument for the affective domain so that evaluation results can be obtained for future classroom purposes. As a result, teachers will be able to do their jobs professionally, in accord with the demands of the curriculum and 21st-century educational challenges.

For educational experts and researchers, the results of this study can be used as reference material for producing assessment instruments for other components of the affective domain, especially the five indicators

(EEC) prescribed by the Ministry of Education and other noble values in the teaching of Ki Hadjar Dewantara.

## References

- Agboola, A., & Tsai, K. C. (2012). Bring character education into classroom. *European Journal of Educational Research*, 1(2), 163–170.
- Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Boston, MA: Allyn and Bacon.
- Department of National Education. (2010). *The Great Dictionary of the Indonesian Language (Kamus Besar Bahasa Indonesia)*. Jakarta: Language Center, Department of National Education.
- Dewantara, K. H. (2013). *Pemikiran, konsepsi, keteladanan, sikap merdeka I (Pendidikan)*. Yogyakarta: UST Press & Majelis Luhur Persatuan Tamansiswa.
- Haryanto, S. (1994). *Pengantar teori pengukuran kepribadian*. Surakarta: Sebelas Maret University Press.
- Kumaidi. (2014). Validitas dan pemvalidasian instrumen penilaian karakter. In *Seminar Psikometri Fakultas Psikologi Universitas Muhammadiyah Surakarta*. Surakarta: Universitas Muhammadiyah Surakarta.
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).
- Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekia.
- Ministry of Education and Culture. (2016). *Modul pengembangan instrumen penilaian oleh pendidikan sekolah menengah pertama*. Jakarta: Ministry of Education and Culture of Republic of Indonesia.
- Ministry of Education and Culture. (2017). *Konsep dan pedoman penguatan pendidikan karakter*. Jakarta: Ministry of Education and Culture of Republic of Indonesia.
- Nunnally Jr., J. C. (1981). *Introduction to psychological measurement*. New York, NY: McGraw-Hill.

- Olatunji, M. O. (2013). Teaching and assessing of affective characteristics: A critical missing link in online education. *International Journal on New Trends in Education and Their Implications*, 4(1), 96–107.
- Regulation of the Minister of Education and Culture of Republic of Indonesia No. 23 of 2016 on Educational Assessment Standard (2016).
- Regulation of the Minister of Education and Culture of Republic of Indonesia No. 53 of 2015 on Learning Outcome Assessment by Educators and educator Units on Primary and Secondary Educational Levels (2015).
- Retno, A., & Haryanto, S. (2016). Pengembangan instrumen pengukuran nilai ulet peserta didik SMA di SMA Negeri 1 Buluspesantren. *Wiyata Dharma: Jurnal Penelitian Dan Evaluasi Pendidikan*, 4(3).
- Setiawan, A. (2017). *Pengembangan instrumen penilaian sikap sosial siswa pada pembelajaran tematik sekolah dasar*. Thesis. Universitas Negeri Yogyakarta, Yogyakarta.
- Soenarno, H. (2012). *Ketamansiswaan 1: Riwayat hidup, perjuangan, dan konsepsi*. Yogyakarta: Majelis Luhur Persatuan Tamansiswa.
- Soenarno, H. (2014). *Ketamansiswaan 3: Pendidikan di Tamansiswa*. Yogyakarta: Majelis Luhur Persatuan Tamansiswa.
- Sugiyono. (2010). *Metode penelitian kuantitatif, kualitatif, dan R & D*. Bandung: Alfabeta.
- Susanto, M. R., & Retnaningsih, R. (2018). Melacak pemikiran avant garde Ki Hadjar Dewantara melalui konsep pendidikan nasional sebagai fenomena quantum leap dalam perspektif filsafat organisme. In *Prosiding Seminar Nasional Pendidikan (Vol. 1)*. Yogyakarta: Direktorat Pascasarjana Pendidikan Universitas Sarjanawiyata Tamansiswa.

## Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics

\*<sup>1</sup>Syukrul Hamdi; <sup>2</sup>Iin Aulia Suganda; <sup>3</sup>Nila Hayati

<sup>1,2,3</sup>Universitas Hamzanwadi

Jl. Cut Nyak Dien No.85, Pancor, Selong, Lombok Timur, West Nusa Tenggara 83611, Indonesia

\*Corresponding Author. E-mail: [syukrulhamdi@hamzanwadi.ac.id](mailto:syukrulhamdi@hamzanwadi.ac.id)

*Submitted: 23 November 2018 | Revised: 04 December 2018 | Accepted: 10 December 2018*

### Abstract

The study was aimed at producing a valid and reliable higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts in the junior secondary school mathematics subject matter. The study is developmental research involving a field try-out of 75 students of Grade VIII. Data were analyzed using classical test theories of difficulty levels, discriminating powers, and functioning distractors. The test validity is assessed using the Aiken formula and reliability is estimated by Cronbach Alpha. Findings show that, of the 20 initial multiple-choice items, 15 were valid and reliable and had the characteristics of good test items with a medium-rated difficulty level average of 0.28, a good-rated discriminating power of 0.31, a good-rated reliability coefficient of 0.79, and all distractors well-functioning.

**Keywords:** *test item development, higher-order thinking skill (HOTS), junior secondary school mathematics education*

### Introduction

Twenty-first-century education does not merely provide access to information for students. It is expected to form generations to be able to act effectively in facing the complex and ever changing world's challenges. It must be able to give new experiences, unique and creative ideas, and develop collaborative attitudes as learners' capital to face the world of work, get along with society, and live the daily lives. *The Partnership for 21st Century Skill* (Warisdiono, et al., 2017, p. 18) explains that learning in educational world must focus in developing the 4C's as competencies which must be acquired to face the 21st century: creativity, critical thinking, communication, collaboration. This has had a great influence on the educational curricula in accommodating 21st-century competencies into the school subject matters, including mathematics.

Mathematics is one of the knowledge fields that have central roles in the development of competencies needed to face the 21st century environments. Mathematics understanding is a readiness centre for the young generation to live in modern society. A proportion of the growth of problems and situations exposed in daily lives, including in the professional contexts, needs a number of levels of mathematics understanding, mathematics thinking, and mathematics tools. Mathematics is an important tool for the young adults in confronting the issues and problems in the personal, professional, societal, and scientific environments in their daily lives (OECD, 2013 in Kurniati, Harimukti, & Jamil, 2016, p. 143). However, the low level of the learners' mathematics knowledge has attracted the attention of the educators and researchers and has always become hot topics of discussions in society.

A number of international evaluations on mathematics learning reveal that Indonesian students have not shown pleasing reality. Indonesia has 1,095 class hours per year but students' competencies are under the average level, as compared to South Korea that has 903 and Japan with 712, sitting in the high level of the world ranking (Rahmawati, 2016, p. 6). The Indonesian's involvement in the international assessment is how its educational achievement is among other countries in the world. Results of the study of Programme for International Student Assessment (PISA) conducted by Organization for Economic Cooperation and Development (OECD), looking at the thinking abilities of students around the 15 years of age in reading, mathematics, and science since 2000, show that the average score of mathematics literacy of Indonesian children is still under the international standard (Indonesia PISA Center, 2013). The mathematics literacy of Indonesian children is, therefore, low. In the PISA study 2015 that took 540,000 15-year old students from 72 countries, Indonesia was at the 63 rank of the 70 countries being assessed with a mathematics score of 386. The international standard score was 490 (OECD, 2016). This shows that the mathematics literacy average score of Indonesian students is still under the international standard score.

Beside PISA, results from another study, Trend in International Mathematics and Science Study (TIMSS) taken by Indonesia since 1999, reported the same thing. The mathematics competences of grade VIII Indonesian students were low (Scientific Literacy, October 24, 2014). The TIMSS study in 2015 showed that Indonesian students scored 397 out of the international standard 500. Indonesia is still under the average rank, 45 out of 50 countries (Mullis, Martin, Foy, & Arora, 2015). Details of the PISA and TIMSS mathematics ranking of Indonesian students can be seen in Table 1.

This condition of Indonesian education in mathematics is frightening. The PISA and TIMSS studies pointed out that the students lacked logic and reasoning in completing test items that demand the competences of analysis, evaluation, and creation.

Table 1. Mathematics ranking of Indonesian students by PISA and TIMSS

Year	PISA	PISA Score	TIMSS	TIMSS Score
1999 / 2000	39 of 41	367	34 of 38	403
2003	38 of 40	360	35 of 46	411
2006 / 2007	50 of 57	391	36 of 49	397
2009	61 of 65	371	-	-
2011 / 2012	64 of 64	375	38 of 42	386
2015	63 of 70	386	45 of 50	397

(Sources: Indonesia PISA Center, 2013; Mullis, Martin, Foy, & Arora, 2012; Mullis et al., 2015; OECD, 2014, 2016; Scientific Literacy, 2014)

The Director of the National Educational Evaluation Centre (NEEC), Nizam (Krisiandi, 2016), stated that Indonesian students are good at answering questions of the memorization type, but poor at application and reasoning. School learning, from daily quizzes to school exams, has not sharpened students' abilities to reason. Nizam also mentioned that learning through the subject matters must not be directed only to knowledge skills but also to competences. In the 21st century, basic literacy (science, mathematics, reading, and technology) and also competencies of critical, creative, communicative, and collaborative thinking must be mastered.

The NEEC researcher, Rahmawati (Krisiandi, 2016), also stated that the students' competences in higher-order thinking are still weak; students must be habituated with higher-order thinking test items. Teachers are expected to develop test items which deal with higher-order thinking. This is not as easy; yet, teachers need to familiarize themselves with high-order test items, items which are used by TIMSS and PISA. This is in agreement with the National Curriculum 2013 that demands learner competencies to communicate and think critically and creatively. The study by Kurniati et al. (2016, p. 154) had the same tone with the NEEC study by Rahmawati stating that the lack of higher-order thinking skills (HOTS) in the students is caused by the inability of the students to understand the subject-matter material and apply it in daily life.

Revision to the Curriculum 2013 in 2017 requires teachers to make a number of improvements. Among others, one is for the teacher to be creative in integrating literacy, 21st 4C skills (creative, critical, communicative, and collaborative), and HOTS in their classroom instruction (Pedia Pendidikan, 2017). Phol (Kurniati et al., 2016, p. 143) stated that the ability to involve analysis, evaluation, and creativity is a higher-order ability. According to Brookhart (2010, p. 29), the HOTS involves logic and reasoning, analysis, evaluation, creation, problem solving, and also judgment. Further, Hamdi, Kartowagiran, and Haryanto (2018, p. 1) stated that, at the third level, which is high level, students' understanding is characterized by the abilities to work with complex materials such as mathematical thinking and reasoning and communicative, critical, creative, interpretative, reflective, generalizing, and mathematical skills.

The use of HOTS items in tests is able to train students to sharpen their abilities and skills that are in line with the 21st-century demands. Through HOTS-based test items, critical thinking skills (creative thinking and doing, creativity, and self-reliance learning), will be built through practices in solving various daily-life real problems (problem-solving) (Warisdiono, et al., 2017, p. 18).

The elevation of higher-order thinking skills has become a priority in the school mathematics learning. Students of the junior secondary levels must be trained toward higher-order thinking in accordance with their age. This can be done by the teacher by giving test items of the HOTS type. For this, it is not enough for the teacher to merely pick up material from the packaged workbooks; but they need to resource to more weighted materials. The problem faced by teachers is that they have insufficient ability to develop test items of the HOTS type.

At school, many teachers still use test items that tend to test students' memory aspects rather than higher-order thinking skills. The test items are directed more to lower-thinking skills (LOTS) of memorization and understanding. On the other hands, what the students need to face the future demands is HOTS. The development of HOTS in stu-

dents is expected to raise students' ability in problem solving, elevate their self-confidence in mathematics, and improve their learning achievement (Butkowski, et al., 1994 in Budiman & Jailani, 2014, p. 142).

A HOTS test item is given through a stimulus. A stimulus can be derived from the recent global issues such as technology, information, science, education, health, and infrastructure. A stimulus can also be raised from the environment such as cultures. It is a fact, however, that test items in the school books lack the involvement of cultural issues. In fact, peoples like the Japanese, Chinese, Koreans, and others have used cultural issues in their mathematics learning which makes far advanced in all fields. Kurumeh stated that the success of the Japanese and Chinese in mathematics learning is because they use ethnomathematics (Supriadi, Arisetyawan, & Tiurlina, 2016, p. 2).

Various cultural products of the Indonesian ancestors show art creativities that contain mathematics elements. The case is the same with the cultural products of the Lombok Sasak tribes. One example is the shield from Ende used in a traditional dance. It is made of thick buffalo leather with a two-dimensional geometric pattern. Another product is the Sasak house architecture with three-dimensional ornaments. Besides, many traditional clothes of Sasak have geometrical pattern motifs and the traditional wedding ceremonies have statistical elements. One example is presented in Figure 1.



Figure 1. Example of cultural products of Lombok Sasak ethnic (Department of National Education, 2000, p. 21)

According to Acho, Imako, and Uloko (Wulandari & Puspawati, 2016, p. 35), students' memory and learning achievement obtained through cultural-based instruction are higher than those obtained through conventional teaching. Besides, the study by Nur and



Palobo (2017, p. 11) shows that contextual instruction using Lombok local cultures as contexts gives a positive and significant influence on the students' problem-solving abilities in mathematics.

In addition, Curriculum 2013 requires teachers to be able to develop HOTS test items in line with local environments. As such, stimuli for the test items will be attractive since they can be directly observed and accepted by students. Besides, the use of local cultures for HOTS test items will increase students' senses of attachment and ownership towards the local potentials of their place. Linking mathematics with cultures will expectedly help students see the connection and application of mathematics not only with other disciplines of science, but also with real life.

The item format developed in the present study is that of the multiple-choice type. According to the opinions of experts and research results, tests of the multiple-choice format can be used for HOTS (Budiman & Jailani, 2014, p. 142). The procedure suggested for the HOTS items is that of a set of items consisting of an input followed by answer options.

Based on the rationalization added with data and supporting evidence presented, a need is felt on developing HOTS test instruments with local cultural contexts in the mathematics subject matter of the junior secondary school to prepare students to face the 21st century. The valid and reliable test instrument can be used to train students' in attaining HOTS, help teachers in testing students' HOTS, and become a reference source for the development of HOTS test items for other base competencies in the syllabus.

## Method

The study was development research. It applied the seven steps of gathering initial information, planning, development of first draft and expert validation, limited-scale try-out/readability, revision of the first draft, field try-out, and revision of the final product.

Initial information gathering was related to the product to be developed. It was done through theoretical reviews covering needs

analyses, reviews of the concepts and theories concerning HOTS and local cultures, and analyses of the core competencies (CC) and base competencies (BC) of the Semester 2 of Grade VIII of junior secondary mathematics in the Curriculum 2013.

In the planning phase, the design of the developed product was outlined through the steps of defining, formulating the objectives, and designing of the initial product. This consisted of formulating the product specification, determining the objectives, and constructing the table of specification for the HOTS test items using Lombok cultures as the contexts.

The developed HOTS test items were constructed using HOTS indicators and also BC indicators. The HOTS indicators were synthesized from Ennis (Komalasari, 2013, p. 266; Bayer, Ellis, Gokhale, Cotton, & Langrehr (Thebooke, n.d.); Torrance (Lestari & Yudhanegara, 2015, p. 89); Budiman & Jailani, 2014, p. 143). The indicators were (1) identifying and relating relevant information from a problem; (2) making accurate conclusion based on the obtained information; (3) finding consistencies/inconsistencies in the product; (4) evaluating the product against determined criteria/standards; (5) synthesizing ideas/strategies for the problem solution; (6) applying the strategy for the problem solving; and (7) developing new alternatives of the problem solving.

In the development of the initial product, a first draft of the HOTS instrument was developed. This consisted of 20 multiple-choice test items.

The draft was then subjected to validation by the expert team from the Department of Mathematics Education. The objective of this assessment was to see whether or not the developed test was acceptable and feasible to be used. Another purpose was to obtain feedback for the improvement of the draft.

After being validated by the experts, the draft was then subjected to the analysis of the results of the item validation. Data were in the form of scores of the test items by the experts. The analysis used Aiken's V formula to calculate the content validity index of the test items.

The next step was trying out the draft in a limited-scale group. The validated and revised draft was tried out in a group of 15 junior secondary school students. The try-out was done to obtain information concerning the testees' ease measure in reading the items, level of attractiveness of the test, and level of testees' interest in the test. The results were converted into percentages wherein  $\geq 60\%$  means positive.

The following phase was the revision of the first draft based on the results of the limited-scale try-out. After being revised, the product was then subjected to the field try-out. The field try-out was conducted in two Grade VIII classes in two junior secondary schools. These schools were MTs. Muallimin NW Pancor and MTs. NW Pancor. This try-out involved 75 students. The resulting data were analyzed empirically by way of classical test parameters.

The final step was the revision of the product. This was done on the second draft that was tried out in the two schools. An item was accepted as a final product if it fulfilled one of the following criteria: (1) The item satisfied all the requirements of difficulty levels, discriminating powers, and functioning distractors; and (2) Easy and difficult items were accepted if they had a discriminating power of the good/medium category and the placement of the distractors was functioning. The items that were accepted were then re-formatted to become final products verified as HOTS items.

## Findings and Discussion


### Findings

Higher-order thinking skills include critical thinking, creative thinking, and problem solving. Problem solving, seen as the main skill in HOTS, is a skill in critically and effectively managing, combining, or developing information in the form of facts or ideas to solve a problem and make a decision or finding a solution to a hard-to-handle situation. A HOTS item is one that requires the ability to apply higher-level thinking. The item is presented using a stimulus. A stimulus can be sourced from global issues such as techno-

logy, information, science, education, health, and infrastructure. A stimulus can also be obtained from the environment such as cultures.

Lombok is one of the islands in Indonesia which retains various cultures from history in the forms of objects, non-objects, traditional habits, ethics, and arts. The diversity of the cultures can be used as stimuli and integrated into the school learning processes, including mathematics. Inheritances of history, architecture, dances, musical instruments, and others contain mathematics elements. One example is the shield from Ende used in a traditional dance. It is made of thick buffalo leather with a two-dimensional geometric pattern. Another product is the Sasak house architecture with three-dimensional ornaments. Besides, many traditional clothes of Sasak have geometrical pattern motifs and the traditional wedding ceremonies have statistical elements. HOTS items can be integrated with cultural elements. One example of the development of test items based on cultural elements, in this case, Lombok, can be seen in Figure 2.

*Gendang Belek* merupakan musik khas Suku Sasak di Lombok. Beberapa peralatan musik *Gendang Belek* memiliki bentuk yang sama dengan ukuran yang berbeda-beda seperti gambar di bawah ini



**Pencsek**                      **Kenceng**                      **Terumpang**

Panjang jari-jari *Kenceng* sama dengan dua kali jari-jari *Pencsek*. Panjang jari-jari *Terumpang* sama dengan tiga kali jari-jari *Pencsek*. Jika  $L_1$ ,  $L_2$ , dan  $L_3$  berturut-turut menyatakan luas alat musik tradisional *Pencsek*, luas alat musik tradisional *Kenceng*, dan luas alat musik tradisional *Terumpang*. Tunjukkanlah mana pernyataan yang benar di bawah ini!

I.  $L_1 + L_2 > L_3$                       II.  $L_1 + L_2 < L_3$   
III.  $L_1 + L_2 = L_3$

Figure 2. Example of HOTS item using the context of Sasak culture

Translation:

*Gendang Belek* is a music instrument specific to Sasak ethnic in Lombok. Some *gendang belek* have the same form but different sizes, as can be seen in the pictures.

The radius of *Kenceng* is twice as large as that of *Pencek*. The radius of *Terumpang* is the same as thrice as that of *Pencek*. If L1, L2, and L3 consequently states the size of *Pencek*, *Kenceng*, and *Terumpang*, which of the following statements is correct?

Figure 2 shows an example of a HOTS test item using a local cultural context of Lombok with a HOTS indicator of critical thinking (Making an accurate conclusion from the information of a situation/problem). To be able to answer the question, the testee needs to be able to recall and understand factual, conceptual, and procedural material about circles. Then, by doing an analysis of the situation (stimulus), the testee determines the strategy in solving the problem.

Other than the example above, there are other cultural inheritances that can be integrated into mathematics. The use of cultural elements in HOTS test items will be able to elevate students' senses of attachment and ownership towards the local potentials of their place. Linking mathematics with cultures will also help students see the connection and application of mathematics not only with other disciplines of science but also with the real world.

#### Instrument Development

The product of the developmental study is a valid and reliable HOTS test instrument, using the local cultures of Lombok as a context, consisting of multiple-choice test items for junior secondary school mathematics. The instrument development passes two assessment phases. The first phase is to assess the validity of the instrument, conducted by three experts of mathematics education. The second involves a limited-scale try-out with 15 testees and a field try-out in two schools with 75 testees.

Validation by experts is to look at the contents of the initial product and obtain feedbacks for revising the first draft. In the process, the experts are given the table of the specification of the test, the test items, and the evaluation sheets. Data of the experts' evaluation are subjected to the Aiken's V formula to find the content validity coefficient. The results can be seen in Table 2.

Table 2. Results of experts' validation

Item Number	Aiken's V Coefficient	Criteria
1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	0.67-1.00	Good to be used
6	0.33	Need revision/deletion

In Table 2, it can be seen that, out of the 20 test items, 19 are feasible for use and one needs revision or deletion. However, there is a number of items which needs to be revised or deleted following the experts' feedbacks. These include the format of the writing, completeness of the stimulus texts, clearer pictures, and suitability with the junior secondary school level.

The results of the readability check in the limited-scale try-out show that the majority of the students give positive responses towards the test, between 75% and 94%. This is strengthened by positive comments written by some students. Sample statements can be seen in Figure 3. Meanwhile, the difficulty levels of the items can be seen in Table 3.

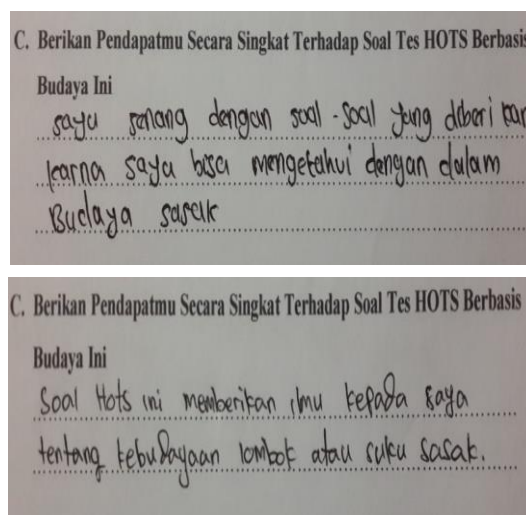


Figure 3. Students' comments on the use of the test instrument

Translation:

Give short opinions about this culture-based HOTS test

*I like the test that is given because I can know about Sasak cultures more deeply*

*This HOTS test give knowledge to me about Lombok cultures or Sasak ethnic*

Table 3. Difficulty levels of the main product test items

Category	Item Number	Total
TK < 0,25 (Difficult)	1, 3, 6, 17	4
0,25 ≤ TK ≤ 0,75 (Medium/Enough)	2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	14
TK ≥ 075 (Easy)	-	0

Table 3 shows that 14 items (77.78%) have the difficulty level in the medium category. Meanwhile, Table 4 shows that seven test items (38.88%) have a discriminating power of the medium category. The spread of the distractors of the main product test items can be seen in Table 5.

Table 4. Discriminating power of the main product test items

Category	Item Number	Total
DP < 0.20 (Poor)	1, 6, 9, 14	4
0.20 ≤ DP < 0.40 (Medium)	4, 12, 13, 15, 16, 17, 18	7
0.40 ≤ DP < 0.70 (Good)	2, 3, 5, 7, 10, 11	6
0.70 ≤ DP ≤ 1.00 (Very Good)	8	1

Table 5. Effectiveness of distractors of the test items

Category	Item Number	Total
Functioning	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	18
Not functioning	-	

In Table 5, it is clear that the distractor distribution of all of the test items is functioning; it means that all the distractors are chosen by 5% of the testees. Based on the results of the analyses of the item characteristics above, the number of items that are accepted and replaced/rejected can be seen in Table 6.

In Table 6, a total of 15 items (83.33%) are accepted and 3 (16.67%) are rejected. The accepted are then reformatted to become the final product test instrument of HOTS in terms of the test validity.

Table 6. Results of analyses of item characteristics

Category	Item Number	Total	Percentage
Accepted	2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18	15	83.33%
Rejected	1, 6, 14	3	16.67%


#### Revision of Final Product

The final product revision is conducted to obtain a test instrument that is valid and reliable. Revision is done by looking at the results of evaluation in the two product try-outs. The revision involves experts' validation, limited-scale try-out, and field try-out.

The experts' validation and product try-outs are used as the main consideration for revision. First, item revision is based on the experts' inputs and suggestions. In general, these include the format of the writing, completeness of the stimulus texts, clearer pictures, and suitability with the junior secondary school level. Figures 4, 5, and 6 show items that are good after revision and items that are rejected.

**Good item after revision**

*Bumbungan* is a traditional house of Sasak Ethnic in Lombok. *Bumbungan* has a steep roof, made from hay with a thickness of about 15 cm. The roof is intentionally let to span to the bottom wall and almost covers the wall. Like the picture below.




The roof bottom is 5.2 in length and the top 5/13 of the roof bottom. Height of the roof is 3/2 of the roof top. The circumference of the roof of the house is ...

A. 12.6 m                      C. 13.6 m  
B. 13.4 m                      D. 14.0 m

Figure 4. Good item after revision

The item in Figure 5 is rejected because the item is not well-formulated and the picture is meaningless (the notes are not clear). Meanwhile, the item in Figure 6 is deleted because it is considered too difficult for junior secondary school age.

**Poor item to be deleted**



**Rudat Dance**

The Lombok-specific *Rudat* dance is usually used to welcome guests, involving 10 dancers. The distance of the most-front dancer and the most-back dancer is ... unit.

A. 3.18                      C. 6.36  
B. 3.56                      D. 10

Figure 5. Poor item that is deleted (1)

**Poor item to be deleted**



**Begasingan Lombok Traditional Game**

In the picture, an arch is made with the center point of *P* and it crosses the line in *Q* point. Then, with the same radius, an arch is made with the center of *Q*, so that it crosses the first arch in *R* point. From the points of *P*, *Q*, and *R*, *PRQ* angle is made. The size of the angle formed by *PRQ* angle is...

A. 30°                      C. 60°  
B. 45°                      D. 75°

Figure 6. Poor Items that are deleted (2)

Second, item revision from the limited-scale try-out is done on the results of the analyses of the item characteristics. Most of the revision deals with discriminating powers and non-functioning distractors.

Third, item revision from the field try-out is done in the same way. All the test items are then verified with the HOTS indicators to make sure that all indicators have been represented. After being verified, all items are reformatted to become the final product of the study.

The field try-out involves 75 students, consisting of 24 from MTs. Muallimin NW Pancor and 51 from Mt. NW Pancor. In general, the achievement of students who take parts in the study can be seen in Figure 7.

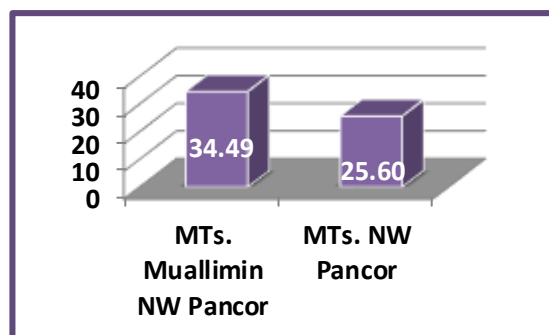


Figure 7. Student achievement profile in mathematics learning viewed from the completed results of the test instrument

Based on Figure 7, it is clear that the average score of students of MTs. Muallimin NW Pancor is higher than that of MTs. NW Pancor. Moreover, a total of 11 students of MTs. Muallimin NW Pancor have scores above the average and 11 have scores below the average. For MTs. NW Pancor, 27 students are above the average score and 24 students are below.

#### Discussion

The product of the study is a valid and reliable HOTS test instrument using Lombok cultures as contexts. It is a fact that, up to the present time, no effort has been done for evidence of test validity and reliability. The development of the instrument begins with the review of HOTS which, according to Brookhart (2010, p. 29), consist of the ability

of logic and reasoning, analysis, evaluation, creation, problem solving, and also decision making (judgment).

It is followed by formulating the item indicators and writing of the test items. Then, the test items are subjected to content validation through expert judgment. This is followed by the Aiken analyses. Before being administered in the field try-out, the items are subjected to a limited-scale try-out for readability. The field try-out involves 75 students from two schools. Finally, item analyses and reliability estimation are conducted. The test instrument development has been conducted following the standard procedure and found that the test is valid and reliable.

The test items developed in the study are those of the multiple-choice type. According to opinions and research results from experts, a multiple-choice test can be used to measure HOTS (Budiman & Jailani, 2014, p. 142). It is suggested that the format of the HOTS test items consist of an introduction followed by response options.

### Conclusion and Suggestions

Based on research findings and discussion, a conclusion is drawn as follows. The final product of the study is a HOTS test instrument using Lombok local cultures as contexts for junior secondary school mathematics consisting of 15 multiple-choice test items with four options. The validity of the test is indicated by the experts' judgment showing that the test is good to be used in the aspects of contents, format, and language. Based on the classical test theories, the instrument fulfills the requirement for reliability shown by a reliability coefficient of 0.79 (good category), with an average score of 0.28 for difficulty levels (medium category), discriminating powers of 0.31 (good category), and functioning distractors.

Based on the conclusion of the study, it is suggested that further research is conducted by analyzing the test items using the IRT as the more modern method. This will expectedly be able to compare the item difficulty levels and the testees' abilities across time and location.

### References

- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria: ASCD.
- Budiman, A., & Jailani. (2014). Pengembangan instrumen asesmen higher order thinking skill (HOTS) pada mata pelajaran matematika SMP kelas VIII semester 1. *Jurnal Riset Pendidikan Matematika*, 1(2), 139–150. <https://doi.org/10.21831/jrpm.v1i2.2671>
- Department of National Education. (2000). *Kain songket Lombok*. Nusa Tenggara Barat: Kantor Wilayah Provinsi Nusa Tenggara Barat Bagian Proyek Pembinaan Permuseuman.
- Hamdi, S., Kartowagiran, B., & Haryanto, H. (2018). Developing a testlet model for mathematics at elementary level. *International Journal of Instruction*, 11(3), 375–390. <https://doi.org/10.12973/iji.2018.11326a>
- Indonesia PISA Center. (2013). *Ranking Indonesia dalam PISA (2000–2012)*. Retrieved February 11, 2018, from [www.Indonesiapisacenter.com/2013/08/ranking-Indonesia-dalam-pisa-2000-2012.html](http://www.Indonesiapisacenter.com/2013/08/ranking-Indonesia-dalam-pisa-2000-2012.html)
- Komalasari, K. (2013). *Pembelajaran kontekstual: Konsep dan aplikasi*. Bandung: PT Rafika Aditama.
- Krisiandi. (2016, December 15). Daya imajinasi siswa lemah. *Kompas*, p. 11. Retrieved from <https://nasional.kompas.com/read/2016/12/15/23091361/daya.imajinasi.siswa.lemah>
- Kurniati, D., Harimukti, R., & Jamil, N. A. (2016). Kemampuan berpikir tingkat tinggi siswa SMP di Kabupaten Jember dalam menyelesaikan soal berstandar PISA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 20(2), 142–155. <https://doi.org/10.21831/pep.v20i2.8058>
- Lestari, K. E., & Yudhanegara, M. R. (2015). *Penelitian pendidikan matematika*. Bandung: PT Rafika Aditama.

- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international result in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2015). *TIMSS 2015 international result in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Nur, A. S., & Palobo, M. (2017). Pengaruh penerapan pendekatan kontekstual berbasis budaya lokal terhadap kemampuan pemecahan masalah matematika. *AKSIOMA: Jurnal Pendidikan Matematika*, 6(1), 1–14.
- OECD. (2014). *PISA 2012 result in focus: What 15 year olds know and what they can do with what they know*. Paris: OECD Publishing.
- OECD. (2016). *PISA 2015 result in focus*. Paris: OECD Publishing.
- Pedia Pendidikan. (2017). *Penjelasan singkat perbedaan RPP K13 Edisi Revisi 2017 dengan RPP K13 Revisi 2016*. Retrieved February 8, 2018, from <http://www.pediapendidikan.com/2017/05/rpp-k13-revisi-2017.html>
- Rahmawati, S. (2016, December 14). Seminar hasil penilaian pendidikan. *Seminar Hasil TIMSS 2015*. Retrieved from [puspen.dik.kemendikbud.go.id/seminar/index.php?folder=hasil\\_seminar\\_puspendik202016](http://puspen.dik.kemendikbud.go.id/seminar/index.php?folder=hasil_seminar_puspendik202016)
- Scientific Literacy. (2014, October 24). *Survei internasional TIMSS (Trends In International Mathematics and Science Study)*. Retrieved February 11, 2018, from [literacyofscientific.blogspot.co.id/2014/10/survei-internasioanl-timms-trends-in.html](http://literacyofscientific.blogspot.co.id/2014/10/survei-internasioanl-timms-trends-in.html).
- Supriadi, S., Arisetyawan, A., & Tiurlina, T. (2016). Mengintegrasikan pembelajaran matematika berbasis budaya Banten pada pendirian SD Laboratorium UPI Kampus Serang. *Mimbar Sekolah Dasar*, 3(1), 1–18. <https://doi.org/10.17509/mimbar-sd.v3i1.2510>
- Thebooke. (n.d.). *Kemampuan berpikir kritis dan kreatif*. Retrieved March 15, 2018, from <http://thebooke.net/do/download-gratis-buku-berpikir-kritis>
- Warisdiono, et al. (2017). *Modul penyusunan higher order thinking skill (HOTS)*. Jakarta: Direktorat Pembinaan SMA, Direktorat Jenderal Pendidikan Dasar dan Menengah Departemen Pendidikan dan Kebudayaan.
- Wulandari, I. G. A. P. A., & Puspawati, K. R. (2016). Budaya dan implikasinya terhadap pembelajaran matematika yang kreatif. *Jurnal Santiaji Pendidikan*, 6(1), 31–37.

## Performance assessment and the factors inhibiting the performance of Buddhist education teachers in the teaching duties

Hesti Sadtyadi

Sekolah Tinggi Agama Buddha Negeri Raden Wijaya  
Jl. Kantil Bulusulur, Wonogiri, Central Java 57615, Indonesia  
E-mail: 15hestisadtyadi@gmail.com

*Submitted: 09 November 2018 | Revised: 27 November 2018 | Accepted: 27 November 2018*

### Abstract

This research aims to examine the components and inhibiting factors of the teacher's work performance in the teaching assignments of Buddhist education teachers. The author believes that the theoretical, as well as the practical problems of Buddhist education teachers, can be solved by examining its components and the inhibiting factors. This research was developmental research that begins by compiling the component of performance instrument and the inhibiting factors instrument through Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). The data were then analyzed through regression analysis. The subjects of research were Buddhist education teachers in Central Java, Indonesia. The results of the research show the values of Anti-image  $> 0.5$  for 7 indicators. Meanwhile, the loading factor was bigger than 0.5 for each component. The model had  $RMSEA \leq 0.05$ , Chi-Square ( $X^2$ )  $> 0.05$ ,  $P = 0.55$ , the GFI was 0.97, which means the model was fit. The teaching performance components of Buddhist education teacher include planning the learning program, implementing the learning program, and evaluating the learning program. The inhibiting factors of the teacher's performance are the additional tasks, the classroom atmosphere, conflict, and work pressure. This research has proved that the inhibiting factors have a negative impact on the teaching performance of Buddhist education teachers.

**Keywords:** *teacher's performance, inhibiting factors, teaching duties*

### Introduction

Teacher takes important roles to establish a good quality of education. Other educational instruments, such as policy, curriculum, learning facilities, or educational technology, were just the supporting devices that will not work properly if the teacher unable to implement their competencies in the classroom. It means the management of the teacher's performance, which contains the elucidation of the teacher's role and responsibilities in the learning activities, becomes the decisive element that determines the teacher's excellence.

The manifestation of teacher's excellence in managing education, as one of the indicators of the teacher's work performance, has been depicted in the study of Suryadarma, Suryahadi, and Sumarto (2005, p. 8). The results of their study upheld the teacher's posi-

tion in influencing the students' progress. Moreover, they also stated that the teacher's position is more important than other factors, for instance, the socioeconomic status and school environment. Their study is supported by the research of Amin and Khan (2009) that showed the importance of the teacher as the key in the education system.

Nevertheless, in particular cases, teacher seems hard to give a maximum impact along with their responsibilities. Sudarwan (2002, p. 168) explains that teachers have not been able to demonstrate their work performance adequately. His research shows that the teacher's performance is not at the optimum level. This is proven by the data of student graduation of 2016/2017 published by the Ministry of Education and Culture (2017), revealing that the average of students drop out is about 1.68 %. The data show that, to a certain degree, the



teacher's work performance has not been optimized in motivating students to study at any level of formal education. The author also believes that these data reflect the current condition of the teachers' performance.

Based on the results of previous research regarding the development of performance instruments for elementary school teachers which was conducted by Sadtyadi and Kartowagiran (2014), the author divided the teachers' performance into two components: the teacher's main tasks and the teacher's functions. These components consist of teaching assignment, educating, guiding, training, and evaluating. Teacher's performance is the result of the works of teachers based on the implementation of their competency and the component of assessment in the form of the main tasks and functions of the teacher.

Considering teachers' the main tasks and functions, the author argues that the performance of Buddhist education teachers is necessary to be assessed to reveal the current condition of their performance level. The research of Sadtyadi (2016), which evaluated the performance of Buddhist education teachers by applying the Teacher Performance Standard model of Stronge and Hindman, has shown the sufficiency of teacher performance based on the evaluation of learning planning, learning implementation, and learning assessment of the teacher. Sadtyadi also suggested assessing the performance of Buddhist education teacher based on the implementation of teachers' performance components.

In terms of performance appraisal conducted by direct assessment to the aspects of teaching assignment, it can afford a self-evaluation for the teacher. Especially if in the assessment process, the teachers are involved in providing self-assessment as the part of self-criticism, self-improvement, and self-evaluation. The process of performance appraisal should not be concerned only to the outcomes, but rather to the effects or the transformation resulted from the process of teacher development. The effects and transformations based on the performance appraisal are expected to be a consideration for the teachers as a fundamental recommendation to elevate their work performance.

The transformation of teacher's work performance can be stimulated by the teacher's introspection of records of accomplishment and various aspects of performance that should be executed. Performance appraisals can also be implemented through peer review or discussion with superiors or colleagues. Precisely, the performance appraisal through peer review and discussion is important as self-assessment and as a medium to evaluate generally the teacher's work performance.

The appraisal of teacher's work performance, particularly for Buddhist education teachers, has a specific characteristic that lies in a number of standards that generally cannot be fulfilled by the teacher, especially related to the standards of infrastructure that are adjacent to the implementation of teaching assignment. The Buddhist education has not entirely used a proper classroom in accordance with the learning schedule. Based on the author's observation, a classroom has been provided for studying Buddhist, but it is not implicitly available. Another problem of Buddhist education is the lack of teaching textbook, because the textbooks and also the teaching material are not sold freely. The textbooks of Buddhist education are sold through certain institutions. In the bookstores, the offered Buddhist books take a specific topic such as Dhamma or meditation, which are commonly used for public readers, and not for Buddhist education students.

Based on the afore-mentioned description, this research is aimed at examining the construction and the inhibiting factors of the Buddhist education teachers' work performance. Theoretically, the assessment of teacher's work performance will be convenient to do by examining its construction. Meanwhile, by knowing the inhibiting factors, the problems of teacher's work performance are easy to overcome, thus, the teacher's performance can be elevated. Practically, this research also tries to arrange the construction as well as the components of teacher's work performance to support the process of teacher improvements, and a reference for policymaking related to improving teacher performance, especially Buddhist education teachers.

## Method

In order to examine the components and inhibiting factors of the teacher's work performance, the methodological frameworks of developmental research model (Borg & Gall, 1983, p. 772) and the Mardapi's developmental stage (Mardapi, 2008, p. 108), were employed as the research method. The subjects of the research were Buddhist education teachers in Central Java Province, Indonesia. In the developmental stage, 52 Buddhist education teachers, consisting of four teachers from Wonogiri Regency, 26 teachers from Semarang Regency, and also 22 teachers from Temanggung Regency, were participated.

The preliminary research was conducted through the literature review, which was used to obtain the related theories and previous research that could be used to support the analysis of research. The further stage was the initial research to adjust the various theories and results of existing research studies, so it would afford a complete study. The initial draft was developed from December 2014 to February 2015. Furthermore, the implementation of the instrument was conducted from August to November 2015, and February to June 2016. At the interval time, deep evaluation, improvements, and readjustments to the instruments were conducted.

The instrument was developed based on the components of teacher's work performance, which are derived from the components of the main tasks and functions of the teacher, specifically the teaching task components. The developed instrument was a non-test instrument, in the form of questionnaires for self-assessment and peer assessment. The developed instrument was generated by considering several points such as instrument specifications based on adequate theoretical studies, grids channel of instruments containing dimensions or components and indicators as well as the number of items from the indicator. The next step was writing instruments along with scaling and scoring systems. Then the review or study of the instrument was carried out. A small-scale trial had been conducted to determine the validity and reliability of the instruments. If the quality of the developed instrument reached the expected cri-

teria, it can be measured against the teacher's work performance. The final step was the interpretation of the measurement to examine the components and the inhibiting factors of teacher's work performance.

### Subject and Setting

The subjects of the initial study conducted at the beginning of December 2014 until February 2015 were Buddhist education teachers and the implementation of the adjusted and developed instrument was conducted from August to November 2015. The first-stage testing was conducted to 52 teachers as the respondents, and the second-stage testing was conducted to 97 teachers as the respondents, carried out during February-June 2016. The research was conducted in three regencies: Wonogiri, Temanggung, and also Semarang. The three regencies are located in Central Java, Indonesia.

### Data, Instrument, and Data Collecting Technique

The data used in this study were mostly quantitative data. The data were collected comprehensively, starting from the instrument arrangement until the product of the instrument and its use, so that the data gained were from religious education teachers, especially Buddhist education teachers. The data were categorized based on arranged instruments: performance assessment instrument and teachers' performance inhibiting factors instrument. The data gained at the initial stage were qualitative data in the form of input from the discussion result and literature study. In addition, the data on the first-stage testing and the second-stage testing were data collected from the use of the instruments of performance assessment and its inhibiting factors, and are in the form of quantitative data.

### Data Analysis Technique

The mixed analysis, consisted of qualitative and quantitative approaches, was used as a data analysis technique. The data of Focus Group Discussion were analyzed qualitatively. Meanwhile, the data regarding the developed instrument were analyzed quantita-

tively. Based on the instruments used, the data were analyzed by factor analysis to obtain appropriate instruments that can be used to compile the instrument of teacher's work performance. The factor analysis was employed through Exploratory Factor Analysis. In the second stage of analysis, the Lisrel program by using Confirmatory Factor Analysis (CFA) for the whole model was implemented to support the analysis. The standard of fitness of the instrument used the criteria proposed by Basuki (2004, p. 12) and Ghozali (2005, p. 325), with the Chi-Square ( $X^2$ ) > 0.05, model fit, RMSEA 0.05 indicating the model fit,  $0.05 < RMSEA \leq 0.08$  shows a reasonable model,  $0.08 < RMSEA \leq 0.1$ , shows sufficient model or (mediocre), and  $RMSEA > 0.1$  indicates a poor fit model. GFI value 9 0.9 is a fitness model. The drafting model began with factor analysis and validity and reliability tests.

Besides, statistical analysis with a qualitative descriptive approach was also employed to analyze the suggestions from Focus Group Discussion and interpret the values from quantitative to qualitative data, so that developed instrument became more valuable. The level of teacher's work performance criteria was determined by a relative scale based on the teacher's qualification. The qualification of teacher's work performance was derived from the literature review, expert suggestions, and author's observation. The qualification can be divided into five categories, such as very good performance, better performance, perform adequately, poor performance, and very poor performance. This technique was also used to conceive the actual performance of the teacher in accordance with the learning objectives, learning methods, and teachers' suitability in the learning process.

#### Instrument Validity and Reliability

Content validity refers to the suitability and readability of the content of the developed instrument with the existing material. The content validity was derived from consideration of fellow researchers, linguist, and Buddhist education teachers. The content validity test aimed to test the readability of the concept and the suitability with the learning objectives.

A factor analysis was employed to test the construct validity of the instrument and find the appropriate composition of the items. As proposed by Kim and Mueller (1986, p. 70), Coakes and Steed (1996, p. 124), Hair, Anderson, Tatham, and Black (2006, p. 129), and Azwar (2013, p. 86), the basic criteria in Stage I refer to the validity of the item and factor loading for each indicator. The items should have a factor loading bigger than 0.3 to be considered for review and revision. Further, in Stage II, the item was tested using CFA, and then the valid items were retained.

The construct validity test was conducted by Exploratory Factor Analysis test with the SPSS program 15.0 for Windows. This program used to determine the correlation between the items, the result by varimax rotation technique, and the factor loading and common factor variance. The reliability test was employed the Cronbach Alpha criteria, with 0.7 for reliability values.

#### Findings and Discussion

##### The Result of the Initial Stage

The initial stage was begun with the literature review to find the previous research and theoretical and empirical studies. Furthermore, the interdisciplinary focus group discussions from the expert of education, linguistics, religious studies, and the Buddhist education teachers were held to obtain the empirical data and recommendations. They participated to measure the validity, reliability, and readability the initial draft of the developed instrument. The empirical data and recommendation were used to revise the initial draft.

The first FGD was designed to examine the content validation that was conducted by analyzing several points such as the texts, language, as well as the suitability with the learning objectives besides of compared with theoretical studies of the instruments. The results of FGD generated the instruments that acceptably with the laws and regulations and the Technical Guidelines for Implementing Teacher's Functional Position and Credit Numbers. In order to examine the inhibiting factors of teacher's work performance, the developed instrument was conducted by con-

sidering Maslach Burnout Inventory (Maslach, Jackson, & Leiter, 1997). Meanwhile, Aiken's V formula also was used to measure the validation. The calculation of content-validity-coefficient was assessed by three Buddhist education teachers, with the results of the test can be concluded that all instruments have high coefficient values, 0.75 to 1. Thus, with limited improvements, the initial design of the developed instrument can be maintained in the next stage.

#### *Test Results of Stage I*

On Stage I, the data were analyzed by considering the factor analysis. The confirmatory approach of the extraction method and the maximum likelihood showed that the instrument was valid and reliable, which was indicated by the values of factor load of each instrument is more than 0.5 that means the items are possible to be used. By noting on the metric component rotated value, which generated a value greater than 0.5, the author indicated the items on the instrument form a certain component in the teaching assignment of Buddhist education teachers. It also forms a component of performance inhibiting instruments. The result of KMO was 0.596, which means that the KMO was miserable.

Table 1 shows the result of the exploratory analysis that shows the work performance of Buddhist education teachers in teaching assignments, includes the teaching programs, the implementation of learning programs, and the evaluation of learning programs. Meanwhile, the components of the inhibiting factors include additional tasks, class atmosphere, conflict, and work pressure.

Table 1. The result of exploratory analysis in anti image

Indicators	Test I	Test II
Planning	.642(a)	.804(a)
Implementing	.607(a)	.714(a)
Evaluating	.553(a)	.659(a)
Additional Task	.619(a)	.687(a)
Classroom	.541(a)	.647(a)
Conflict	.566(a)	.713(a)
Workpressure	.658(a)	.759(a)

Source: The author's processed data

#### *The Development of the Instrument of Work Performance and Its Inhibiting Factors*

The second FGD was the advanced of the previous stage for content validation. The participant of the second FGD reviewed the instruments from the various provisions of the instruments arrangement. The Aiken's V formula was also employed to calculate the content-validity-coefficient of each item in the developed instrument. The test had indicated that all instruments had coefficient values of 0.85 to 1, which means the items had high coefficient values, thus, the instrument was ready to be used.

The result of the second FGD showed that the teacher's work performance in teaching assignments could be prepared by designing an appropriate teaching program, implementing the learning program, and evaluating the learning program. Whereas, the inhibiting factors of teacher's work performance are the additional tasks, classroom atmosphere, the conflict, and the work pressure in accordance with the previous factor analysis.

#### *The Results of Stage II*

#### *Analysis of Validity, Reliability and GOF of the Developed Instrument*

In Stage II, the author held a larger test than the previous stage by analyzing the teacher's work performance of 97 teachers from Wonogiri, Semarang and Temanggung regencies, Central Java, Indonesia. The analysis shows that the data had a normal distribution. In Stage II, the author reiterated the same analysis on the previous stage by testing the validity and reliability of the developed instruments. Several suggestions on Stage II had highlighted same points that the teacher's work performance of the Buddhist education teachers could be explained by three components: planning the teaching programs, implementing learning programs, and evaluating the learning programs. Whereas, the inhibiting factor of the teacher's work performance can be explained into four indicators: the additional tasks, classroom atmosphere, the conflicts, and the work pressure. The model has similarities from the Stage I. The value of Anti-image test was above 0.5. The analysis of

the Rotated Component Matrix generated 7 indicators. Based on the value of each Anti-image, which had a value of more than 0.5 and the loading factor value for each component was more than 0.5, had indicated that each item had no double dimensions.

Table 2. The result of exploratory analysis

Number	Test I	Test II
	Components	
	1	1
Planning	0.708267	0.792478
Implementing	0.732193	0.839202
Evaluating	0.819713	0.883535
	2	2
Additional task	0.739685	0.829552
Classroom	0.614886	0.540466
Conflict	0.720257	0.804357
Work pressure	0.545565	0.651139

Source : The author's processed data

The value of the Cronbach Alpha's showed the reliability value of the developed instrument was 0.8. It means the developed instrument was reliable. Likewise, the value of inhibiting factor instruments has the Cronbach's Alpha 0.7, which indicated that the developed instrument was reliable. The value of KMO was bigger than 0.5, which equal to 0.710. This value indicated that the developed instrument could be enhanced in the next stages. By using EFA, it could be explained

that the loading factors from each indicator are form two components. The first component was called the performance component, composing of three indicators of the teaching assignment components. The second component was the performance barrier, which consists of four indicators that compose from the component of performance inhibiting component.

The CFA analyzed seven indicators that composed from the component of the work performance and the inhibiting factor, with the highly significant loading factor. Thus, the teacher's work performance (*Kinerja*) could be arranged by using three indicators: the teaching programs (*Ajar1*), the implementation of learning programs (*Ajar2*), and the evaluation of learning programs (*Ajar3*). Whereas, four indicators could explain the inhibiting factors: the additional tasks (*TGSTAM*), the classroom atmosphere (*Kelas*), the conflict (*Konflik*), and the work pressure (*TEKKERJA*). Overall, the results of the Stage I was similar with the Stage II.

Confirmatory Factor Analysis (CFA) was employed for the entirety model. The criteria of a fit model showed that the model has  $RMSEA \leq 0.05$ , the Chi-Square ( $X^2$ )  $> 0.05$ ,  $P = 0.55$ , which all indicated that the model was fit. Likewise, the value of GFI was equal to 0.97, which indicated that the model was fit.

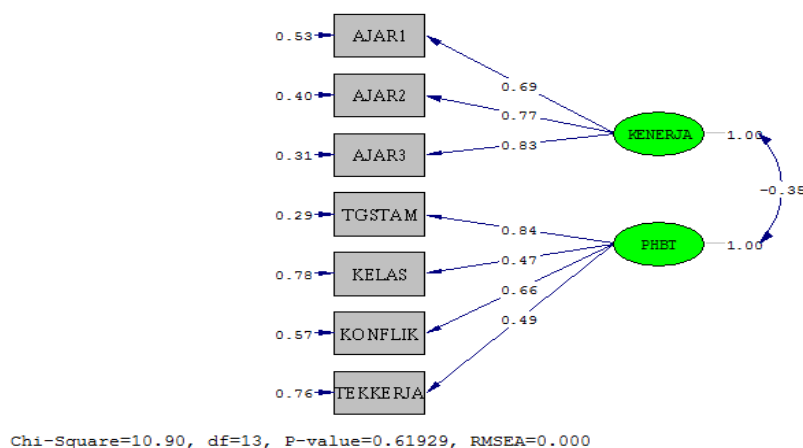


Figure 1. The result of Confirmatory Factor Analysis (CFA)

### The Result of Implementation Stage

Based on the application of performance appraisal instruments and inhibiting factors, it can be illustrated that the Buddhist education teachers, in achieving performance, have significant obstacles, which are around 40%, while 14%, which have few obstacles. The tabulation of the inhibiting factors is illustrated in Figure 2.

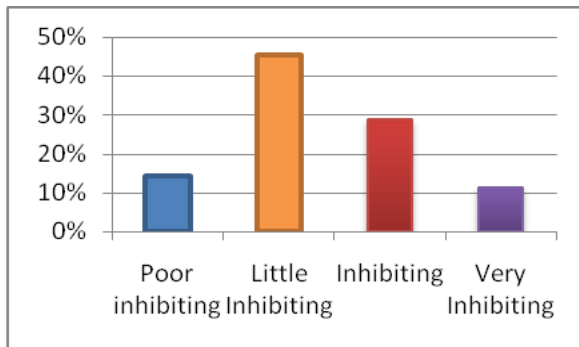


Figure 2. The tabulation of the inhibiting performance factors of Buddhist education teachers

Based on the teaching performance instrument, the performance of Buddhist education teachers is tabulated in Figure 3. The teacher who has good teaching performance ranged from 31 %, the teachers with less performance ranged from 31 %, while the teacher with the poor performance was about 37 %. These results show that there are inhibiting factors in the work performance of Buddhist education teachers.

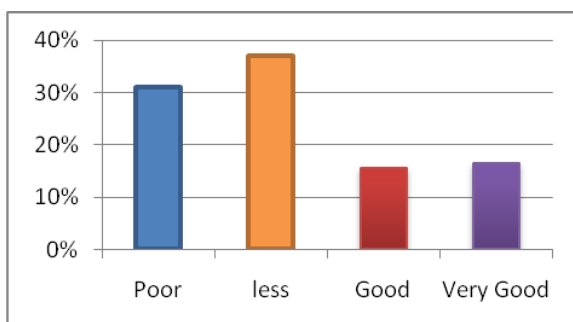


Figure 3. The level of teacher's work performance of Buddhist education teachers

The inhibiting factors of the Buddhist education teachers were analyzed using the regression analysis. The analysis showed the influence of the inhibiting factors on the perfor-

mance of Buddhist education teachers. By using Lisrel program, through the Structural Equation Model (SEM) analysis technique, the numbers of the influence of the inhibiting factors on teacher performance is -0.28 with the t-test indicated that the value of the t-count was greater than the t-table. It means that the regression coefficient was significant, thus, it can be explained that there was a negative influence of the inhibiting factors on the performance of Buddhist education teachers.

From the afore-mentioned description, it is indicated that the performance of Buddhist education teachers in teaching assignments can be prepared through teaching assignment indicators, which include planning the teaching programs, implementing the learning programs, and follow-up the learning programs. Meanwhile, four indicators can explain the inhibiting factor of the teacher's work performance: the additional tasks, classroom atmosphere, the conflict, and the work pressure. The additional task generally referred to as some additional responsibilities outside of the main task area, or not in accordance with their basic competencies, such the administrative tasks. The problem of additional task commonly can be found in the primary school teachers in which the school structure does not have administrative staff. Therefore, the teachers have many additional tasks that potentially can hinder their work performance. The second component is the classroom atmosphere that relates to the infrastructure used by the teacher in the learning process. There are still the shortcomings relate to the supporting media or infrastructure to implement the Buddhist education in the school. The conflict generates a negatively affects for the teacher's work performance. It relates to the importance of communication, work relations, and internal problems of the teachers. Moreover, the component of work pressure relates to many pressures and obstacles that have been faced by the teachers. Some of the Buddhist education teachers teach far away from their hometown. In sum, those inhibiting components can be a barrier to the performance of Buddhist education teachers.

## Conclusion and Suggestions

The components of the teacher's work performance of Buddhist education teachers consist of planning the teaching programs, implementing learning programs, and evaluating learning programs. Meanwhile, the inhibiting factors of the teacher's work performance are the additional tasks, the classroom atmosphere, the conflict, and the work pressure. Those inhibiting factors can negatively affect the work performance of Buddhist education teachers. In order to overcome as well as improve the teacher's work performance, the teacher must be able to reduce the inhibiting factors by noticing each dimension of the inhibitor indicators.

## References

- Amin, H. U., & Khan, A. R. (2009). Acquiring knowledge for evaluation of teachers' performance in higher education - using a questionnaire. (*IJCSIS*) *International Journal of Computer Science and Information Security* (Vol. 2).
- Azwar, S. (2013). *Penyusunan skala psikologi* (2nd ed.). Yogyakarta: Pustaka Pelajar.
- Basuki, H. (2004). *Analisis faktor konfirmatori (Confirmatory Factor Analysis) dalam materi pelatihan SEM (Structural Equation Modeling) angkatan IV, Surabaya*. Surabaya: Lembaga Penelitian Universitas Airlangga.
- Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). New York, NY: Longman.
- Coakes, S. J., & Steed, L. G. (1996). *SPSS version 14.0 for Windows: Analysis without anguish*. Melbourne: Jacaranda Wiley.
- Ghozali, I. (2005). *Structural equation modeling: Teori, konsep, dan aplikasi dengan program Lisrel 8.80*. Semarang: Badan Penerbit Universitas Diponegoro.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Kim, J.-O., & Mueller, C. W. (1986). *Factor analysis: Statistical methods and practical issues*. London: Sage Publications.
- Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekia.
- Maslach, C., Jackson, S. E., & Leiter, M. P. (1997). *Evaluating stress: A book of resources*. (C. P. Zalaquett & R. J. Wood, Eds.) (3rd ed.). Lanham, Md., & London: The Scarecrow Press.
- Ministry of Education and Culture. (2017). *Ikhtisar data pendidikan tahun 2016/2017*. Jakarta. Retrieved from [http://publikasi.data.kemdikbud.go.id/uploadDir/isi\\_FC1DCA36-A9D8-4688-8E5F-0FB5ED1DE869\\_.pdf](http://publikasi.data.kemdikbud.go.id/uploadDir/isi_FC1DCA36-A9D8-4688-8E5F-0FB5ED1DE869_.pdf)
- Sadtyadi, H. (2016). *Evaluasi kinerja guru Pendidikan Agama Buddha dan alumni Prodi Dharmacarya*. Wonogiri: STABN Raden Wijaya.
- 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>
- Sudarwan, D. (2002). *Inovasi pendidikan dalam upaya peningkatan profesionalisme tenaga kependidikan*. Bandung: Pustaka Setia.
- Suryadarma, D., Suryahadi, A., & Sumarto, S. (2005). *Penentu kinerja murid sekolah dasar di Indonesia*. Jakarta: Semeru, Yertas Verja.

## Mapping of physics problem-solving skills of senior high school students using PhysProSS-CAT

\*<sup>1</sup>Edi Istiyono; <sup>2</sup>Wipsar Sunu Brams Dwandaru; <sup>3</sup>Revnika Faizah

<sup>1</sup>Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta

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

<sup>2,3</sup>Department of Physics Education, Universitas Negeri Yogyakarta

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

\*Corresponding Author. E-mail: [edi\\_istiyono@uny.ac.id](mailto:edi_istiyono@uny.ac.id)

*Submitted: 04 December 2018 | Revised: 19 December 2018 | Accepted: 20 December 2018*

### Abstract

Evaluation using computerized adaptive tests (CAT) is an alternative to paper-based tests (PBT). This study was aimed at mapping physics problem-solving skills using PhysProSS-CAT on the basis of the item response theory (IRT). The study was conducted in Sleman Regency, Yogyakarta, involving 156 students of Grade XI of senior high school. Sampling was done using stratified random sampling technique. The results of the study show that the PhysProSS-CAT is able to accurately measure physics problem-solving skills. Students' competences in physics problem solving can be mapped as 6% of the very high category, 4% of the high category, 36% of the medium category, 36% of the low category, and 18% of the very low category. This shows that the majority of the students' competences in physics problem solving lies within the categories of medium and low.

**Keywords:** *assessment, problem-solving skill, CAT*

### Introduction

One of the 21st-century learning and innovation skills is the ability related to critical thinking, problem solving, technology, and information (Daryanto & Karim, 2017). Technology is an integral aspect of the development of a nation. The more advanced the cultures of a nation, the more varied and complicated the technology that is used. Problem solving is a cognitive process directed to the attainment of an objective when there is a solution method to solve a problem (Bueno, 2014). Physics learning highly needs problem-solving skills; it is, therefore, necessary to have an evaluation as one of the efforts in elevating the learners' thinking skills.

Nitko and Brookhart (2011, p. 3) define evaluation as a process to obtain information for making decisions concerning the learners, curriculum, program, school, and educational policy. Evaluation instruments used in learn-

ing covers tests and non-tests (Nitko & Brookhart, 2011). Test-type instruments can be further grouped into objective tests and non-objective tests. Objective tests can be in the form of multiple-choice, short answers, matching, and objective essays. Non-objective tests can be open essays, work performance or observation, and portfolios or project tasks (Mundilarto, 2010, p. 52). Multiple-choice test items can be used to assess learning more complex outcomes which are concerned with the aspects of recall, understanding, application, analysis, synthesis, and also evaluation (Arifin, 2016, p. 138). The administering of the test can be done in two modes: paper-pencil and computer-based test (CBT). The paper-pencil test is paper-based test (PBT) as has been done for long, while CBT is computer-based (Pakpahan, 2016, p. 24).

PBT is based on the assumption that learners with the same level of age and education have the same level of competences. In



reality, there is, however, a significant variation (Bagus, 2012, pp. 45–46). The PBT model has many shortcomings especially related to deviating behaviors, such as frauds, discussions, sharing of answer keys, or even teachers or schools giving out answers keys with the intention that the teachers or schools are not regarded as failing in the running of education and learning by the society (Balan, Sudarmin, & Kustiono, 2017, p. 37). Further, Retnawati (2014, p. 190) states that Indonesia is a big archipelago consisting tens of provinces. As such, distribution of test packages from the centre to the regions faces many obstacles including, for example, during the national examination (NE). This causes, among others, test administration to be impartial and tests results not valid in that they do not represent the real competences of the students. These limitations of PBT can be overcome by testing using the computer.

Computer-based testing has some advantages, including: there is no need to wait for weeks for testees to receive their scores; scores can be obtained immediately. CBT also provides the facility for giving each testee test items that are pre-arranged to give the testee the freedom to select the next test item (Miller, Linn, & Gronlund, 2009, p. 12). According to Luecht and Sireci (2011), the CBT model can be categorized into: (1) computerized fixed tests (CFT); (2) linear-on-the-fly tests (LOFT); (3) computerized adaptive tests (CAT); (4) stratified computerized adaptive tests (AS); (5) content-constrained CAT with shadow tests; (6) test-based CAT and multi-stage computerized mastery tests (combined); and (7) computer-adaptive multistage tests.

Each model has its own advantages and disadvantages. CBT gives more advantages than PBT does in that, among others, its scoring system is automatic and it reduces the burdens on the part of the testees (Riley & Carle, 2012). However, CBT is similar to PBT in that it may not be able to measure the testees' abilities accurately since there is still a potential of fraud in its administration. CBT makes the testees respond to all of the items so that there is inefficiency in the use of time.

There are two theories in assessment that have been empirically and technologically

developed. These are classical test theory (CTT) and item response theory (IRT). Both CTT and IRT widely represent two different frames of assessment. In views of the CTT, scoring of a test is done partially, using the steps that need to be taken in answering a test item correctly. Scoring is conducted step by step, each testee's item score is obtained by summing up the score in each step, and achievement is estimated from raw scores. This scoring model may not be appropriate since the difficulty level of each step is not taken into consideration (Istiyono, Mardapi, & Suparno, 2014, p. 4). In the item level, the CTT model is relatively simple; CTT does not demand a complex theoretical model to relate a testee's success in responding to a test item. On the contrary, CTT collectively considers a group of testees for a particular item. IRT has been developed and important to complement CTT in the design, interpretation, and evaluation of a test or examination. IRT has a strong mathematical basis and relies on a complex algorithm more efficiently calculated on the computer (Adedoyin, 2010, p. 108). IRT supports the use of the computer in educational testing. IRT can be used to provide any item saved in the computer independently, so that the computer select a test from item banks, manage the procedure of the item administering, or design a model for a new computer-based item-response test (Masters & Keeves, 1999, p. 139; van der Linden & Glas, 2003). Thus, a test which uses CAT is highly suitable with the item response theory (IRT).

Hambleton, Swaminathan, and Rogers (1991, p. 9) propose three assumptions underlying the item response theory, including: (1) the chance for answering an item is not dependent on that for another item (local independence), (2) an item measures one competence dimension (unidimensional), and (3) the response pattern of each item can be represented in an item characteristic curve. The weaknesses of the classical theory are tackled up by these three assumptions. Hambleton et al. (1991) identify four limitations of the classical theory. First, item statistics such as difficulty levels and discriminating powers are restricted by specific observed samples that are

obtained; i.e. they depend on the group and test. Second, reliability is defined by parallel-test concepts, which are difficult to realize in practice. This is due to the fact that individuals can never be the same in the second test since they may forget, earn new competences, or have different motivation and anxiety levels. Third, standard errors of measurement are assumed to be the same for all subject matters and variabilities in errors are not being considered. Fourth, the classical theory reflects focus on the test-level information to put item-level information aside. Test-level information is an additive process, that is, the amount of information across the item, and item-level information is the information only for certain items. These limitations show that the classical theory deals with individual score totals and not each testee's competences in the individual level.

A CAT is based on the item response theory. Hambleton and Swaminathan (1985, p. 48), state that there are three types of scoring systems: dichotomous, polytomous, and continuous. Of the three, dichotomous system is the most used in the educational evaluation. The models that can be used for the dichotomous data are latent linearity, perfect scale, latent distance, Ogive one-two-three normal parameter, one-two-three logistic parameter, and four logistic parameter (Barton & Lord, 1981; Guttman, 1944; Lazarsfeld & Henry, 1968; Lord, 1952). The dichotomous model is only suitable for items with two-category scores such as true/false. For items with more than two score categories, the polytomous system is used.

The polytomous scoring system has a number of models, such as nominal response, graded response, partial credit model, and others (Bock, 1972; Geoff N. Masters, 1982; Samejima, 1969). The partial credit model (PCM) has been developed in order to analyze the test items which require multiple-step responses, wherein the items follow the partial credit model patterns so that individuals with higher competences will score higher than those who have lower competences (Istiyono, 2017, p. 2). Therefore, it is reasonable that the partial credit model is used for multiple-choice tests.

A CAT is based on the principles that items must be selected by a consideration that they must measure the testees' competences. Generally, an item is selected in that it gives the most information to estimate the testee's competences. Then, based on the true/false response pattern, the competence level is supposed to return and the item is selected on the basis of the newly estimated competence. These processes are then continued up to a certain precision of the obtained testee's competences (Hambleton & Zaal, 1991). Based on the discussion of these facts, a need is felt on the development of a test that will measure the testees' competences in problem solving. The computerized adaptive test (CAT) has been developed as a CBT alternative to examine PBT tests and provide better tests items and shorter tests in accordance with each test. CAT is a testing system which is more advanced than CBT (Hadi, 2013, p. 12). In accordance with Suyoso, Istiyono, and Subroto (2017), computer-based evaluation is needed more and can help teachers in conducting an evaluation in their subject-matter teaching. In the 21st century, more is emphasized on the higher-order thinking cognitive domain such as HOTS Bloomian, HOTS Marzonian, critical thinking, creative thinking and problem solving (Brookhart, 2010; Heong et al., 2011; Schraw & Robinson, 2011). Testees interact directly with the computer containing the test items of the subject matter. They work on answering test items through the computer as they do in PBT through writing. The number of items is the same that in PBT and item characteristics do not function as they do in CAT (Pakpahan, 2016, pp. 26–27).

The use of CAT does not require items in a great number since the computer is able to give the items in accordance with the testees' competence levels. On the contrary, PBT, which is developed by classical theories, needs items in a great number since it needs to measure the testees' optimum competences repeatedly (Gregory, 2014). According to Weiss (2004, p. 82), CAT is a technology that is viable to have the potentials to give a better assessment, in smaller testing time, for various application in counseling and education. In these two fields, there are needs to measure

individuals' changes. There are so many varieties in the evaluation applications, and one that is able to make use of the superiority of assessment applications which are good and efficient is that which applies the CAT technologies.

## Method

The study was conducted in State Senior High School in Sleman Regency, Yogyakarta Province, during the even semester of the 2017-2018 academic year. The subjects of the study were 156 students of the Physics Department selected by a stratified random sampling technique taking the higher, medium, and lower groups into consideration based on the students' scores of the National Examination in Physics. The size of the sample was determined from the population using the 1-PL formula that ended with 150 to 250 students (Linacre, 2006).

Data collection was conducted by a test that was used to map students' competences in problem solving in the field of physics. The research participants were asked to take the PhysProSS-CAT test which was the product of this research development.

The PhysProSS-CAT consists of items that have undergone development in the forms of multiple-choice items with reasons. The material is related to the balance of solid things, elasticity and Hooke law, static fluid, dynamic fluid, and temperature and calorie. The development of the instrument was based on the Curriculum 2013 which had been revised on the aspects and sub-aspects of problem-solving skills (Ministry of Education and Culture, 2013). The aspects included identification, planning, implementation, and evaluation. The sub-aspects included identifying, differentiating, planning, formulating, sequencing, connecting, applying, checking, and criticizing. The test was developed into four sets of test items, 180 in total with nine anchor items.

The test items had the characteristics that fulfilled the requirements for testing. These requirements were as follows: (a) Based on the results of the content validation by the evaluation experts, the test was content-wise valid with Aiken's V value of 0.97; (b) Based

on the empirical evidence, the test had a fit with the Partial Credit Model (PCM) poly-atomic data with four categories with a mean score and INFIT MNSQ standard deviation of  $1.00 \pm 0.25$ ; (c) Based on the Cronbach Alpha reliability estimation values, all items were regarded as reliable at the measure of 0.93; (d) Based on the levels of difficulty, the test was regarded as good with a range of -1.23 to 1.50; and (e) On the information function and SEM, the test was stated to be able to estimate competences on the range between -2 and 1.6.

The scoring of the test used the partial credit model (PCM) technique which was a development of the 1-PL model and was of the Rash family. Meanwhile, the results of the physics problem-solving test used the computerized adaptive test (CAT) categorized in the form of levels adapted from (Azwar, 2010). The categories are shown in Table 1.

Table 1. Intervals of students' problem-solving skills

No	Skill Interval	Level
1	$Mi + 1.5SBi < \theta$	VeryHigh
2	$Mi + 0.5SBi < \theta \leq Mi + 1.5SBi$	High
3	$Mi + 0.5SBi < \theta \leq Mi - 0.5SBi$	Medium
4	$Mi - 1.5SBi < \theta \leq Mi - 0.5SBi$	Low
5	$\theta < Mi - 1.5SBi$	Very Low

## Findings and Discussion

### Findings

The level of students' competences in problem solving is directly in comparison with the level of item difficulty. The higher the students' theta values, the more difficult the items; the lower the theta, the lower the item difficulty. Students respond to an item whose difficulty level is comparable with their competence level. The first item is one with a medium level of difficulty. If the students answer it correctly, the test will give them a more difficult item; and if they get it wrong, the test will give them a less difficult item. The exposed items have been fitted with the problem-solving aspects, namely identification, planning, implementation, and evaluation. The presentation of an item using CAT can be seen in Figure 1.

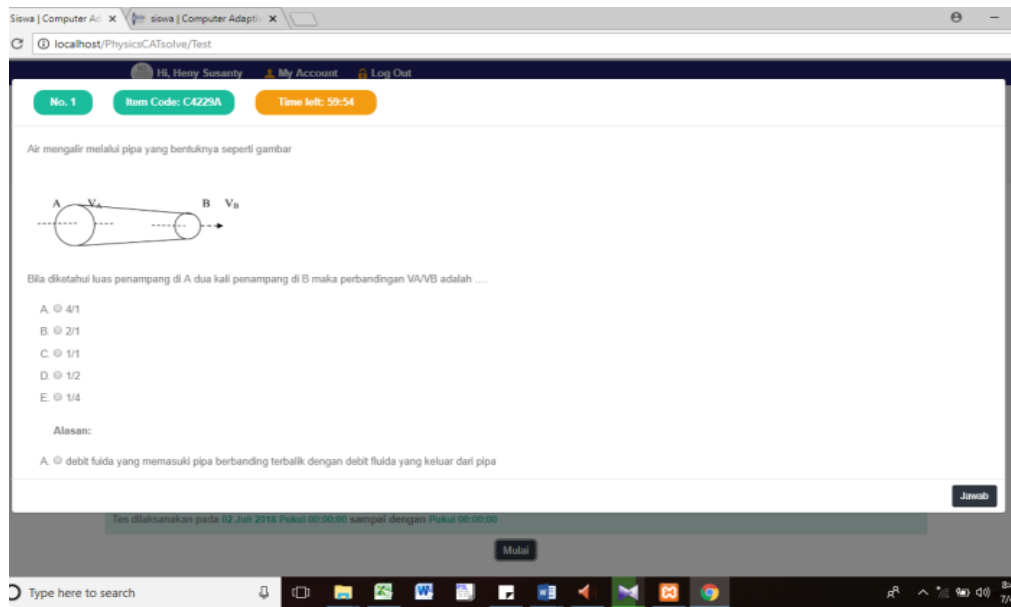


Figure 1. Test item appearance

No	NIS	Nama	Kelas	Guru Pengampu	$\theta$	Soal Dikerjakan	Nilai	Waktu
1	001	Achmad Saifudin	MIPA	Revnika Faizah	0.06	17 Butir	51.00	28 : 55
2	002	Alif Muayyarah	MIPA	Revnika Faizah	0.17	18 Butir	52.83	36 : 22
3	003	Anisa Sholihah Suhartati	MIPA	Revnika Faizah	0	10 Butir	50.00	30 : 45
4	004	Aprina Dwi Hastari	MIPA	Revnika Faizah	0.22	18 Butir	53.67	32 : 52
5	005	Ayu Permata Sholihah	MIPA	Revnika Faizah	0.06	9 Butir	51.00	14 : 00
6	006	Erma Puslita Sari	MIPA	Revnika Faizah	0.17	16 Butir	52.83	26 : 18
7	007	Erwan Sidik Prasista	MIPA	Revnika Faizah	0.22	18 Butir	53.67	32 : 05
8	008	Fadlan Kharisma Aji Nugroho	MIPA	Revnika Faizah	0.09	18 Butir	51.50	03 : 10
9	009	Hestian Agung Prayoga	MIPA	Revnika Faizah	0.17	20 Butir	52.83	28 : 37
10	010	Ibnu Subarkah	MIPA	Revnika Faizah	0.09	15 Butir	51.50	34 : 56
11	011	Ignasia Acrista Cahya Inna	MIPA	Revnika Faizah	0.06	11 Butir	51.00	36 : 07
12	012	Ihsan Ahmad Badrianto	MIPA	Revnika Faizah	0.22	18 Butir	53.67	41 : 50
13	013	Muliyah Nur Hamida	MIPA	Revnika Faizah	0.17	16 Butir	52.83	37 : 13
14	014	Putra Al	MIPA	Revnika Faizah	0.03	16 Butir	53.50	35 : 40

Figure 2. Recapitulation report appearance

In Figure 1, a PhysProSS-CAT test item can be seen in the multiple-choice format with reasons. The testees are asked to select the correct answer and give the reasons for selecting it. After a testee completes the test on the CAT, a recapitulation report from the computer will appear on the screen, as presented in Figure 2.

The recapitulation report can be immediately seen by the administrator, teacher, and student. The administrator can see all the reports of all the test takers. The teacher can see only the reports of his students. The report is in the form of theta scores representing the

students' competences. The students' competence level ( $\theta$ ) is categorized into very high, high, medium, low, or very low in a five-level scale (Azwar, 2010, p. 63) as can be seen in Table 2.

Table 2. Problem-solving skill scale conversion

No	Interval	Competence Level
1	$0.27 \leq \theta$	Very High
2	$0.21 < \theta \leq 0.27$	High
3	$0.16 < \theta \leq 0.21$	Medium
4	$0.10 < \theta \leq 0.16$	Low
5	$\theta \leq 0.10$	Very Low

In Table 3 and Figure 3, of the 156 students taking the CAT test, ten are in the very high category, six are in the high, 56 are in the medium, 56 are in the low, and 28 are in the very low. In percentages, 6% of the students are in the very high category, 4% in the high, 36 % in the medium, 36% in the low, and 18% in the very low. It means that most students' competence levels are in the medium and low categories.

Table 3. Mapping results of competence levels in three state senior high schools

No	Competence Level	Number of Students	Percentage (%)
1	Very High	10	6.41
2	High	6	3.85
3	Medium	56	35.90
4	Low	56	35.90
5	Very Low	28	17.95

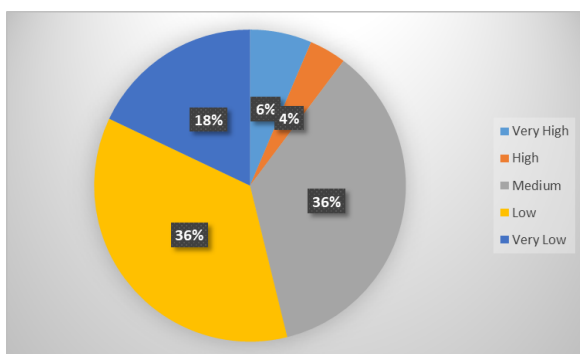


Figure 3. Mapping results of competence levels in three state senior high schools

Mapping is done on the three schools based on the scores which are obtained from the national examination (NE) in Physics, categorized as: high, medium, and low. The results of the mapping are presented in Table 4, Table 5, and Table 6.

Table 4. Mapping of problem-solving competence levels in Senior High School A

No	Competence Category	Number of Students	Percentage (%)
1	Very High	5	7.81
2	High	4	6.25
3	Medium	23	35.94
4	Low	19	29.69
5	Very Low	13	20.31
	Total	64	100.00

Table 5. Mapping of problem-solving competence levels in Senior High School B

No	Competence Category	Number of Students	Percentage (%)
1	Very High	2	3.33
2	High	2	3.33
3	Medium	25	41.67
4	Low	22	36.67
5	Very Low	9	15.00
	Total	Total	100.00

Table 6. Mapping of problem-solving competence levels in Senior High School C

No	Competence Category	Number of Students	Percentage (%)
1	Very High	3	9.38
2	High	2	6.25
3	Medium	9	28.13
4	Low	15	46.88
5	Very Low	3	9.38
	Total	32	100.00

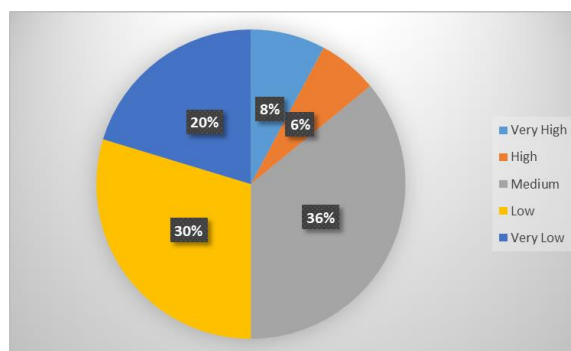


Figure 4. Mapping of problem-solving competence levels in Senior High School A

Shown in Figure 4, in State Senior High School A, of the 64 students, 8% are in the very high category, 6% very high, 36 % medium, 30% low, and 20% very low. It indicates that most students' competence in this school are in the 'medium' category.

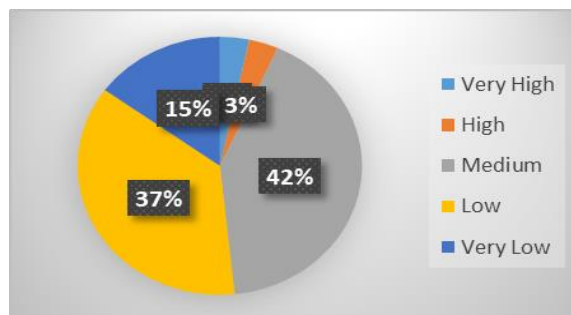


Figure 5. Mapping of problem-solving competence levels in Senior High School B

Based on Figure 5, in State Senior High School B, of the 60 students participating in the study, 3% are in the very high category, 3% very high, 42% medium, 37% low, and 15% very low. It indicates that most students in this school are in the 'medium' category.

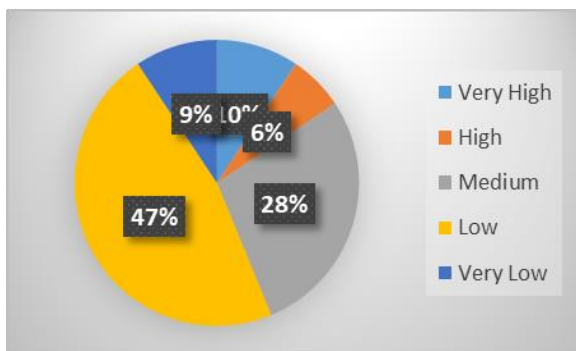


Figure 6. Mapping of problem-solving competence levels in senior high school C

As seen from Figure 6, in State Senior High School C, 32 students participated in the study and 10% of them are in the very high category, 6% very high, 28% medium, 47% low, and 9% very low. It indicates that most students in this school are in the 'low' category.

### Discussions

Based on the findings of the research, it is clear that the PhysProSS-CAT test has been quite well and accurately able to map students' competences in Physics problem solving. The CAT-based instrument has been able to select

the items in accordance with the students' competence levels. In this case, students of School A who are high in the national examination are dominantly in the medium category, but have the highest score in the problem-solving test. In the B school, which is medium in the national examination, the students are dominantly at the medium and low categories. Meanwhile, in School C, with a low level of national examination results, the students are dominantly low in their problem solving competence. This means that mapping has been done well in matching test items with students' levels of competence.

The results of the overall mapping of the 156 students participating in the study show that many of the students are in the medium category. This can be traced from the factors of students' motivation, instructional processes, and evaluation practices. In this relation, only the evaluation factor will be discussed further. Accurate evaluation will be able to support students to learn using higher-order thinking (Istiyono et al., 2014, p. 2). The learning processes and evaluation are supposed to deal with higher-order thinking, including problem solving, in order that the students' skills in problem solving improve. In time, the need is felt to develop evaluation that will be able to measure these students' skills. Ultimately, this will help in realizing students' learning achievements.

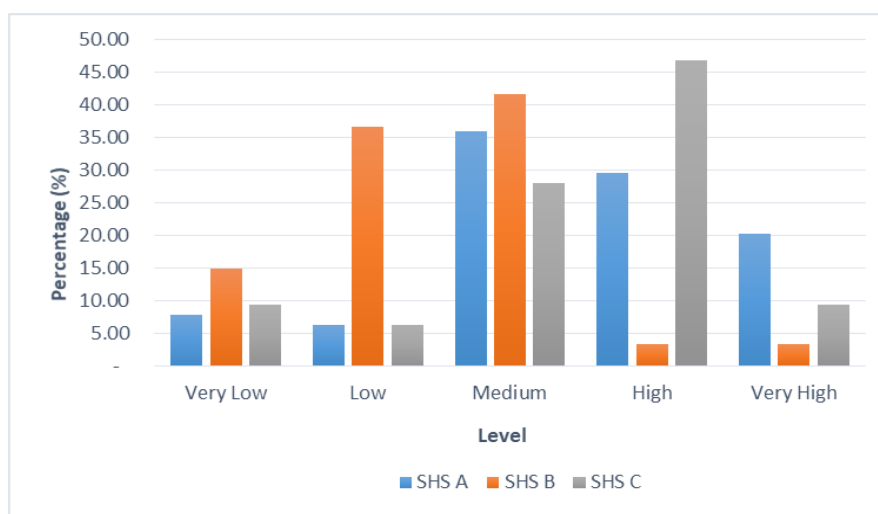


Figure 7. Mapping of the students' problem-solving competences in three schools

As shown in Figure 7, the PhysProSS-CAT test results have categorized students' competences in Physics problem solving into very high, high, medium, low, and very low. Further, the results give accurate information about students' problem-solving skills. The students of State Senior High School A with high national examination scores have the most students with medium problem-solving skills, School B with medium national examination scores has many students with medium problem-solving skills, and School C with low national examination scores has most students with low problem-solving skills.

Meanwhile, Figure 8 presents the recapitulation report of the test results. It consists of scores, test items answered, and time. In the time of the test administration, most students completed 18 to 25 test items, in 35 to 50 minutes out of the total items of 154. The minimum items to be completed were nine items, and the shortest time was 14 minutes. The maximum items completed were 25 and the longest time was 58 minutes. Students did not need to complete all the items but only those within their competences. This is in line with Gregory (2014) stating that CAT testing does not need too many items since, in the computer-based testing, the computer provides test items that are within the range of the testee's competences.

Departing from the weaknesses of the paper-based testing (PBT) mode, in which all testees take all items without considering their skill differences, the computer-based testing

(CBT), on the other hand, is designed using the adaptive mode. In this mode, next items are given on the basis of the testee's competence in completing the previous items (Istiyono, 2013). It is, therefore, reasonable to use the computerized adaptive test (CAT) as an alternative technique for testing since it gives a better estimation result and using a shorter test to be adjusted to the testee's competence. Further, testees do not have to answer all questions, and this saves testing time. In accordance with Huang, Chen, and Wang (2012), the superiority of the CAT over the PBT is that the CAT is able to achieve the same precision with fewer items and shorter time. In CAT, the testee needs only to click on the correct answers until the computer finds and determines his most accurate estimate of his competences to terminate the test and gives his score. CAT is most suitable for such tests for selection and one of a large scale.

The use of PhysProSS-CAT can minimize frauds since testees do different items and have different numbers of items to complete the test; the CAT program gives different items to testees in accordance with their levels of competences. Safety and confidentiality of the items are guarded. On its turn, results of the testing will be reliable. In PBT and CBT testing, chances are abound for frauds to take place for the opposite reasons that testees take the same test with relatively the same items.

No	NIS	Nama	Kelas	Guru Pengampu	θ	Soal Dikerjakan	Nilai	Waktu
1	001	Achmad Saifuldin	MPA	Revnika Faizah	0.06	17 Butir	51.00	28 : 55
2	002	Abi Musyarah	MPA	Revnika Faizah	0.17	18 Butir	52.83	36 : 22
3	003	Anisa Sholihah Subartati	MPA	Revnika Faizah	0	18 Butir	50.00	30 : 45
4	004	Aprina Dwi Hastari	MPA	Revnika Faizah	0.22	18 Butir	53.67	32 : 52
5	005	Ayu Permata Sholihah	MPA	Revnika Faizah	0.06	9 Butir	51.00	14 : 00
6	006	Erma Puspta Sari	MPA	Revnika Faizah	0.17	16 Butir	52.83	26 : 18
7	007	Erwan Sidik Prasista	MPA	Revnika Faizah	0.22	18 Butir	53.67	32 : 05
8	008	Fadlan Kharisma Aj Nugroho	MPA	Revnika Faizah	0.09	18 Butir	51.50	03 : 10
9	009	Hestian Agung Prayoga	MPA	Revnika Faizah	0.17	20 Butir	52.83	28 : 37

Figure 8. Recapitulation report of the PhysProSS-CAT test results

The PhysProSS-CAT can do its testing functions safely, fast, and accurately showing the accurate competences of the testees. For this reason, the test helps much in competence mapping for various purposes. The immediate issuance of the test results helps the teacher map the students' competences in a short time. The teacher can also immediately evaluate and plan for further programs.

In line with the opinions proposed by van der Linden and Glas (2003), a number of reasons for switching to the CAT type are: (1) CAT makes it possible for testees to schedule their own testing in accordance with their preferences; (2) Testing is administered in a comfortable atmosphere with fewer people around than there are in conventional paper-pencil testing; (3) CAT processes the data and gives out the results fast; and (4) Test items and materials are more varied in levels and sizes.

It is possible for teachers to select a test from a variety of choices but testing must be done in accordance with the needs and situations. In a school with adequate facilities for computers, the CAT type testing is more preferable. For the assessment of higher-thinking skills, more specifically, the CAT model is more appropriate since it measures competences accurately and efficiently and saves energy and time of the administration. This is supported by Jiao, Macready, Liu, and Cho (2012) stating that computerized captive testing achieves higher accuracy of the measurement and provides efficient administering of the assessment. In view of the superiority of PhysProSS-CAT, it is suitable for testing individuals' competences in such testing for selection and the final examination. The test saves time, and energy and minimizes frauds.

## Conclusion and Suggestions

### Conclusion

Based on the results of the study, it can be shown that the PhysProSS-CAT is able to accurately map the students' competences in problem solving in the physics field. In percentages, students' competences can be rated as very high (6%), high (4%), medium (36%), low (36%), and very low (18%). This means

that the majority of the students' competences are within the categories of medium and low. On the average, of the total 154 items provided in the test, students complete between 18 and 25 test items in a time range of 35 to 50 minutes. Meanwhile, the minimum number of items responded is 9 and the time needed is 14 minutes; and the maximum number is 25 and the maximum time 58 minutes. Therefore, PhysProSS-CAT is able to map problem-solving competences accurately, efficiently, and saves time and energy.

### Suggestions

In the administering of CATs, including PhysProSS-CAT, it is recommended that administrators provide items with difficulty levels that are more normal in distribution. In relation to the technical facilities, it is suggested that administrators use adequate numbers of items to anticipate troubles in the computer webs since testees access the same items in the same time.

## References

- Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Educational Sciences*, 2(2), 107–113. <https://doi.org/10.1080/09751122.2010.11889987>
- Arifin, Z. (2016). *Evaluasi pembelajaran: Prinsip, teknik, dan prosedur* (8th ed.). Jakarta: Remaja Rosdakarya.
- Azwar, S. (2010). *Metode penelitian*. Yogyakarta: Pustaka Pelajar.
- Bagus, H. C. (2012). The national exam administration by using computerized adaptive testing (CAT) model. *Jurnal Pendidikan Dan Kebudayaan*, 18(1), 45–53.
- Balan, Y. A., Sudarmin, S., & Kustiono, K. (2017). Pengembangan model computer-based test (CBT) berbasis Adobe Flash untuk sekolah menengah kejuruan. *Innovative Journal of Curriculum and Educational Technology*, 6(1), 36–44. <https://doi.org/10.1186/2229-0443-1-3-60>



- Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. ETS Research Report Series (Vol. 1981). Princeton, NJ: John Wiley & Sons. <https://doi.org/10.1002/j.2333-8504.1981.tb01255.x>
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51. <https://doi.org/10.1007/BF02291411>
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.
- Bueno, P. M. (2014). Assessment of achievement in problem-solving skills in a General Chemistry course. *Journal of Technology and Science Education*, 4(4), 260–269. <https://doi.org/10.3926/jotse.100>
- Daryanto, & Karim, S. (2017). *Pembelajaran abad 21*. Yogyakarta: Gava Media.
- Gregory, R. J. (2014). *Psychological testing: History, principles and applications* (7th ed.). Wheaton, IL: Pearson.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9(2), 139–150. <https://doi.org/10.2307/2086306>
- Hadi, H. (2013). *Pengembangan Computerized Adaptive Test berbasis web*. Yogyakarta: Aswaja Pressindo.
- 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.
- Hambleton, R. K., & Zaal, J. N. (1991). *Advances in educational and psychological testing*. Boston, MA: Kluwer Academic.
- Heong, Y. M., Othman, W. B., Yunus, J. B. M., Kiong, T. T., Hassan, R. Bin, & Mohamad, M. M. B. (2011). The level of Marzano higher order thinking skills among technical education students. *International Journal of Social Science and Humanity*, 1(2), 121–125.
- Huang, H.-Y., Chen, P.-H., & Wang, W.-C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement*, 36(8), 689–706. <https://doi.org/10.1177/0146621612459552>
- Istiyono, E. (2013). *Pengembangan instrumen untuk mengukur kemampuan berpikir tingkat tinggi dalam mata pelajaran Fisika di SMA*. Yogyakarta: Department of Physics Education, Universitas Negeri Yogyakarta.
- Istiyono, E. (2017). The analysis of senior high school students' physics HOTS in Bantul District measured using PhysReMChoTHOTS. In *AIP Conference Proceedings* (Vol. 1868, p. 070008). AIP Publishing LLC. <https://doi.org/10.1063/1.4995184>
- Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (Phys-THOTS) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 18(1), 1–12. <https://doi.org/10.21831/pep.v18i1.2120>
- Jiao, H., Macready, G., Liu, J., & Cho, Y. (2012). A mixture Rasch model-based computerized adaptive test for latent class identification. *Applied Psychological Measurement*, 36(6), 469–493. <https://doi.org/10.1177/0146621612450068>
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York, NY: Houghton Mifflin.
- Linacre, J. M. (2006). *WINSTEP: Rasch-model computer programs*. Chicago, IL: Winstep.com.
- Lord, F. (1952). *A theory of test scores*. Richmond, VA: Psychometric Corporation.
- Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing*. New York, NY: The College Board.

- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Masters, G. N., & Keeves, J. P. (1999). *Advances in measurement in educational research and assessment* (1st ed.). Amsterdam: Pergamon.
- Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). The role of measurement and assessment in teaching. In *Measurement and assessment in teaching* (10th ed., pp. 29–31). Upper Saddle River, NJ: Pearson Education.
- Ministry of Education and Culture. (2013). *Pengembangan kurikulum 2013*. Jakarta: Kementerian Pendidikan dan Kebudayaan.
- Mundilarto. (2010). *Penilaian hasil belajar Fisika*. Yogyakarta: Pusat Pengembangan Instruksional Sains (P2IS) Jurdik Fisika FPMIPA UNY.
- Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston, MA: Pearson Education.
- Pakpahan, R. (2016). Model ujian nasional berbasis komputer: Manfaat dan tantangan. *Jurnal Pendidikan Dan Kebudayaan*, 1(1), 19–35. <https://doi.org/10.24832/jpnk.v1i1.225>
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.
- Riley, B. B., & Carle, A. C. (2012). Comparison of two Bayesian methods to detect mode effects between paper-based and computerized adaptive assessments: A preliminary Monte Carlo study. *BMC Medical Research Methodology*, 12, 124. <https://doi.org/10.1186/1471-2288-12-124>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. In *Psychometrika Monograph, No. 17*. Richmond, VA: Psychometric Society.
- Schraw, G. J., & Robinson, D. H. (2011). *Assessment of higher order thinking skills: Current perspectives on cognition, learning, and instruction*. Charlotte, NC: Information Age Publishing.
- Suyoso, S., Istiyono, E., & Subroto, S. (2017). Pengembangan instrumen asesmen pengetahuan fisika berbasis komputer untuk meningkatkan kesiapan peserta didik dalam menghadapi ujian nasional berbasis komputer. *Jurnal Pendidikan Matematika Dan Sains*, 5(1), 89–97. <https://doi.org/10.21831/jpms.v5i1.12461>
- van der Linden, W. J., & Glas, C. A. W. (2003). *Computerized adaptive testing: Theory and practice*. London: Kluwer Academic Publisher.
- Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70–84. <https://doi.org/10.1080/07481756.2004.11909751>

## An evaluation of internship program by using Kirkpatrick evaluation model

<sup>\*1</sup>Lathifa Rosiana Dewi; <sup>2</sup>Badrun Kartowagiran

<sup>1,2</sup>Department of Educational Research and Evaluation, Graduate School of Universitas Negeri Yogyakarta

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

<sup>\*</sup>Corresponding Author. E-mail: [lathifarosianadewi@gmail.com](mailto:lathifarosianadewi@gmail.com)

*Submitted: 20 December 2018 | Revised: 21 December 2018 | Accepted: 22 December 2018*

### Abstract

This study was aimed at evaluating an internship program using Kirkpatrick's evaluation program. The subjects of the study were students of batch 2015 and instructors. Slovin formula was used to calculate the sample. A questionnaire and teaching assessment sheet were used as instruments for collecting data. This study used content validity and exploratory factor analysis as the validity of the test. Reliability was estimated by Cronbach's Alpha. The results of this study showed that (1) in facility, the level of satisfaction was in the 'very satisfactory' category (77.01%); (2) in instructor, the level of satisfaction was in the 'very satisfactory' category (82.76%); (3) in schedule, the level of satisfaction was in the 'satisfactory' category (50.57%); (4) in material, the level of satisfaction was in the 'very satisfactory' category (89.66%); and (5) in students' teaching abilities. The improvement was in the 'very satisfactory' category.

**Keywords:** *program evaluation, internship, Kirkpatrick model*

### Introduction

Teachers' quality determines the quality of education. Teachers are said to be qualified when they have competencies to plan, teach, evaluate, guide, train, research, and conduct community service (article 39 of Law of Republic of Indonesia No. 20 of 2003). According to Jailani (2014), there are some teachers who are not qualified to teach; in public elementary schools 78.93%, private elementary schools 71.06%, public secondary schools 45.88%, private secondary schools 39.01%; public high schools 34.71%, and private high school 35.27%. This may give unfavorable effects to the educational practices in Indonesia. Meanwhile, teachers in Indonesia still have an important role in the national education. Teachers, therefore, are expected to have good competencies.

Teachers who have good competencies are believed to have good abilities in teaching. This statement is supported by Ardiansyah

(2013) who states that teachers who have good competencies can teach well. There are four competencies which should be owned by the teacher namely pedagogic competence, personal competence, professional competence, and social competence. Pedagogic competence is the teacher's competence in managing teaching and learning processes. Their ability in managing the class, arrange the students' seats, and others are examples of pedagogic competences. Personal competence is competence to influence students to have good attitudes. Professional competence is a teacher's competence in mastering the material. The last is social competence, where teachers should be able to have good interaction with the students, other teachers, and parents. These competencies can lead to the success of the teaching and learning process.

Hallo and Munadi (2014) mention the same thing that teachers have important roles in the success of teaching and learning process. The success of the teaching and learn-

ing process cannot be realized if they do not have good competencies. This can be the reason why teachers should have good competencies since they still are prospective teachers. This is aimed at making them ready when they should become teachers in the field. If they do not have good competencies, they will be just teachers who only transfer knowledge.

Nowadays, some teachers only transfer knowledge to the students. They just deliver the class material without knowing whether their students understand the material or not. Teachers should play their role to teach, evaluate the teaching and learning process, and improve anything that needs to be improved there. This should be realized by teachers from the very first time they are in the teaching and learning process. This can happen when they are trained to be a teacher while they are in the university. Each university has a program called Teaching Training Internship (TTI). This is a program where prospective teachers train to be a real teacher while they are in university.

TTI is a program which is held in the last semester of the curriculum. This program trains the prospective teachers to teach and do anything real teachers do in the classroom. This program aims to build the prospective teachers' characters so that they are ready to be teachers. Mardiyono (2006) argues that this program focusses on the prospective teachers' abilities in teaching in the classroom and doing school administration. This means that prospective teachers learn about not only how to manage the classroom and deliver the material, but also how to do school administration. It is in line with Kiggundu and Nayimuli (2009) who insist that teaching training is the activity to integrate the theory obtained from the class with practice. Some teacher training institutions implement this program, however, some do not. They have another program called internship programs.

Both internship and TTI have the same characteristics in that they train the prospective teachers to be real teachers. The aim of this internship program is to give students experience in teaching. This program lets the students in each batch teach in an addressed

school. In the initial phase, they will be trained to create lesson plans and develop class material. They will then apply what they have learned in the teaching practice in the classroom. This program is a mandatory program which means that each student teacher should take it in a year. There is an instructor who comes from the addressed school. The instructor is an English teacher at that school. The instructor should guide the student teacher how to plan a lesson, create instructional material, manage the class, and do many other things that teachers do in the classroom.

This program is divided into two. In the first semester, students should attend the debriefing. Debriefing means that students are guided to create a lesson plan, develop instructional material, and complete classroom activities. At the end of the semester, students are expected to submit the lesson plan class material. In the second semester, students do the teaching practice at the assigned school.

This program has been running for some years, but it has not been evaluated in an appropriate way. This means that the evaluation process in the program just merely gives how many students come and the strengths and weaknesses of the students. However, it has not been reported. Considering its importance, the program should be evaluated.

There are many approaches that can be applied to conduct a program evaluation. Fitzpatrick, Sanders, and Worthen (2011, p. 114) explain that the differences in evaluation approaches come from the background, experience, and worldview of the authors. This means that each approach is affected by the author. This means that an author can choose the approach which is appropriate for the evaluation process.

One of the evaluation models that can be used is Kirkpatrick's evaluation model. This model aims to evaluate the training program. There are four levels in this evaluation model namely reaction, learning, behavior, and result. Kirkpatrick and Kirkpatrick (2006, p. 21) mention that reaction assesses the satisfactory level of the program; learning assesses what knowledge has been obtained and improved; behavior assesses the changes of the

trainees' behavior after the program; and result assesses the final result, focusing on the benefit for the institution.

### Evaluation Principles

An evaluation is a systematic process which gives out information about program achievement. It means that evaluation gives information whether the objective has been achieved or not. Evaluation is a systematic process to gather data, information, and interpretation so that this can be used as the basis for policy making, decision making, or creating another program as the results of the evaluation. This can be information that can be used to revise, stop, or continue the program (Abrory & Kartowagiran, 2014).

Evaluation is different from research in terms of objectives. While research is aimed at obtaining new theories, evaluation is not. People cannot get new theories from evaluation. What people obtain from evaluation is merely information about the success of a program. Besides, evaluation can give information on the impact, or effectiveness, of a program (Stufflebeam, Madaus, & Kellaghan, 2002). It indicates that evaluation has the same method with research, but the result is really different. Research does not create a new theory but information. The information is really useful for policymaking.

In doing an evaluation process, the evaluator should follow the standards that need to be done. This is in line with Yarbrough, Shulha, Hopson, and Caruthers (2011) who believe that there are four standards that should be followed namely utility, accuracy, feasibility, and propriety. The explanation of these standards is as follows. (1) Utility means that the information which is obtained from evaluation should be useful and practical. In other words, the information can be used as a basis for decision making and for the success of the program. (2) Accuracy means that the information which is gathered should fulfill the requirements for rules of data gathering. In this case, the process of information gathering should be conducted in the right way of research in terms of instrumentation, validity, reliability, measurement, and generality. (3) Feasibility means that an evaluation study

should be proper both in the politic or cost-effectiveness. This means that, when doing an evaluation, everything should be considered. Politics means that there is no interest while doing the evaluation. For example, policy-making requires evaluation and, thus, evaluation is developed. Besides, cost-effectiveness should be considered so that there is no wasted cost. (4) Propriety means that evaluation should be done legally. This means that evaluation cannot be done in secret. The code of ethics of evaluation should be obeyed.

Evaluation is a process to measure a program, make a decision, and know the usefulness of a program. Evaluation is done when the decision maker or stakeholders are curious about the success of the program (Irambona & Kumaidi, 2015). Evaluation has an important role in the running of a program. Without evaluation, people do not know whether the program is successful or not so that follow-ups can be taken.

### Kirkpatrick's Evaluation Model

Kirkpatrick's evaluation model was employed to evaluate a training program. There are four stages in this evaluation model, including: reaction, learning, behavior, and result. These four stages can be described as follows (Kirkpatrick & Kirkpatrick, 2006, p. 21).

#### *Reaction*

In this stage, the researchers measure the level of participants' satisfaction with the program. Training programs are considered successful if the trainees are happy with the program so that they are motivated to learn. Interest, attention, and motivation of participants in following the course of training are indicators of the success of the program. In this first stage, trainees will be given a questionnaire of satisfaction on matters relating to training such as materials, instructors, training environment, and consumption in the training.

#### *Learning*

Learning can be defined as a change of attitude, improvement of knowledge, and or enhancement of the skills of the participants

after the program. There are three components to be measured in this evaluation: what knowledge has been learned, what attitude has changed, and what skills have been developed or improved. To measure all three components, then, it takes a test.

#### *Behavior*

In this evaluation, what is assessed is the attitude change of the trainees after returning from the program. The focus in this level is whether or not the trainee applies what has been obtained from the program.

#### *Result*

Evaluation at this stage is at the final stage. It is focused on the final results after the participants follow the program.

#### *Internship*

An internship is a program which is implemented in order to prepare prospective teachers to become teachers who have good skills. Inside is a professional preparation stage where a student has gained knowledge to be applied in the field with the supervision of several interested parties and within a certain period of time (Hamalik, 1990). Thus, an internship program is a program in which a student does science applications that have been obtained. In education, internship can be interpreted as the application of competences which are possessed by a teacher in school.

There are several objectives of holding an internship program of education as expressed by Hamalik (1990). These include developing a more comprehensive view to the intern about education, equipping the intern with experience about the implementation and responsibility of education as a teacher, enabling the intern to get knowledge from supervisors in school, and providing an overview to the intern about the professional code of ethics of a teacher.

In recent literature, internship is defined as an experiential learning that integrates both the theory and knowledge which are acquired in the classroom with practice (Kiser, 2016). The purpose of holding an internship is to gain valuable experience about the application

of science that has been obtained previously and make connections between the science and the field of profession based on the future career goals. Kiser (2016) mentions several important things in the internship, that is, the time spent during the internship, how time is used, the quality of the internship, and the application of the previous learning.

Based on the importance of internship evaluation, the research objective is to find out five levels of satisfaction towards the components of the internship program. These are levels of satisfaction towards (1) facilities, (2) instructors, (3) scheduling, (4) content material, and (5) students' improvement.

#### **Method**

The study was conducted in the vicinity of Muhammadiyah University of Yogyakarta (or *Universitas Muhammadiyah Yogyakarta* - UMY). Of the four Kirkpatrick's model, only two are conducted: reaction and learning. For the first level, the study is intended to find out the satisfaction level towards the program seen from facilities, instructors, scheduling, and material. For the second, the study is intended to find out the students' improvement of teaching abilities.

The subjects of the study were students of English education department batch 2015 and some instructors. The sample for this study consisted of 87 of 103 students. The number of respondents was calculated by using the Slovin formula.

A questionnaire was used to gather data about the satisfaction level towards the internship program. There were four aspects namely facilities, instructors, material, and schedule. Meanwhile, improvement of teaching abilities was obtained by using performance sheets. In addition, students and teachers in each school were interviewed to gather additional information.

The validity measures implemented in the study were of content and construct. Content validity is one which confirms what the instrument is supposed to measure (Azwar, 2015, p. 111). The questionnaire and interview guidelines were judged by three experts, and the data were subjected to the Aiken formula. All instruments were valid because the

Aiken value was higher than 0.7. It is in line with Azwar (2015, p. 149) who mentions that coefficient value can be said to be valid when the value is higher than 0.35. For the construct validity, factor analysis was used. There were four aspects in the questionnaire: facilities, instructors, schedule, and material. From the results of the construct validity measures, one item in the facilities and schedule aspect which should be dropped. The questionnaire reliability was estimated using Cronbach's Alpha. There were 36 items. The reliability value was 0.844. This can be said to be reliable.

For the quantitative data of the students' survey, the descriptive statistics proposed by Azwar (2017, p. 148) as presented in Table 1 was employed. After analyzing the quantitative data, the results were interpreted qualitatively. The results from the quantitative analyses were then cross-checked with the students and teachers before a conclusion was made.

Table 1. Normal curve statistics for students' satisfaction

Score X	Categories
$X > M + 1.5 SD$	Very satisfactory
$M + 0.5 SD < X \leq M + 1.5 SD$	Satisfactory
$M - 0.5 SD < X \leq M + 0.5 SD$	Fairly satisfactory
$M - 1.5 SD < X \leq M - 0.5 SD$	Less satisfactory
$X \leq M - 1.5 SD$	Not satisfactory

Notes:

M: Ideal mean of the concerned component in this research.

$[ M = \frac{1}{2} (\text{highest ideal score} + \text{lowest ideal score}) ]$

X: the total point scored by each respondent regarding to each item/component to evaluate.

SD: Ideal standard deviation of each component.

$[ SD = \frac{1}{6} (\text{highest ideal score} - \text{lowest ideal score}) ]$

#### Students' Satisfaction toward Facilities

In this section, each student scored five points as an ideal minimum score and the maximum ideal score was 25. Thus, the ideal mean was 15, and the standard deviation became 3.33. The facilities were judged satisfactory if the mean score belongs to the first category (Very satisfactory). The criteria are defined in Table 2.

Table 2. Evaluation criteria of facilities, instructor, schedule, and material

Score X	Categories
$X > M + 1.5 SD$	Very satisfactory
$M + 0.5 SD < X \leq M + 1.5 SD$	Satisfactory
$M - 0.5 SD < X \leq M + 0.5 SD$	Fairly satisfactory
$M - 1.5 SD < X \leq M - 0.5 SD$	Less satisfactory
$X \leq M - 1.5 SD$	Not satisfactory

#### Students' Satisfaction Level toward Instructor

There are 20 questions used in this instructor aspect. Based on the criteria, the ideal minimum score was 20 and the ideal maximum score was 100. Thus, the ideal mean was 60 and the ideal standard deviation was 13.3. The instructor was considered to be satisfied if the mean score belongs to the first category (Very satisfactory). Then, the very satisfactory category was converted to a percentage.

#### Students' Satisfaction Level toward Schedule

The schedule aspect included two questions. The ideal minimum score of this aspect was 2 and the ideal maximum score was 10. Thus, the mean ideal of this aspect was 6 and the standard deviation was 1.33. The schedule was judged to be satisfactory if the mean score belongs to the first category (Very satisfactory).

#### Students' Satisfaction Level toward Material

There were seven questions in the material aspect. Based on the criteria, the ideal minimum score was 7 and the ideal maximum score was 35. Thus, the ideal mean was 21 and the standard deviation was 4.67. The material was considered to be satisfactory if the mean score belongs to the first category (Very satisfactory).

#### Students' Teaching Ability Improvement

To measure students teaching ability improvement, the instructors were asked to fill the performance sheet. There were 'increase' and 'not increase' category. The instructor should fill the sheet by putting check marks. For 'increase' category, there were five improvement categories as mentioned in Table 3. Then, each category was converted to a percentage.

Table 3. Evaluation criteria of students' teaching ability improvement

Score X	Categories
$X > M + 1.5 SD$	Very high
$M + 0.5 SD < X \leq M + 1.5 SD$	High
$M - 0.5 SD < X \leq M + 0.5 SD$	Fairly high
$M - 1.5 SD < X \leq M - 0.5 SD$	Less high
$X \leq M - 1.5 SD$	Not high

### Findings and Discussion

Students' satisfaction becomes the most important aspect of any program. In this internship program, students' satisfaction will affect student's motivation and this can lead to the program success. Badu (2013) asserts that program effectiveness is where the training program is fun and enjoyable so that students can get a high motivation to learn.

Evaluation of reaction for the internship program was measured based on the students' satisfaction toward the program. There were 34 statements in the questionnaire, grouped into four aspects namely facilities, instructor, schedule, and material. Each aspect has a different number of statements. The facility aspect has five statements, instructor aspect has 20 statements, schedule aspect has two statements, and the material aspect has seven statements.

The indicator that represents the level of satisfaction toward the program is comfort and suitability. Comfort means that the rooms were well equipped. This can be known from the using of media, air conditioner, and air freshener. Suitability means the readiness of the room. Two statements for this suitability factor is the readiness of room before it was used and room capacity was suitable for students' number. The result showed that 77.01% of the students reported that facilities were in the very satisfactory category; 21.84% satisfactory category; and 1.15% fairly satisfactory category. Each item in the facility aspect then was categorized 'very satisfactory' and 'satisfactory'. Four items (the using of air conditioner, media, room readiness, and also room suitability to the student's number were in the very satisfactory of fresheners was only in the satisfactory category. Based on the interview, students said that the room which was used for the coaching was well equipped.

However, the using of fresheners was less. Vonny (2016) states that facilities can give satisfaction. This means that when students were asked about satisfaction, they will mention facilities aspect as one of the indicators. The implication from this study is that the better the facilities, then, the higher the increase. From the study, it can be concluded that a program can be said as satisfying where the facilities are good. The internship program can be regarded as successful because more than 50% of the students stated that the room for coaching has been equipped by the good facilities.

Instructor becomes one of the most important roles in a coaching program. The instructor should be selected carefully because they can give either good or bad effects for trainees. Instructors of internship programs need to be evaluated because they give important material before students do the teaching practice. There were 20 statements to measure the students' level of satisfaction toward the instructor. These statements include the instructor's readiness before the coaching, the delivery strategy, the delivery of materials, the ability to communicate orally, the ability to communicate in writing, and the use of media.

A total of 82.76% of the students stated that the instructor aspect was in the very satisfactory category. They mentioned that instructors' abilities in delivering the material were good so that they could understand the material well. Besides, they deliver the material in detail and a fun way. Students enjoyed joining the coaching and they could understand the material well.

This study found that students felt satisfied with the use of media and teaching video. This means that the instructor did not give them the teaching video as only an example. Some students reported that their instructor did not use the media often. This can lead to the conclusion that they just talked in the class without doing anything. Putri and Kartika (2016) report the same thing that the highest level of satisfaction was attached to instructors who have good abilities in delivering material and who can be fun too. For example, the instructor used jokes while delivering the material.



The internship program was scheduled for eight sessions in the semester. Each group had a different schedule based on the agreement between students and instructors. This was revealed by the interview with students and instructor. They said that the internship schedule was flexible so that each group had a different schedule. This means that a group may complete the internship program in only two months but the others may not.

This aspect actually included three items, but one item should be deleted due to the factor analyses. These items were the time to start the coaching and the time to end the coaching. These items can represent students' satisfaction levels because the schedule is one of the crucial things. When the coaching was not based on the schedule, this can affect the students' responses.

Students' level of satisfaction toward the schedule was only categorized by 'satisfactory'. A number of 50.57% of the students mentioned that the schedule was in the satisfactory category. Students reported that instructors used time in each coaching. Students felt useless because the coaching time did not give them any information. Zahro and Wu (2016) state that time allocation in a program should be evaluated so that there would be an improvement of the schedule for the next coaching. To anticipate the instructor who has not kept the right schedule, there should be a team for monitoring the internship program. It is in accordance with Rohani (2015) who mention that the needs of a quality control team will give a good supervisory function. Supervisors should check the coaching time in a week, for example. They cannot just come then go, but they should be there along the coaching time. This aimed to decrease the bias. It means that when students do the best to teach, then, there is no supervisor who does not come to supervise, giving students disadvantages. It is in line with Sahraini and Madya (2015) who report that teachers who have good abilities in teaching will not be appreciated because the evaluation has no regular schedule.

Material becomes one of the most important aspects of evaluation. The better the material, the better the impacts it gives to the

trainees. There are seven statements which are divided into two factors, namely material suitability with learning and material conformity to students' needs. These items are material conformity with the lesson plan, the systematics of material delivery, the interrelationship within the material, the suitability of the material with the curriculum used in the partner school, the way the selection of teaching materials, the way of choosing learning strategies, and how to manage the class.

In this study, evaluation toward material was in the very satisfactory category, as high as 89.66%. This was confirmed by the students' interviews. They mentioned that the instructor gave the lesson plan before coaching so that they knew what will be done in the coaching. Besides, the instructor gave the suitable material for them like curriculum, syllabus, and lesson plan which was used in each school. It leads to the students' understanding of what should be written in the lesson plan and what should be done in the teaching practice. In other words, the material was really useful for their needs in the teaching practice. The material in the internship program has been fitted to the students' need in both coaching and teaching practice. Utomo and Tehupeioro (2014) mention the same thing about the importance of aligning the material delivered to the students with the program objective. Program coordinators should keep this right. This means that the material which was suitable for the students' needs should be kept, while material which was not used for the internship program could be considered to be deleted.

Evaluation in learning is conducted to assess what has been learned by students, what kind of ability which improved, and what has changed (Kirkpatrick & Kirkpatrick, 2006, p. 21). This evaluation only focused on the improvement of ability in teaching. After coaching, students should do the teaching practice three times. They did the teaching practice in a class for the instructor to give them a grade.

The evaluation result was that the students' abilities in the practice teaching improved by a high-level category. This means that their teaching has changed in each time

of the teaching practice. There were five schools, coded School 1, School 2, School 3, School 4, and School 5.

From the descriptive data, it can be interpreted that 79% of students who did their teaching practice in School 1 showed improvement in their teaching abilities. Students at School 2 gave a higher score of 92%. At School 3, improvement was marked by 72%. Students at School 4 improved their teaching ability as much as 92%. Students at School 5 showed the lowest percentage of 55%. This was supported by the qualitative data from the interviews with instructors. They stated that some students have learned well but the others have not. Students who have not improved were those who did not change their way of teaching. On the average, however, it is indicated that students' ability in teaching improved by the high level of category. More than 50% of the students improved their ability in teaching by the high-level category in each school. The internship program can be said to be successful because there was an improvement in the students' teaching abilities. It is in line with Al Yahya and Norsiah (2013) who stated that ability improvement is an indication of success in a program.

## Conclusion and Recommendations

### Conclusion

Based on the findings of the reaction aspect, it can be concluded that three aspects have occupied the 'very satisfactory' category. These were facilities, instructor, and material. On the other hand, the schedule aspect did not obtain the 'very satisfactory' category. This was mostly caused by the fact that the instructor was over-timed in each coaching.

For the learning aspect, there were more than 50% of students in each school who had the 'high level' category of improvement. It can be concluded that students understood the material well so that they could apply the material learned from the coaching in the teaching practice.

### Recommendations

Some recommendations are proposed for program coordinators. The program co-

ordinators should monitor the internship program from the beginning until the end. This means that they should know what the strengths and what weaknesses of the program are. Coordinators can come to the coaching session in each school or they can just interview students about what has missed in the program.

Besides, coordinators should evaluate the internship program periodically. It has been known that evaluation can be done before the program, whilst program, and at the end of the program. It is highly recommended that the coordinators have a team for such periodical evaluations. This can prevent the program from various difficulties and weaknesses.

Coordinators, lecturers, and instructors can create the criteria of the success of the internship program. It means that there should be specific criteria to measure the success of the program. This will help them in giving a quality evaluation to the program.

## References

- Abroy, M., & Kartowagiran, B. (2014). Evaluasi implementasi Kurikulum 2013 pada pembelajaran matematika SMP negeri kelas VII di Kabupaten Sleman. *Jurnal Evaluasi Pendidikan*, 2(1), 50–59.
- Al Yahya, M. S., & Norsiah, B. M. (2013). Evaluation of effectiveness of training and development: The Kirkpatrick model. *Asian Journal of Business and Management Sciences* (Vol. 2).
- Ardiansyah, J. (2013). Peningkatan kompetensi guru bidang pendidikan di Kabupaten Tana Tidung. *EJournal Pemerintahan Integratif*, 1(1), 38–50.
- Azwar, S. (2015). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.
- Azwar, S. (2017). *Penyusunan skala psikologi*. Yogyakarta: Pustaka Pelajar.
- Badu, S. Q. (2013). The implementation of Kirkpatrick's evaluation model in the learning of initial value and boundary condition problems. *International Journal of Learning and Development*, 3(5), 74–88. <https://doi.org/10.5296/ijld.v3i5.4386>

- Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Boston, MA: Pearson Education.
- Hallo, D., & Munadi, S. (2014). Evaluasi kinerja guru SMA yang bersertifikat profesional di Kabupaten Halmahera Barat. *Jurnal Evaluasi Pendidikan*, 2(2), 111–122.
- Hamalik, O. (1990). *Sistem internship kependidikan teori dan praktek*. Bandung: Mandar Maju.
- Irambona, A., & Kumaidi, K. (2015). The effectiveness of English teaching program in senior high school: A case study. *REiD (Research and Evaluation in Education)*, 1(2), 114–128. <https://doi.org/10.21831/reid.v1i2.6666>
- Jailani, M. S. (2014). Guru profesional dan tantangan dunia pendidikan. *Al-Ta'lim Journal*, 21(1), 1–9. <https://doi.org/10.15548/jt.v21i1.66>
- Kiggundu, E., & Nayimuli, S. (2009). Teaching practice: A make or break phase for student teachers. *South African Journal of Education*, 29, 345–358.
- Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs* (3rd ed.). San Francisco, CA: Berrett-Koehler.
- Kiser, P. M. (2016). *The human services internship: Getting the most from your experience*. Boston, MA: Cengage Learning.
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).
- Mardiyono, S. (2006). Praktik pengalaman lapangan terpadu dalam peningkatan kualitas calon guru. *Jurnal Cakrawala Pendidikan*, 25(1), 57–72. <https://doi.org/10.21831/cp.v0i1.392>
- Putri, Y. E., & Kartika, L. (2016). Evaluasi efektivitas pelatihan marketing skills pada PT XYZ. *KOLEGIAL*, 4(2), 11–22.
- Rohani, E. (2015). Analisis kepuasan peserta pelatihan pertolongan pertama gawat darurat obstetri dan neonatal (PPGDON) di balai pengembangan tenaga kesehatan (BPTK) Mataram menggunakan metode servqual. *Media Bina Ilmiah*, 9(2), 37–45.
- Sahraini, S., & Madya, S. (2015). Model evaluasi internal kompetensi guru bahasa Inggris (Model\_EIKGBI) SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 19(2), 156–167. <https://doi.org/10.21831/pep.v19i2.5576>
- Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2002). *Evaluation models: Viewpoints on educational and human services evaluation*. New York, NY: Kluwer Academic Publishers.
- Utomo, A. P., & Tehupeiry, K. P. (2014). Evaluasi pelatihan dengan metode Kirkpatrick analysis. *Jurnal Telematika*, 9(2), 37–41.
- Vonny, R. P. E. (2016). Pengaruh pelatihan, fasilitas kerja dan kompensasi terhadap kepuasan kerja karyawan pada PT United Tractors cabang Manado. *Jurnal Berkala Ilmiah Efisiensi*, 16(3), 407–418.
- Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: SAGE.
- Zahro, S., & Wu, M. (2016). Implementing of the employees training evaluation using Kirkpatrick's model in tourism industry - A case study. *International Journal of Innovation and Applied Studies*, 17(3), 1042–1049.

## Comparing the methods of vertical equating for the math learning achievement tests for junior high school students

\*<sup>1</sup>Chairun Nisa; <sup>2</sup>Heri Retnawati

<sup>1</sup>Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta  
Jl. Colombo No. 1, Depok, Sleman, Yogyakarta 55281, Indonesia

<sup>2</sup>Department of Mathematics Education, Universitas Negeri Yogyakarta  
Jl. Colombo No. 1, Depok, Sleman, Yogyakarta 55281, Indonesia

\*Corresponding Author. E-mail: [c.nisa.258@gmail.com](mailto:c.nisa.258@gmail.com)

*Submitted: 12 April 2018 | Revised: 20 December 2018 | Accepted: 21 December 2018*

### Abstract

Developing the students' mathematical ability needs to be carried out to improve the teaching process. This is very important for continuous education. This study aimed to describe: (1) the characteristics of the mathematics achievement tests for grades VII and VIII; (2) the equity constant of the vertical equating result of the mathematics achievement; (3) the accuracy of the mean & mean method, mean and sigma, Haebara characteristics curve, Stocking & Lord characteristics curve methods in the vertical equating of the tests for grades VII and VIII. The data were the students' scores for the Higher Order Thinking tests collected with the anchor test design. The analysis technique utilized was the descriptive quantitative analysis. The findings of the study show that: (1) the learning achievement tests for grades VII and VIII have the difficulty level (location) in the fair category (0.190 and 0.451), and the discrimination index (slope) in the category of good with the mean of 0.700 and 0.633; (2) the vertical equating result shows an equation of  $Y' = 0.88X - 0.27$  with the mean and mean method,  $Y' = 0.19X - 0.02$  with the mean and sigma method,  $Y' = 0.38X - 0.12$  with the Haebara characteristics curve method, and  $Y' = 0.57X - 0.18$  with the Stocking and Lord characteristics curve; (3) the lowest Root Mean Square Different (RMSD) belongs to the mean and mean method, followed by the Stocking and Lord characteristics curve method, mean and sigma method, and the Haebara characteristics curve method.

**Keywords:** *equating method, vertical equating, HOT, mathematics*

### Introduction

Science and technology development in Indonesia has brought about changes in almost every aspect of human life. The development demands that different problems be solved through the effort to master science and technology. In order to be able to contribute to the global competition in the 21st century, human being needs to develop the self-quality so that they can compete with others. The human resource quality is influenced by education.

The improvement of the quality of education becomes an essential pillar for the development of education in Indonesia. Quality education will result in competitive human

resources as stated in the Law of Republic of Indonesia No. 20 of 2003 on National Education System. It is stated that the national education functions to develop the capability, character, and civilization of the nation for enhancing its intellectual capacity, and is aimed at developing learners' potentials so that they become persons imbued with human values who are faithful and pious to the one and only God; who possess morals and noble character; who are healthy, knowledgeable, competent, creative, independent; and as citizens, are democratic and responsible.

The Partnership for 21st Century Skills (P21) argues that teaching should focus on developing critical thinking, communication, collaboration, and creativity as students' skills

in the 21st century. The 4C's skills are a part of the higher order thinking skills. Therefore, students need to develop their higher order thinking skills in every educational process.

Based on the data from different surveys, it is found out that students' achievement in maths in PISA and TIMSS is still low. Thus, holistic and continuous efforts need to be made to improve the quality of education from all parties including students, teachers, principals, and the government. According to Mardapi (2012, p. 12), efforts to improve the quality of education in educational institutions can be made by improving the quality of the teaching and assessment system. This means that teaching is closely related to assessments. Teachers, as an important component in education, should be able to carry out their duties and play their roles as stated in Law No. 14 of 2005 of Republic of Indonesia about Teachers and Lecturers. Teachers are expected to be able to develop students' potentials, through both the teaching process in the class and the assessment model used. The assessment model used by teachers can actually provide information on the teaching process and learning achievement.

A good assessment can be carried out by collecting accurate data related to students' learning achievement and this can make the class assessment process beneficial to the students, that is, it can improve the students' motivation and learning achievement (Stiggins & Chappuis, 2012, p. 3). Therefore, learning achievement assessment is expected to be able to provide information about the students' ability development. The information can be a reference to know the quality of the learning achievement at the class, school, or national levels. This can be used as a study to improve the quality of Indonesian education.

The test in mathematics has different characteristics from that in other subjects. The mathematics materials are hierarchical and closely related to each other. This means that the students' mastery of previous materials becomes the basis for continuing to and understanding of the materials in the next level. Teachers are expected to be able to write a good test and also to use the test to connect the students' learning achievement in different

grades so that the information about the students' ability development can be known.

In addition to knowing the characteristics of the test items used, teachers are expected to make sure that, in order that the information about the students' ability development is accurate, the test items should be in the students' ability level (Gagné, 1977, p. 158). The use of test items which are beyond the students' ability will make the students unable to answer the questions so that teachers will not be able to find out the information about the students' development. Students of the same age and grade may not have the same development.

The scores of two different tests from two or more different groups can be compared when the items are equal and are based on the same scale (Kolen & Brennan, 2004, p. 5). The equating between scores can be done statistically. A statistical analysis is carried out to the scores of two different tests to be adjusted on the same scale. The statistical process used to produce a single scale from the scores of two different tests with the same scale is called equating (Kolen & Brennan, 1995, p. 5). Hambleton, Swaminathan, and Rogers (1991, p. 123) state that equating is a process to transform the score  $X$  to the test score matrix  $Y$  or vice versa, so that the result of the equating process can be compared.

There are two kinds of equating process which can be conducted to test scores: horizontal equating and vertical equating. Horizontal equating is the equating carried out to test scores which have equal difficulty index at the same grade, while vertical equating is the equating process carried out to reveal the students' ability measured by test instruments which have different difficulty index and on different grades, but they measure the same trait (Crocker & Algina, 2008, p. 456; Hambleton & Swaminathan, 1985, p. 197). Thus, the vertical equating can be used by teachers to reveal the students' ability development although the students are in different grades and they have different abilities provided that the tests measure the same traits.

The equating using the Item Response Theory approach can be carried out using different methods. They are the mean-and-mean

method, the mean-and-sigma method, and the characteristic curve transformation (Kolen & Brennan, 2004). Several previous studies carried out used the classical approach and Item Response Theory on elementary school students by Antara and Bastari (2015); equating using the IRT approach with the mean-and-mean method, mean-and-sigma, Haebara, and Stocking-and-Lord methods for the mixed model by Kartono (2008), and equating using the IRT approach on mixed tests by Uysal and Kilmen (2016). Previous studies showed the accuracy of different methods. The utilization of different equating methods resulted in different equating results, so to find out an accurate result, it is necessary to choose the appropriate design and method in accordance with the condition. Therefore, teachers will be able to find accurate information about the students' ability development.

The scoring model used was Generalized Partial Credit Model (GPCM). This is because GPCM is an alternative scoring in the teaching assessment (Istiyono, 2016).

## Method

This is a study of vertical equating in general using the quantitative approach. In the instrument development part, the researchers developed a mathematics HOTS instrument using the mixed model for grades VII and VIII of junior secondary schools administered in the even semester. Revision based on the expert's suggestions was carried out after the readability testing and content validation by an expert. The revised instrument was then tried out in one junior secondary school which was not the sample of the study, that is, SMPN 3 Lubuk Pakam. The data from the try out were analyzed using the IRT approach using the Parscale program to find out the characteristics of the developed items so that the items would be good items. The good items were then set into a mathematics test for grades VII and VIII.

The research was carried out in Deli Serdang District, Indonesia, especially in public junior secondary schools in the district in the academic year of 2016/2017. The study was conducted from May to June 2017. The population was students' response on mathe-

matics test. There were 51 schools taken as the sample using the stratified proportional random sampling technique. This was done because the population had levels, that is, grade VII and grade VIII. The students in each grade were then selected proportionally. The schools were categorized into high, middle, and low categories based on the national examination scores in the previous year (data obtained from *Dinas Pendidikan Pemuda dan Olahraga*). Five schools were selected to be the sample. The sample consisted of 1009 students, including 505 grade VII students and 504 grade VIII students.

The HOTS test instrument on mathematics used was the GPCM analysis model. The test consisted of 15 items each set consisting of 10 items in the form of multiple choice and five items in the form of essay items. The multiple choice items were used as this kind of items is more objective and reliable in finding out the students' response, nor influenced by the subjectivity of the scorers. Meanwhile, the essay items were used to find out the students' higher order thinking skills. The equating design used was the common item non-equivalent groups (Hambleton & Swaminathan, 1985). Both tests had the same items as the anchor. Four multiple choice items (26.7%) were used as the anchor. It was based on the theory which states that the minimum items for the anchor is 20% (Kolen & Brennan, 1995, p. 248).

Unidimension testing was conducted by the factor analysis in SPSS 22. Data can be analyzed using the factor analysis when they meet two criteria: Kaiser-Meyer-Olkin Measure Sample of Adequacy (KMO-MSA) and Bartlett's Sphericity Test. KMO-MSA test was needed to see the sample adequacy and Bartlett's test was used to see the normality of the analyzed data. Field (2000, pp. 453–469) states that further analysis can be carried out when the KMO has the sig. < 0.05 and the MSA is > 0.05. Hambleton and Swaminathan (1985, p. 16) state that the unidimension testing was met when the test only measures one dominant dimension, that is, the same ability. The unidimension aspect can be seen from the eigenvalue obtained from each test and the unidimension criterion can be seen from

the scree plot formed. The local independence assumption testing functions to find out that the students' ability is independent from the items. This means that the students' answer to one item is not influenced by the answer to another item. The conformity of the model was done to know the appropriate model with the analyzed data. The conformity testing was meant to know that the items used were appropriate with the model used. The way used to know the conformity of the model was by comparing the chi-square observed and the chi-square table with a certain degree of freedom. Then, the parameter estimation and ability estimation were carried out with the appropriate model. The parameter estimation and the ability estimation were analyzed using the Parscale program.

Vertical equating was carried out based on the result of the item characteristics analysis using IRTEQ program (Han, 2009). The test equating was done by making an equation using the mean-and-mean, mean-and-sigma, Haebara characteristics curve, and Stocking-and-Lord characteristics curve methods to see the equating of the test for grades VII and VIII based on the difficulty index and the discrimination index in the test anchor.

The next step was finding the smallest error of the used equating methods. The accuracy of the method can be seen by calculating the RMSD for each method.

$$RMSD(\theta) = \sqrt{\frac{\sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2}{N}}$$

Notes:

N = the number of the testees,

$\hat{\theta}_i$  = the first students' ability after being equated

$\theta_i$  = the first students' ability before being equated.

## Findings and Discussion

### Findings

Based on the item response theory used, it is necessary to find out the assumption of the item response theory. When the assumptions were met, it is possible to conduct further item response theory analysis. There are two assumptions, i.e. unidimension assumption, and local independence assumption.

### *The Unidimension and Local Independence Assumptions*

Before testing the unidimension assumption, it is necessary to find out the adequacy of the sample through KMO-MSA and Bartlett's test of Sphericity for the normality of the data which were used. The empirical analysis for the test for Grade VII shows that the KMO-MSA value is 0.933 with the Bartlett's test significance of 0.000. Meanwhile, for the test for grade VIII, the KMO-MSA value is 0.867 with the Bartlett's test significance of 0.000. Based on the result of the analysis, it is indicated that both instruments for grade VII and grade VIII have the KMO-MSA >0.05 and the Bartlett's test significance of <0.05, so that both tests meet the assumptions. This means that the unidimension test can be carried out.

The result of the analysis shows that the factor formed having the eigenvalue of > 1 in the test for grade VII is only one factor with the value of 5.121. The factor formed having the value > 1 is a factor that can be maintained and can be used as an indicator of a trait (Wagiran, 2014, p. 302). The eigenvalue is the highest value among the other eigenvalues so that it is indicated that the mathematics higher order thinking skill test instrument for Grade VII is unidimensional.

For the test for Grade VIII, there are four components having the eigenvalue >1, so that it is indicated that the mathematics test for Grade VIII formed four factors. The test scree plot of Grade VIII test shows that the eigenvalue became slopy starting from the second factor. Other information from the result of the analysis shows that the one dominant factor has the highest eigenvalue, that is, 4.936, so that it is indicated that the mathematics higher order thinking skill test for Grade VIII is unidimensional.

### *Local Independence*

After being proven that the test is unidimensional, the local independence assumption is automatically proven, too (Retnawati, 2014, p. 7). Therefore, the local independence assumption for the tests for grades VII and VIII is met.

*Test Item Analysis*

The item analysis was carried out using the item response theory. The fitness of the model on the output PH2 in the Parscale program can be seen from the item fit statistics. In order to determine the appropriate model, the data were analyzed using the 3-logistic parameter model.

The analysis was carried out by comparing the 1-parameter logistic model, the 2-parameter logistic model, and the 3-parameter logistic model. An item is said to fit with a model when the chi-square observed is lower than the chi-square table or the significance level  $< \alpha$ . The result of the test instrument model fitness analysis for the test for grade VII and grade VIII can be seen in Table 1.

The result of the analysis using the Parscale program shows that the mathematics test instrument for grades VII and VIII is most appropriate using the IRT analysis with the 2-parameter logistic model. This is based on the fact that the highest number of the items fitting the model is in the 2-parameter model.

The result of the analysis using the Parscale program provides information about the item parameter based on the item difficulty index (b) and the discrimination index (a). The difficulty index can be said to be good when it is in the range of -2 to +2 (Baker, 2001, p. 22; DeMars, 2010, p. 21; Hambleton et al., 1991, p. 5).

The result of the analysis of the item difficulty index for grade VII and grade VIII are presented in Table 2. In addition, the result of the item discrimination index for grade VII and grade VIII is presented in Table 3.

From Table 2, it can be seen that the item having the highest difficulty index for Grade VIII is item no 1 with a logit of 0.792 while the item having the lowest difficulty index is item no 6 with a logit of -0.807. The anchor items in the tests both for grades VII and VIII are used for further analysis. The highest difficulty index for the grade VIII test is on item no 9 with a logit of 1.456, while the item with the lowest difficulty index is item no 6 with a logit of -1.095.

Table 1. Model fitness analysis result

No	Model	Grade VII test		Grade VIII test	
		Number of items (prop > 0.05)		Number of items (prop > 0.05)	
1	1 PL	11		9	
2	2 PL	14		13	
3	3 PL	7		7	

Table 2. The analysis of the item difficulty index (location)

Grade VII test			Grade VIII test		
Item	Difficulty index	Category	Item	Difficulty index	Category
1	0.792	Good	1*	0.114	Good
2	-0.017	Good	2*	-0.424	Good
3	0.105	Good	3*	0.114	Good
4	-0.005	Good	4*	0.777	Good
5	0.570	Good	5	0.538	Good
6	-0.807	Good	6	-1.095	Good
7*	0.044	Good	7	0.219	Good
8*	0.072	Good	8	1.145	Good
9*	0.108	Good	9	1.456	Good
10*	-0.036	Good	10	0.092	Good
11	0.498	Good	11	0.863	Good
12	0.393	Good	12	1.043	Good
13	0.592	Good	13	0.552	Good
14	0.069	Good	14	1.164	Good
15	0.470	Good	15	0.207	Good
Mean	0.190			0.451	

\*: Anchor



Table 3 shows that all items in the test for grade VII have good discrimination index. The item having the highest discrimination index is item no 6 with a logit of 0.943, and the item having the lowest discrimination index is item no 11 with a logit of 0.041. The item having the highest discrimination index of the test for grade VIII is item no 1 with a logit of 1.377, and the item with the lowest discrimination index is item no 12 with a logit of 0.147. The item anchor in the tests for grades VII and VIII is used for further analysis.

The test set function would be higher when the test items had high information function. Standard error measurement (SEM) is closely related to the information function.

The higher the information function, the smaller the SEM, and vice versa. The relation between the information function and the SEM is presented in Figure 1 and Figure 2.

Figure 1 shows that the mathematics higher order thinking skill test for Grade VII has a low score in the range between -1.4 and +1.9. This means that the test would provide higher information when it was used to measure the students' ability in the range between -1.4 and +1.9. Figure 2 shows that the test has a higher information function compared with the standard estimation error in the range between -1.2 and +2.5. Therefore, the mathematics tests were appropriate for students having the ability in the range between -1.2 and +2.5.

Table 3. The analysis of the discrimination index parameter

Grade VII test			Grade VIII test		
Item	Discrimination Index	Category	Item	Discrimination Index	Category
1	0.703	Good	1*	1.377	Good
2	0.670	Good	2*	0.313	Good
3	0.767	Good	3*	0.157	Good
4	0.820	Good	4*	0.756	Good
5	0.596	Good	5	0.736	Good
6	0.943	Good	6	0.171	Good
7*	0.606	Good	7	0.844	Good
8*	0.786	Good	8	0.745	Good
9*	0.743	Good	9	0.478	Good
10*	0.812	Good	10	1.015	Good
11	0.401	Good	11	0.928	Good
12	0.834	Good	12	0.147	Good
13	0.745	Good	13	0.352	Good
14	0.683	Good	14	0.696	Good
15	0.405	Good	15	0.791	Good
Mean	0.700		Mean	0.633	

\*: Anchor

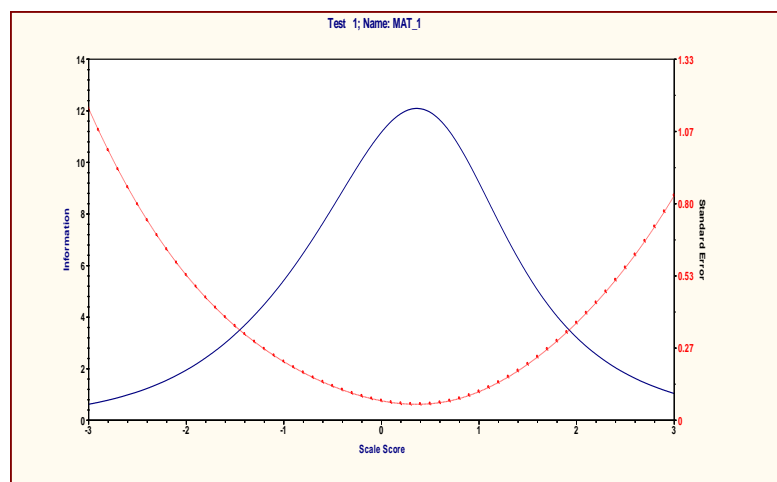


Figure 1. The relation between the information function and SEM of the test for Grade VII

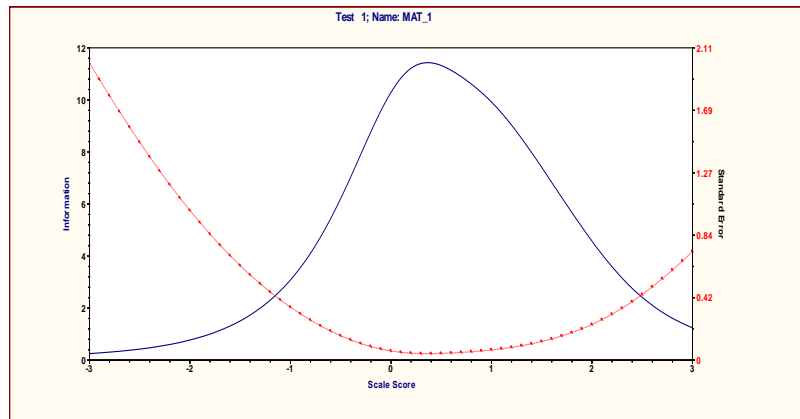


Figure 2. The relation between the information function and SEM of the test for Grade VIII

Note: ----- = SEM  
 ————— = Test information

Table 4. The mean of the slope and the standard error of the equating result

Method	Equating	Slope mean	SD
Mean & mean	Slope of class VII (X)	0.700	0.150
	Slope of class VII (Y*)	0.797	0.171
	Slope of class VIII (Y)	0.634	0.357
	WIT's scale (Y*)	536.241	7.780
	WIT's scale (Y)	528.835	16.233
Mean & sigma	Slope of class VII (X)	0.700	0.150
	Slope of class VII (Y*)	3.689	0.792
	Slope of class VIII (Y)	0.634	0.357
	WIT's scale (Y*)	667.855	36.033
	WIT's scale (Y)	528.835	16.233
TCC Haebara	Slope of class VII (X)	0.700	0.150
	Slope of class VII (Y*)	1.845	0.396
	Slope of class VIII (Y)	0.634	0.357
	WIT's scale (Y*)	583.928	18.017
	WIT's scale (Y)	528.835	16.223
TCC Stocking & Lord	Slope of class VII (X)	0.700	0.150
	Slope of class VII (Y*)	1.230	0.264
	Slope of class VIII (Y)	0.634	0.357
	WIT's scale (Y*)	555.952	12.011
	WIT's scale (Y)	528.835	16.223

### The Equating Result

This vertical equating employed a mixed model with a 2-logistic parameter. The 2-logistic parameter includes equating the difficulty index parameter (b) and the discrimination index parameter (a). The equating method used in this study was the mean-and-mean method, mean-and-sigma method, Haebara characteristics curve, and Stocking-and-Lord characteristics curve. The vertical equating using IRTEQ produced the conversion equation: (1)  $Y'=0.88X-0.27$  in the mean-and-

mean method; (2)  $Y'=0.19X-0.02$  in the mean-and-sigma method; (3)  $Y'=0.38X-0.12$  in the Haebara characteristics curve method; and (4)  $Y'=0.57-0.18$  in the Stocking-and-Lord characteristics curve.

The equating result with the parameter of slope (a) and location (b) in the mean-and-mean method, mean-and-sigma method, the Haebara characteristics curve, and also the Stocking-and-Lord characteristics curve is presented in Table 4 and Table 5, while the result of the calculation of the equating accuracy is presented in Table 6.

Table 5. The location mean and the standard error of the equating result

Method	Equating	Mean of Location	SD
Mean & mean	Location of class VII (X)	0.190	0.385
	Location of class VII (Y*)	-0.103	0.339
	Location of class VIII (Y)	0.451	0.669
	WIT's scale (Y*)	495.317	15.406
	WIT's scale (Y)	520.521	30.448
Mean & sigma	Location of class VII (X)	0.190	0.385
	Location of class VII (Y*)	0.061	0.073
	Location of class VIII (Y)	0.451	0.669
	WIT's scale (Y*)	500.731	3.326
	WIT's scale (Y)	520.521	30.448
TCC Haebara	Location of class VII(X)	0.190	0.385
	Location of class VII(Y*)	-0.048	0.146
	Location of class VIII(Y)	0.451	0.669
	WIT's scale (Y*)	497.823	6.653
	WIT's scale (Y)	520.521	30.448
TCC Stocking & Lord	Location of class VII(X)	0.190	0.385
	Location of class VII(Y*)	-0.072	0.219
	Location of class VIII(Y)	0.451	0.669
	WIT's scale (Y*)	496.734	9.979
	WIT's scale (Y)	520.521	30.448

Table 6. The calculation result of RMSD

Equating	Equating method	RMSD
Class VII to Class VIII	Mean & Mean	0.2955
Class VII to Class VIII	Mean & Sigma	0.8102
Class VII to Class VIII	TCC Haebara	0.631
Class VII to Class VIII	TCC Stocking & Lord	0.466

## Discussion

The equating in this study used the mean-and-mean method, the mean-and-sigma method, the Haebara characteristics curve, and Stocking-and-Lord characteristics curve. The sample used consist of 505 grade VII students and 504 grade VIII students. This was based on the minimum sample measure in item response theory with the 2-logistic parameter, that is, 500 respondents (DeMars, 2010, p. 34). The item anchor used was four items or 26.7%. The number of the anchor influences the test equating result (Kartono, 2008, p. 303). The anchor must be at least 20% of the number of test items (Kolen & Brennan, 2014, p. 288). The test characteristics based on item response theory resulted in the mean of the parameter of the item difficulty index or location (b) in the good category in the range between  $-2 < b < 2$ , that is 0.190 and 0.451 successively. The mean of the discrimination index or slope (a) for grade VII

and Grade VIII was 0.701 and 0.634 successively. Based on the item difficulty index, these items were in a good category because they lied in the range  $-2 < b < 2$ .

The calculation result of the equating constant based on anchor items results in some equations. The equations obtained using the mean-and-mean, mean-and-sigma, Haebara characteristics curve, and Stocking-and-Lord characteristics curve methods are  $Y' = 0.88X - 0.27$ ,  $Y' = 0.19X - 0.02$ ,  $Y' = 0.38X - 0.12$ , and  $Y' = 0.57X - 0.18$  successively.

The findings of the research show that the score conversion of the parameter of location and slope indicate consistent results in the mean-and-mean method, mean-and-sigma method, Haebara characteristics curve method, and also Stocking-and-Lord characteristics curve method. An examples of the equating result using the mean-and-mean method can be seen in item no 5 for grade VII. The equation of the parameter location of Grade VII to Grade VIII is:

$$b^* = 0.88 (b_x) - 0.27.$$

Item no 5 for grade VII has the difficulty index (location) of 0.57 logit. Thus, after being equated to Grade VIII location, it becomes  $b^*=0.232$ . This means that item no 5 Grade VII has the location of 0.57 logit being equal with the location value of 0.232 in the item for Grade VIII. The result of the equating of the item difficulty index (location) shows that the item difficulty index for Grade VIII experiences a decrease after being converted to Grade VIII. If the item no 5 for Grade VII was done by students in Grade VIII, the VIII grade students would find it easier.

The result of the equating of location in the four methods shows that the test for Grade VIII is more difficult than the test for Grade VII. This information can be seen from the comparison table of the parameter scores of the item difficulty index before and after being equated in each method. The mean score of the item difficulty index parameter decreases when converted to a higher scale. This means that the test for Grade VII is easier when it is done by Grade VIII students. On the other hand, the test for Grade VIII students would be more difficult when it was done by Grade VII students.

The equating result of the discrimination index parameter scores (slope) can be illustrated in one of the items of the test for Grade VII. The discrimination index parameter conversion equation from Grade VII to Grade VIII with the mean-and-mean method is:

$$a^* = \frac{a_x}{0.88}$$

Item no 5 for Grade VII has the discrimination index of 0.596 logit. Therefore, after being equated to Grade VIII, the discrimination index would be  $b^*=0.677$ . This means that the discrimination index for Grade VII, that is, 0.596 is equal with the discrimination index of 0.677 for Grade VIII. The equated discrimination index of item no 5 increases after being equated to Grade VIII. All the four methods provide consistent information. This can be seen on the comparison of the mean of the discrimination index of the items for Grade VII and for Grade VIII which has been equated (in WIT's scale). It is indicated

that after being equated, the test instrument for Grade VII has a higher discrimination index than the test instrument for Grade VIII.

The method which provides the smallest error in the equating using 2-logistic parameter is the mean-and-mean method. This, then, is followed by the Stocking-and-Lord method, the Haebara method, and the mean-and-sigma method successively. Baker and Al-Karni (1991) state that the mean-and-mean method has better accuracy than the Stocking-and-Lord method. A study which was conducted by Kartono (2008) concludes that the equating using the mean and mean method is one level better than the mean and sigma method. Sugeng (2010, p. 289) states that the mean and mean method tends to provide more accurate information than the IRT vertical equating using the partial credit model. Uysal and Kilmen (2016) present their study stating that the Stocking-and-Lord method has a smaller error than the Haebara method. The mean-and-sigma method results in the biggest error, while the sample size and the distribution of the students' ability influences the RMSD value (Kilmen & Demirtasli, 2012, p. 130).

## Conclusion and Suggestions

### Conclusion

The characteristics based in the item response theory results in the mean of the parameter value of the item difficulty index or location (b) which is categorized as good in the range of  $-2 < b < 2$ . The parameter values are 0.190 and 0.451. The mean of the discrimination index parameter value or slope (a) for the tests for Grades VII and VIII are 0.701 and 0.634 successively.

The equating results in four equations based on the method used, that are,  $Y^*=0.88X-0.27$  using the mean-and-mean method,  $Y^*=0.19X-0.02$  using the mean-and-sigma method,  $Y^*=0.38X-0.12$  using the Haebara characteristics curve method, and  $Y^*=0.57X-0.18$  using the Stocking-and-Lord characteristics curve method.

The calculation of the equating accuracy results in the Root Mean Square Difference (RMSD) of the mean-and-mean method, the

mean-and-sigma method, the Haebara characteristics curve method, and the Stocking-and-Lord characteristics curve method of 0.2955, 0.8102, 0.6315, and 0.466 successively. The mean-and-mean provides the smallest RMSD, followed by the Stocking-and-Lord characteristics curve method, the Haebara characteristics curve method, and the mean-and-sigma method.

#### Suggestions

The study related to the students' mathematics higher order thinking skill development is still limited, that is, it is only concerned with the ability development of grade VII and grade VIII students. Further studies need to be carried out for the ability development from grade VII to grade VIII and from grade VIII to grade IX. In addition, it is suggested that the use of the methods be studied further with different logistic parameters and different lengths of tests to get more accurate information.

The students' mathematics higher order thinking skill ability in Deli Serdang District in this study is still low. Teachers are expected to teach materials with varied cognitive domain as suggested by the curriculum implemented in the schools.

The school principals play an important role in the advancement of the educational institution so that it is suggested that every year a test to know the students' higher order thinking skill development be administered. In addition, it is also necessary for the school to provide a kind of training for the teachers to analyze test items using the classic and modern analyses so that they can develop better items to depict the students' ability in different grades.

#### References

Antara, A. A. P., & Bastari, B. (2015). Penyetaraan vertikal dengan pendekatan klasik dan item response theory pada siswa sekolah dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 19(1), 13–24. <https://doi.org/10.21831/pep.v19i1.4551>

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

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, 28(2), 147–162. <https://doi.org/10.1111/j.1745-3984.1991.tb00350.x>

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. New York, NY: Oxford University Press.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage Publications.

Gagne, R. M. (1977). *The conditions of learning*. New York, NY: Holt, Rinehart, and Winston.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Newburg Park, LA: Sage Publication ICC.

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

Han, K. T. (2009). IRTEQ: Windows application that implements item response theory scaling and equating. *Applied Psychological Measurement*, 33(6), 491–493. <https://doi.org/10.1177/0146621608319513>

Istiyono, E. (2016). The application of GPCM on MMC test as a fair alternative assessment model in physics learning. In *Proceeding of the 3rd International Conference on Research, Implementation and Education of Mathematics and Science (ICRIEMS), 16-17 May 2017* (pp. 25–30). Yogyakarta: Universitas Negeri Yogyakarta. Retrieved from <http://seminar.uny.ac.id/icriems/sites/>

- seminar.uny.ac.id/icriems/files/prosidin  
g/PE-04.pdf
- Kartono, K. (2008). Penyetaraan tes model campuran butir dikotomus dan politomus pada tes prestasi belajar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 12(2), 302–320. <https://doi.org/10.21831/pep.v12i2.1433>
- Kilmen, S., & Demirtasli, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences*, 46, 130–134. <https://doi.org/10.1016/J.SBSPRO.2012.05.081>
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer New York.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer New York.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.
- Law No. 14 of 2005 of Republic of Indonesia about Teachers and Lecturers (2005).
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).
- Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.
- Stiggins, R. J., & Chappuis, J. (2012). *An introduction to student-involved assessment for learning*. Boston, MA: Pearson.
- Sugeng, S. (2010). Penyetaraan vertikal model kredit parsial soal matematika SMP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 14(2), 289–308. <https://doi.org/10.21831/pep.v14i2.1083>
- Uysal, I., & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, 8(2), 1–11.
- Wagiran. (2014). *Metode penelitian pendidikan: Teori dan implementasi*. Yogyakarta: Deepublish.



## 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].
- All Author(s)' names and identity(es) must be completely embedded in the form filled in by the corresponding author: email; affiliation; and country. [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](mailto:reid.ppsuny@uny.ac.id) or [reid.ppsuny@gmail.com](mailto:reid.ppsuny@gmail.com)