REiD (RESEARCH AND EVALUATION IN EDUCATION)
Vol. 5, No. 2, December 2019

Creative thinking ability and cognitive knowledge: Big Five personality
--Jonni Sitorus; Nirwana Anas; Ermaliana Waruhu

Estimation of college students' ability on real analysis course using Rasch model
--Isnaini; Wikan Budi Utami; Purwo Susongko; Herani Tri Lestiani

An authentic assessment model to assess kindergarten students' character
--Umi Faizah; Darmiyati Zuchdi; Yasir Alsamiri

Psychometric characteristic of positive affect scale within the academic setting
--Kartika Nur Fathiyah; Asmadi Alsa; Diana Setiyawati

Evaluation of the implementation of Batik-skills training program
--Hendro Prasetyono; Dedeh Kurniasari; Laila Desnaranti

The respondent factors on the digital questionnaire responses
--Muhardis; Burhanuddin Tola; Herwindo Haribowo

The effectiveness of Game-Based Science Learning (GBSL) to improve students' academic achievement: A meta-analysis of current research from 2010 to 2017
--Heru Setiawan; Shane Phillipson

Parallel tests viewed from the arrangement of item numbers and alternative answers
--Badrun Kartowagiran; Djemari Mardapi; Dian Normalitasari Purnama; Kriswantoro

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

9 772460 699001

REiD

Research and Evaluation in Education

Vol. 5, No. 2, December 2019

Vol. 5, No. 2, December 2019

ISSN 2460-6995

Research and Evaluation in Education

REiD

Publisher:
PROGRAM PASCASARJANA
UNIVERSITAS NEGERI YOGYAKARTA

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

Publisher
Program Pascasarjana Universitas Negeri Yogyakarta

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

Editorial
Department of Educational Research and Evaluation, Graduate School of Yogyakarta State University
2nd Floor of General Lecture Building, Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281
Telephone: 0274 586168 ext. 229 or 0274 550836, Facsimile: 0274 520326
E-mail: reid.ppsuny@uny.ac.id, reid.ppsuny@gmail.com

Foreword

We are very pleased that REiD (Research and Evaluation in Education) is releasing its tenth edition. We are also very excited that the journal has been attracting papers from the foreign countries, Saudi Arabia and Australia. 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 natural sciences, mathematics, language education, social sciences, and communication program, with focuses on assessment and evaluation.

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

Yogyakarta, December 2019

Editor in Chief

TABLE OF CONTENT

Jonni Sitorus *Nirwana Anas* *Ermaliana Waruhu*	Creative thinking ability and cognitive knowledge: Big Five personality	85-94
Isnaini *Wikan Budi Utami* *Purwo Susongko* *Herani Tri Lestiani*	Estimation of college students' ability on real analysis course using Rasch model	95-102
Umi Faizah *Darmiyati Zuchdi* *Yasir Alsamiri*	An authentic assessment model to assess kindergarten students' character	103-119
Kartika Nur Fathiyah *Asmadi Alsa* *Diana Setiyawati*	Psychometric characteristic of positive affect scale within the academic setting	120-129
Hendro Prasetyono *Dedeh Kurniasari* *Laila Desnaranti*	Evaluation of the implementation of Batik-skills training program	130-143
Muhardis *Burhanuddin Tola* *Herwindo Haribowo*	The respondent factors on the digital questionnaire responses	144-151
Heru Setiawan *Shane Phillipson*	The effectiveness of Game-Based Science Learning (GBSL) to improve students' academic achievement: A meta-analysis of current research from 2010 to 2017	152-168
Badrun Kartowagiran *Djemari Mardapi* *Dian Normalitasari Purnama* *Kriswantoro*	Parallel tests viewed from the arrangement of item numbers and alternative answers	169-182

REiD (Research and Evaluation in Education), 5(2), 2019, 85-94
Available online at: http://journal.uny.ac.id/index.php/reid

Creative thinking ability and cognitive knowledge: Big Five personality

***[1]Jonni Sitorus; [2]Nirwana Anas; [3]Ermaliana Waruhu**
[1]Badan Penelitian dan Pengembangan Provinsi Sumatera Utara
Jl. Sisingamangaraja No.198, Siti Rejo I, Kota Medan, Sumatera Utara 20216, Indonesia
[2]Faculty of Tarbiyah Science and Teacher Training, Universitas Islam Negeri Sumatera Utara
Jl. William Iskandar Ps. V, Kenangan Baru, Kab. Deli Serdang, Sumatera Utara 20371, Indonesia
[3]Sekolah Dasar Negeri 014648 Padang Mahondang
Padang Mahondang, Pulau Rakyat, Kabupaten Asahan, Sumatera Utara 21273, Indonesia
*Corresponding Author. E-mail: sitorus_jonni@yahoo.co.id

Submitted: 04 January 2019 | *Revised: 27 May 2019* | *Accepted: 05 August 2019*

Abstract

This research aims at describing the ability and level of student's creative thinking and student's cognitive knowledge. It is qualitative research to search for data and information. Operationally, this research was conducted with some steps namely: (1) giving a set of big five personality test to 215 students to determine their personality type, (2) giving a set of creative thinking test of 215 students to measure the ability and level of their creative thinking, and (3) choosing one student randomly from each student's personality type to be interviewed to search their cognitive knowledge. The results show that every student has a creative thinking ability, but the level of creative thinking varies. The category of student's creative thinking ability based on Big Five Personalities is 'moderate or high'. The level of student's creative thinking based on the big five personality is 'very creative, creative, quite creative or less creative'. The student's cognitive knowledge based on the big five personality is drawing, designing, ascertaining, dividing, reasoning, analogy, imagining, utilizing, solving, understanding, determining, mentioning, and using trial and error.

Keywords: *creative thinking ability, cognitive knowledge, Big Five personality, creative thinking level, novel answer*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.22848

Introduction

Creativity is not only a result but can also blossom other cognitive functions, such as cognitive thinking (Silvia, 2008). Most of the cognitive theory models focus on the ability to think and solve problems creatively. The cognitive theory model provides a place for psychometric procedures to understand various cognitive abilities, including creative thinking ability (Batey & Furnham, 2006).

Creative thinking is the process of understanding difficulty, problem, information gaps, loose elements, and inconsistency; formulating the problem clearly; supposing or formulating hypotheses about deficiency; examining the hypotheses and possibilities of revising and re-examining or redefining the problem, and ultimately communicating the results. Creative thinking is an individual ability based on its uniqueness to generate worth and novel ideas. The formulation of creativity that emphasizes creative thinking ability is known as the major impetus for the research of creativity (Santrock, 2003).

The ability to think creatively is closely related to intelligence abilities or cognitive traits (Setiawan, 2016). Cognitive traits include fluency, flexibility, originality, elaboration, and also many affective traits (Setiawan, 2017; Wolfradt & Pretz, 2001): curiosity, courage to take risks, challenged by plurality, and imaginative. The primary aptitude traits which are related to creativity, and typically called the characteristics of creative thinking ability (Carson, Peterson, & Higgins, 2005), namely: sensitivity to problems; fluency, includes the fluency of word, expressional, and ideational; flexibility, includes the spontaneous and adaptive flexibility; originality; elaboration; and redefinition. Creativity as an associative function is the ability to connect the objects, experiences, knowledge, and prior information to something new (Mumford, 2003).

Batey, Furnham, and Safiullina (2010) state that there is a positive and negative relationship between creativity with the dimensions of the Big Five personality. Creativity is positively correlated with extraversion and openness dimensions and is negatively related to agreeableness, conscientiousness, and neuroticism. Individuals with an openness dimension have creative characteristics, broad interests, curious, original, and imaginative. Neuroticism has an anxious, nervous, emotional, insecure, and incompetent dimension.

Creative Thinking Ability

Creative thinking is a process of constructing ideas to gain something new in insights, approach, perspective or way of understanding the problem (Grieshober, 2004; Isaksen, Dorval, & Treffinger, 2000; Martin, 2009; McGregor, 2007). Some indicators of creative thinking are fluency, flexibility, novelty, productivity, impact, success, efficiency, coherence (Briggs & Davis, 2008; Martin, 2009; Santrock, 2007; Sternberg, 2012). Creative thinking is a combination of logical and divergent thinking based on intuition consciousness by caring for fluency, flexibility, and novelty (Pehkonen & Törner, 2004; Siswono, 2004).

Everyone has the potential to think creatively, but the level of creative thinking for each person is different (Alenikov, 2002; Neethling, 2000). Siswono (2004) classifies five creative thinking levels: level 4 (very creative), the student can solve the problem by finding more than one novel solution; level 3 (creative), the student can solve the problem by only finding one novel solution; level 2 (quite creative), the student can solve the problem by finding more than one flexible solution; level 1 (less creative), the student can solve the problem by only finding one flexible solution; and level 0 (not creative), the student is unable to solve the problem.

Fluency traits include sparking many ideas, answers, problem-solving, or questions fluently, providing many ways or suggestions for doing things, and always think of more than one answer. The flexibility traits include generating various ideas, answers, or questions, being able to see a problem from different perspectives, searching for many different alternatives or directions, and being able to change the approach or way of thinking. The originality traits include generating something new and unique, thinking of unconventional ways to express oneself, and being able to make unusual combinations of parts or elements.

Novelty is not idea really new, but new for the student (Briggs & Davis, 2008). The novelty concept must be returned to the student's knowledge condition and cannot be generalized to all conditions. Choi (2004) informs that novelty relates to a new experience, where the novelty level is an incompatibility function between the past and the present experience. The novelty of the concept of problem-solving is the student's ability to solve problems by giving several different and correct answers or one unusual answer, which is adjusted to student's knowledge level. Different answer refers to the answer looks different and does not follow a certain pattern.

Big Five Personality

Personality is a dynamic organization or composition from the psychophysical system as unique individual characteristics (feeling, thought, behavior, physical, intelligence, or mood), settled at someone to adjust to the environment (Feist & Feist, 2006).

One of the approaches to measure psychology personality type is the Big Five Personality, which has five personality dimensions, namely: extraversion, agreeableness, conscientiousness, neuroticism, and openness (Friedman & Schustack, 2008). Raymond B.

Cattell is a first theorist in measuring the personality, which is then developed into the basic form of personality structure, better known as the Big Five Personality nowadays.

The characteristics of the Big Five Personality are: (1) extraversion, a high-score individual tends to be full of affection, cheerful, talkative, gregarious, and loving. Conversely, a low-score individual tends to be self-contained, quiet, passive, and lack the ability to express feeling; (2) agreeableness; a high-score individual tends to have full trust, generous, receptive and kind-hearted. A low-score individual tends to be suspicious, stingy, unfriendly, irritable, more aggressive, critics, and less cooperative; (3) conscientiousness, a high-score individual tends to be hardworking, meticulous, timely, and diligent. A low-score individual tends to be irregular, lax, lazy, aimless, and easily give up when getting difficulty; (4) neuroticism, a high-score individual tends to be anxious, temperamental, self-pitying, self-aware, emotional, and prone to stress disorders. A low-score individual tends to be happier and content, calm, ordinary, self-satisfied, and unemotional; (5) openness, refers to how individual to adjust oneself to a new situation and idea. A high-score individual tends to be easy to tolerate and absorb information, focus, and be alert to feeling, thought, and impulsivity. A low-score individual tends to be narrow-minded, conservative, and does not like change.

Batey et al. (2010) state that there are positive and negative linkages between creativity and the dimension of the Big Five personality. Creativity is positively associated with extraversion and openness dimensions and is negatively related to agreeableness, conscientiousness, and neuroticism.

Cognitive Knowledge

Anderson et al. (2001) state that cognitive taxonomy as a revision of Bloom's Taxonomy refers to memorizing, that is recognizing and recalling; understanding, that is interpreting, exemplifying, classifying, summarizing, comparing and explaining; applying, that is executing and implementing; analyzing, that is differentiating, organizing and attributing; evaluating, that is checking and critique; and creating, that is generating, planning, and producing (Krathwohl, 2002; Smith, 2008).

Moreover, de Lange (2003) asserts that student's cognitive knowledge in the process of mathematics learning is to produce great ideas to solve the mathematical problems; create a mathematical model created by students to solve problems of student's learning creativity; bring up various problem solving; express ideas; connect the mathematics concepts with everyday life; and use mathematics and mathematical mindset in everyday life in various sciences through the practice of acting and mathematical activities on the basis of logical, rational, critical, creative, accurate, honest, effective and efficient.

According to Galbraith and Stillman (Ee & Widjaja, 2013; Stillman, 2015), students' cognitive knowledge when given the problems are to understand and structuralize the problems; simplify and interpret the context; assume, formulate and perform the mathematization process; verify the results by comparing, critique, validating, communicating (Rahayu (2015), justifying, and report on writing; and revise the incorrect answer based on the revision results.

Method

This study is qualitative research. The researchers used the basic statistics (mean & percentage) to get the student's creative thinking ability data and then interviewed some students to get the student's cognitive knowledge data. The research was conducted in March 2017 for seven primary schools in North Sumatra Province, Indonesia.

Population and Sample

The number of research population is 611 sixth class students from seven primary schools in North Sumatra Province. The number of research samples is 215 students chosen randomly, a minimum of 10% of the population (Cohen, Manion, & Morrison, 2007). The sample consists of 98 female students and 117 male students. They are around 12-13 years old.

Research Instrument and Data Collection Technique

Data were collected in two ways, namely: test and in-depth interview. The research instruments are a set of creative thinking tests, big five personality test, and interview guidelines. The researchers used the standard Big Five Personality test and creative thinking test, so they do not need to be validated. The creative thinking test consists of an open-ended and problem-solving item focused on two-dimensional figure material in class VI of primary school for measuring the ability and level of student's creative thinking. The Big Five personality test was used for determining student's personality type. The interview guideline was used for searching student's cognitive knowledge.

Data Analysis Technique

Qualitative and quantitative data were analyzed qualitatively by some phases, namely: coding each data and information obtained from interviews and tests; determining the similarity of data and information obtained from interviews and tests based on different contexts; collaborating on differences in data and information obtained from interviews and tests; classifying and categorizing data and information obtained from interviews and tests, and looking for relationships between each categorization

Research Procedure

First, the researchers gave a set of big five personality test (Mayer, 2003, 2005) to 215 students to determine their personality type. The results are presented in Table 1.

Second, the researchers gave a set of creative thinking tests of 215 students to measure the ability and level of their creative thinking. The creative thinking ability is measured from student's answer fluency. The researchers gave the score of creative thinking ability without differentiating the creative thinking indicator. The score of one correct answer is 1, two correct answers are 2, and so on. The researchers then converted the score of value to categorize the student's creative thinking ability based on 'scale 5', namely: very low (0-54), low (>54-64), moderate (>64-79), high (>79-89) and very high (>89-100). The creative thinking level is measured from student's answer flexibility and novelty. According to Siswono (2004), the creative thinking level is categorized into five, namely: level 4 (very creative), student is able to solve the problem by giving more than one novel answer; level 3 (creative), student is able to solve the problem by only giving one novel answer; level 2 (quite creative), student is able to solve the problem by giving more than one flexible answer; level 1 (less creative), student is able to solve the problem by only giving one flexible answer; and level 0 (not creative), student is unable to solve the problem.

Third, the researchers chose one student randomly from each student's personality type, as shown in Table 1, as a key informant in this research to be interviewed to search their cognitive knowledge. The six students as research informants must have a creative thinking level 'very creative or creative'. The researchers interviewed them by using exploratory and confirmatory approaches.

Table 1. Student's personality type based on big five personality

No.	Trends of Student's Personality Type	Number of Student (Person)	Percentage (%)
1.	Extraversion	28	13.02
2.	Agreeableness	21	9.77
3.	Extraversion + agreeableness + openness	53	24.65
4.	Extraversion + conscientiousness + openness	48	22.33
5.	Extraversion + neoroticism + openness	41	19.07
6.	Agreeableness + openness	24	11.16
	Total	215	100

Findings and Discussion

Findings

Student's creative thinking ability based on the Big Five personality is shown in Table 2, in which, student's creative thinking ability with personality types of extraversion, agreeableness, or extraversion + neuroticism + openness is under overall mean value (75.50). Student's creative thinking ability that has personality types of extraversion + agreeableness + openness, extraversion + conscientiousness + openness, or agreeableness + openness is over the overall mean value (75.50). The mean difference of student's creative thinking ability for each personality type is shown in Table 3. Student's creative thinking level based on the Big Five personality can be seen in Table 4.

Table 2. Student's creative thinking ability based on big five personality

No.	Trends of Student's Personality Type	Mean Value	Category of Creative Thinking Ability
1.	Extraversion	65.82	Moderate
2.	Agreeableness	70.04	Moderate
3.	Extraversion + agreeableness + openness	83.73	High
4.	Extraversion + conscientiousness + openness	78.66	Moderate
5.	Agreeableness + openness	80.51	High
6.	Extraversion + neoroticism + openness	74.22	Moderate
	Overall mean value	75.50	Moderate

Table 3. Mean difference in student's creative thinking ability for each personality type trend

No.	Trends of Student's Personality Type		Mean Difference
1.	Extraversion	Agreeableness	4.22
		Extraversion + agreeableness + openness	17.91
		Extraversion + conscientiousness + openness	12.84
		Agreeableness + openness	14.69
		Extraversion + neoroticism + openness	8.4
2.	Agreeableness	Extraversion + agreeableness + openness	13.69
		Extraversion + conscientiousness + openness	8.62
		Agreeableness + openness	10.47
		Extraversion + neoroticism + openness	4.18
3.	Extraversion + agreeableness + openness	Extraversion + conscientiousness + openness	5.07
		Agreeableness + openness	3.22
		Extraversion + neoroticism + openness	9.51
4.	Extraversion + conscientiousness + openness	Agreeableness + openness	1.85
		Extraversion + neoroticism + openness	4.44
5.	Agreeableness + openness	Extraversion + neoroticism + openness	6.29

Table 4. The number of student based on creative thinking level

No.	Trends of Student's Personality Type	The Number of Student based on Creative Thinking Level (Person)					Total
		Level 4 (> 1 novel answer)	Level 3 (1 novel answer)	Level 2 (> 1 flexible answer)	Level 1 (1 flexible answer)	Level 0 (none)	
1.	Extraversion	6	18	3	1	-	28
2.	Agreeableness	3	12	4	2	-	21
3.	Extraversion + agreeableness + openness	20	15	18	-	-	53
4.	Extraversion + conscientiousness + openness	18	19	3	8	-	48
5.	Extraversion + neoroticism + openness	22	17	2	-	-	41
6.	Agreeableness + openness	14	8	1	1	-	24
	Total	83	89	31	12	-	215

Note: - = No student

Referring to Table 4, students with personality type 'extraversion, agreeableness, Extraversion + conscientiousness + openness, or agreeableness + openness' are very creative, creative, quite creative, or less creative. The students having personality type 'extraversion + agreeableness + openness, or extraversion + neuroticism + openness' are very creative, creative, or quite creative. Overall, there are 83 very creative students (38.60%), 89 creative students (41.40%), 31 quite creative students (14.42%), and 12 less creative students (5.58%). No student is uncreative.

Student's flexible or novel answers from their answer sheets as their creative products can be counted and decided in the following ways. One of the problems solved by students is to divide a rectangle into two equal-area parts of unique and various forms. For instance, students divided a rectangle into two equal rectangle area parts; two equal triangle area parts; two equal trapezoidal area parts; two equal two-dimensional area parts shaped zigzag and circular. It means that the student got five correct answer alternatives: rectangle, triangle, trapezoidal, two-dimensional zigzag shape, and two-dimensional circular shaped. The number of student's flexible answer is two, namely: triangle and trapezoidal because their shapes are different from the original one, but not unique. The number of student's novel answer is two, namely: two-dimensional zigzag shape, and two-dimensional a circular shape, because they are unique. The rectangle divided into two equal rectangle area parts is not flexible nor novel answer because the shape is the same as the original one.

To search for data and information on student's cognitive knowledge by in-depth interviews, the researchers chose one student from each personality type, as shown in Table 1. The student with personality type 'extraversion' is called Student S1; 'agreeableness' is called Student S2; 'extraversion + agreeableness + openness' is called Student S3; 'extraversion + conscientiousness + openness' is called Student S4; 'extraversion + neuroticism + openness' is called Student S5; and 'agreeableness + openness' is called Student S6.

Based on interview with Student S1, she could draw and design various two-dimensional figures of unique shapes by cutting, folding, and measuring the rectangle into two parts in equal size and area. When the researchers asked her how she ascertained the two parts equal, she just said that if the two-dimensional figure is divided into two parts in equal size, they must have equal area; it means she uses her math reasoning. When the researchers asked her how she is able to draw and design the two-dimensional polygon figure, she just said that she used her imagination to create creativity, it means she imagined the relevant things to find creative ideas.

Based on interview results from Student S2, he divided a rectangle into two parts of the unique two-dimensional figure in the equal area by utilizing his intuitions ability. He had no relevant learning experience previously. He did not also cheat or ask his friends. It means that he solved the problem by his own conscience. Student S3 divided a rectangle into two parts of trapezoidal in the equal area as one of his answers on the answer sheet. He said that a trapezoidal has a pair of facing lines of a parallel position. It means that he really understands the concept of trapezoidal.

Student S4 & Student S5 divided a rectangle into two parts of the triangle in the equal area as one of their alternative answers. They determined the two-dimensional figure area by using formula. They understand the concept of the triangle by mentioning that one of the angles of the right triangle is 900. They also divided a rectangle into two parts of the unique two-dimensional figure in the equal area as another answer by using a trial and error system. They also utilized their intuition ability to find creative ideas. Student S6 divided a rectangle into two parts of the two-dimensional figure in the equal area as one of her alternative answers by utilizing her prior knowledge and previous learning experience. According to her, their teacher ever taught and gave a similar problem, meaning that she made an analogy to solve the problem. Referring to student's cognitive knowledge description, the researchers try to summarize them, shown in Table 5.

Table 5. Student's cognitive knowledge

No.	Student's Cognitive Knowledge
1.	Draw and design various two-dimentional figures in unique shapes
2.	Ascertain the two parts of unique two-dimentional figure of equal area
3.	Divide the two-dimentional figures in two equal size parts
4.	Reasoning
5.	Imagine the relevant things to create creativity and find creative ideas
6.	Utilize the intuition ability, prior knowledge and previous learning experience
7.	Solve the problem by his own conscience
8.	Understand the concept of two-dimentional figures
9.	Determine the two-dimentional figures area by using formula
10.	Mention the characteristic of two-dimentional figure
11.	Use trial and error system to determine the unique two-dimentional figure in equal area
12.	Analogy

Discussion

One of the results and findings of this research is that the students have the creativity and the creative thinking ability to find the novel answers. It is in line with the opinion of Munandar (1999) that creativity is defined as the ability to create new combinations based on the existing data, information, or elements, and find possibly many answers to one problem, where the emphasis is on the quantity, usability, and diversity of answers. Creativity is the ability to reflect the answer originality.

The ability to draw the unique and novel two-dimensional constructed through student's creative ideas at the research findings is also in line with Isaksen's et al. opinion (Grieshober, 2004) that the creative thinking is an idea-building process that emphasizes on the indicators of fluency, flexibility, novelty, and elaboration. Creative thinking tends to the acquisition of insights, approaches, perspectives, or new ways of under-standing one mathematics problem (McGregor, 2007). Martin (2009), an inflexible-thinking individual uneasily changes his/her ideas or views even though he/she knows any contradiction between the belonging of a new idea.

According to Sharp (Briggs & Davis, 2008), novelty is not idea really new, but new for students. It is also found in this research where the student's answers are only the pentagon, which is actually not two-dimensional, really original from the student's new idea, but the student himself who only drew such two-dimensional in the class. It means that the student's answer has been categorized as a new one if compared to other students' answers. When the student finds this solution to problems with the first time, he has found something new, at least for himself.

Every student has a creative thinking ability, but the level of creative thinking varies. It can be seen by the creation of evidence of certain people in extraordinary technology and knowledge. On the other hand, some people cannot be creative; they have no knowledge or skills at all or only use others' creativity. This state indicates the level or degree of creativity or the creative ability of someone is different. The level of someone's creative thinking can be viewed as a continuum from the lowest to the highest one. If an individual is taken randomly, we can place him/her in the continuum of the creative thinking level. However, because the number of discreet individuals is considerable, the approach to know the degree of creative thinking is a discrete and hierarchical classification.

Students' personality types influence the ability and level of their creative thinking as the research findings. It is in line with the opinion of Ivcevic and Mayer (2006), stating that personality types can differentiate someone's creative thinking ability. An individual's creativity may differ based on his/her personality differences. Personality can be defined as a psychological attributes system that describes how someone feels, thinks, interacts with the social world, and regulates behavior (Funder, 2001; Mayer, 2005).

In the last few decades, the Big Five Personality has become the dominant model for describing broad personality traits. The openness of the Big Five Personality is theoretically and empirically defined as a general disposition of creativity.

Creativity is also related to a narrower nature in the area of emotion and motivation, cognition, social expression, and self-regulation. The behavior of emotion and motivation in the creative thinking process offers an opportunity to be creative and can be a source of creative ideas. For example, motivated people intrinsically engage in activities because of their happiness in creating or enjoying the opportunity for expression. Another behavior related to creativity is hypomania, which can enhance creativity, creative potential (e.g., self-perceived creativity), and creative behavior (e.g., involvement in creative activities). The mood increases the awareness, fluency, and flexibility of thinking.

Cognitive knowledge enables someone to generate creativity. One of the student's cognitive knowledge as the research findings is the reasoning ability. The students use their reasoning ability and can be improved by the creative and innovative learning approaches and require them to be more active and skilled in the learning process. The learning approach factor gives a significant influence on the improvement of students' mathematics reasoning ability, either whole or based on the subgroup of students.

The student's ability to determine the two-dimensional area on the research findings is in line with the opinion of Van de Walle, Karp, and Bay-Williams (2008). According to them, a common mistake that often made by students is the using of incorrect-formula to conceptualize the height and pedestal of two-dimensional. According to Bahr and Bossé (2008), students must learn mathematics by understanding, actively build new knowledge from previous experience and knowledge. Learning by understanding is important to enabling students to solve the new problems that will inevitably face in the future.

The student's mathematics intuition as the research findings is in line with some opinions which state that the creative thinking in the mathematics subject is a combination of the logical and divergent thinking based on intuition with the indicators of fluency, flexibility, and novelty, one of the creative personal characters is characterized by an intuition ability — an individual needs two mathematical thinking skills, namely: the abilities of intuition and analytic thinking.

Conclusion

Based on the findings and discussion, the researchers conclude some points of the research. The category of student's creative thinking ability based on the Big Five personality is 'moderate or high'. The level of student's creative thinking based on the Big Five personality is 'very creative, creative, quite creative or less creative'. The student's cognitive knowledge based on the Big Five personality is drawing, designing, ascertaining, dividing, reasoning, analogy, imagining, utilizing, solving, understanding, determining, mentioning and using trial and error.

References

Alenikov, A. (Ed.). (2002). *The future of creativity*. Bensenville, IL: Scholastic Testing Press.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., … Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Bahr, D. L., & Bossé, M. J. (2008). The state of balance between procedural knowledge and conceptual understanding in mathematics teacher education. *International Journal of Mathematics Teaching and Learning*, 1–28. Retrieved from http://hdl.lib.byu.edu/1877/2880

Batey, M., & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered lterature. *Genetic, Social, and General Psychology Monographs*, *132*(4), 355–429. https://doi.org/10.3200/MONO.132.4.355-430

Batey, M., Furnham, A., & Safiullina, X. (2010). Intelligence, general knowledge and personality as predictors of creativity. *Learning and Individual Differences*, *20*(5), 532–535. https://doi.org/10.1016/j.lindif.2010.04.008

Briggs, M., & Davis, S. (2008). *Creative teaching: Mathematics in the early years and primary classroom*. London: Routledge.

Carson, S. H., Peterson, J. B., & Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, *17*(1), 37–50. https://doi.org/10.1207/s15326934crj1701_4

Choi, J. N. (2004). Individual and contextual predictors of creative performance: The mediating role of psychological processes. *Creativity Research Journal*, *16*(2–3), 187–199. https://doi.org/10.1080/10400419.2004.9651452

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). New York, NY: Routledge.

de Lange, J. (2003). Mathematics for literacy. In The National Council on Education and the Disciplines (Ed.), *Quantitative literacy: Why numeracy matters for schools and colleges* (pp. 75–89). Retrieved from https://www.maa.org/sites/default/files/pdf/QL/WhyNumeracyMatters.pdf

Ee, D. N. K., & Widjaja, W. (2013). Mathematical modelling in the primary school: Elements in teacher education. *5th Redesigning Pedagogy International Conference*. Singapore: National Institute of Education.

Feist, J., & Feist, G. J. (2006). *Theories of personality*. Boston, MA: McGraw-Hill Education.

Friedman, H. S., & Schustack, M. W. (Eds.). (2008). *The personality reader* (2nd ed.). Boston, MA: Allyn and Bacon.

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*(1), 197–221. https://doi.org/10.1146/annurev.psych.52.1.197

Grieshober, W. E. (2004). *Continuing a dictionary of creative term and definitions*. Buffalo, NY.

Isaksen, S. G., Dorval, K. B., & Treffinger, D. J. (2000). *Creative approaches to problem solving: A framework for change* (2nd ed.). Dubuque, IA: Kendall/Hunt.

Ivcevic, Z., & Mayer, J. D. (2006). Creative types and personality. *Imagination, Cognition and Personality*, *26*(1), 65–86. https://doi.org/10.2190/0615-6262-G582-853U

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Martin, P. N. (2009). Societal transformation and reference services in the academic library: Theoretical foundations for re-envisioning reference. *Library Philosophy and Practice*, (May), 1–8. Retrieved from https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1265&context=libphilprac

Mayer, J. D. (2003). Structural divisions of personality and the classification of traits. *Review of General Psychology*, *7*(4), 381–401. https://doi.org/10.1037/1089-2680.7.4.381

Mayer, J. D. (2005). A tale of two visions: Can a new view of personality help integrate psychology? *American Psychologist*, *60*(4), 294–307. https://doi.org/10.1037/0003-066X.60.4.294

McGregor, S. L. T. (2007). International Journal of Consumer Studies: Decade review (1997-2006). *International Journal of Consumer Studies*, *31*(1), 2–18. https://doi.org/10.1111/j.1470-6431.2006.00566.x

Mumford, M. D. (2003). Where have we been, where are we going? Taking stock in creativity research. *Creativity Research Journal*, *15*(2–3), 107–120. https://doi.org/10.1080/10400419.2003.9651403

Munandar, U. (1999). *Kreativitas dan keterbakatan, strategi mewujudkan potensi* *kreatif dan bakat*. Jakarta: Gramedia Pustaka Utama.

Neethling, K. (2000). The beyonders. In E. P. Torrance (Ed.), *On the edge and keeping on the edge* (pp. 153–166). Bensenville, IL: Scholastic Testing Press.

Pehkonen, E., & Törner, G. (2004). Methodological considerations on investigating teachers' beliefs of mathematics and its teaching. *NOMAD - Nordic Studies in Mathematics Education*, *9*(1), 21–49.

Rahayu, S. (2015). Pembelajaran matematika dengan pendekatan PMRI memang beda. *Buletin*, *VI*(February). Retrieved from http://www.pmri.or.id/main.php

Santrock, J. W. (2003). *Psychology* (7th ed.). New York, NY: McGraw-Hill.

Santrock, J. W. (2007). *Child development* (11th ed.). New York, NY: McGraw-Hill.

Setiawan, R. (2016). Construct of creative thinking assessment on divergent and convergent ability. *International Journal of Advance Research and Innovative Ideas in Education*, *2*(4), 1034–1041. Retrieved from http://ijariie.com/FormDetails.aspx?MenuScriptId=1904

Setiawan, R. (2017). The influence of income, experience, and academic qualification on the early childhood education teachers' creativity in Semarang, Indonesia. *International Journal of Instruction*, *10*(4), 39–50. https://doi.org/10.12973/iji.2017.1043a

Silvia, P. J. (2008). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual Differences*, *44*(4), 1012–1021. https://doi.org/10.1016/j.paid.2007.10.027

Siswono, T. T. (2004). *Pendekatan pembelajaran matematika*. Jakarta: Departemen Pendidikan Nasional Republik Indonesia.

Smith, M. K. (2008). Howard Gardner, multiple intelligences and education. Retrieved from The Encyclopedia of Pedagogy and Informal Education (Infed.org) website: https://infed.org/mobi/howard-gardner-multiple-intelligences-and-education/

Sternberg, R. J. (2012). The assessment of creativity: An investment-based approach. *Creativity Research Journal*, *24*(1), 3–12. https://doi.org/10.1080/10400419.2012.652925

Stillman, G. A. (2015). Applications and modelling research in secondary classrooms: What have we learnt? In S. J. Cho (Ed.), *Selected regular lectures from the 12th International Congress on Mathematical Education* (pp. 791–805). https://doi.org/10.1007/978-3-319-17187-6

Van de Walle, J. A., Karp, K. S., & Bay-Williams, J. M. (2008). *Elementary and middle school mathematics: Teaching developmentally*. Boston, MA: Allyn and Bacon.

Wolfradt, U., & Pretz, J. E. (2001). Individual differences in creativity: Personality, story writing, and hobbies. *European Journal of Personality*, *15*(4), 297–310. https://doi.org/10.1002/per.409

REiD (Research and Evaluation in Education), 5(2), 2019, 95-102
Available online at: http://journal.uny.ac.id/index.php/reid

Estimation of college students' ability on real analysis course using Rasch model

[1]Isnaini; *[2]Wikan Budi Utami; [3]Purwo Susongko; [4]Herani Tri Lestiani
[1,2]Mathematics Education Department, Universitas Pancasakti Tegal
Jl. Halmahera Km. 1, Mintaragen, Kec. Tegal Tim., Kota Tegal, Jawa Tengah 52121, Indonesia
[3]Department of Natural Science Education, Universitas Pancasakti Tegal
Jl. Halmahera Km. 1, Mintaragen, Kec. Tegal Tim., Kota Tegal, Jawa Tengah 52121, Indonesia
[4]Mathematics Education Department, Institut Agama Islam Negeri Syekh Nurjati Cirebon
Jl. Perjuangan By Pass Sunyaragi, Kota Cirebon, Jawa Barat 45132, Indonesia
*Corresponding Author. E-mail: wikan.piti@gmail.com

Submitted: 24 August 2018 | Revised: 15 March 2019 | Accepted: 21 August 2019

Abstract

This study is aimed at estimating the difficulty level of essay tests and the accuracy of students' ability in Real Analysis essay test using the Rasch model with the QUEST program and R 3.0.3 package eRm program. The population in this study was all students of the Department of Mathematics Education, Universitas Pancasakti Tegal in the academic year 2016/2017, who were enrolled in the Real Analysis course. The data were analyzed using the R 3.0.3 package eRm program and QUEST program. The students' ability was obtained from the result of the course final exam of the first Real Analysis course. The analysis shows that: (1) by using Rasch model for partial credit scoring, the difficulty level shows that 100% of essay questions in Real Analysis final exam is categorized as difficult, (2) the estimation of students' ability in Real Analysis course using Rasch Model with CML method is better than the estimation of students' ability using Rasch Model with JML approach.

Keywords: *estimation of ability, level of difficulty, Rasch Model, Item Response Theory*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.20924

Introduction

One important component in the formation of quality human resources is education. The most important factor to be able to compete globally in the 21st century is education. According to Mardapi (2012, p. 12), efforts to improve the quality of education can be pursued through improving the quality of learning and the quality of the assessment system. Thus, in the process of education in Higher Education, for example in learning mathematics must strive to implement the learning process and assessment as well as possible. A good process of learning mathematics can certainly be done by providing flexibility for students to develop and explore their abilities.

Today, education in Indonesia is still considered very low, especially for mathematics. Even though mathematics is the main science taught from elementary school to university. This indication can be seen from the low student achievement in each academic year. Ironically, mathematics is a subject that is not liked. Many students are afraid of mathematics. For them, math is like a frightening enemy they want to avoid. Schwartz (2005, p. 1) suggests the basic success of mathematics education is to support the development of intelligence in mathematics from a variety of life conditions. Student's mathematical skills in living conditions at the School can be seen when students take the test. The implementation of the test is basically to assess the success of students during the learning process.

The test is very necessary so that the educator in this case the lecturer can know the student's learning achievement after being given the subject matter in the learning process. Therefore, making a good test needs to be pursued by considering the ability of students, so that the tests carried out as a measuring tool to test student achievement can reflect/ describe the true abilities of students.

Students of the Mathematics Education program at Universitas Pancasakti Tegal all this time consider the most difficult subjects to be Real Analysis. Real Analysis comprises deductive and axiomatic topics. Previous observation on the performance of students of Universitas Pancasakti revealed the students' ability in this course is relatively low. It is indicated by their ability to prove a convergent sequence yet, they found it difficult in solving some problems related to convergent sequence as there are many theorems are included.

Student learning evaluation activities are one of the important tasks that must be done by lecturers. In the field of education, evaluation of student learning achievements is conducted to determine the progress of students in the curriculum that has been taught. One effort to evaluate students is to give examinations in the middle of the semester and at the end of the semester. However, sometimes giving questions that are too difficult or too easy causes it to be difficult for lecturers to distinguish students' abilities. Therefore, an analysis of exam questions is needed in the hope that the exam results present the ability of students.

Evaluation is a series of activities in improving the quality, performance, or productivity of an institution in carrying out its program. Through evaluation, information about what has been achieved and which have not will be obtained, then this information is used to improve a program. According to Tyler (1950), evaluation is a process of determining the extent to which educational goals have been achieved. According to Griffin and Nix (1991), evaluation is a judgment on the value of the measurement results or implications of the measurement results. Tyler emphasizes the achievement of the objectives of a program, while Griffin and Nix emphasize the use of assessment results. Thus, the focus of evaluation is a program or group, and there is a judgment element in determining the success of a program (Mardapi, 2012, p. 4).

The form of real analysis subject evaluation is the midterm and the final semester examination. The test is in the form of a description test, the advantages of the description form test are easy in the preparation. This form of description will also train students in expressing opinions both systematically and logically (Buckley, Winkel, & Leary, 2004). A lecturer will be able to find out where the weaknesses of the students are in the material that has been taught so that they will give input on what things must be improved. Scoring on the description form tests takes a long time and is relatively more difficult so the form of the description test is difficult to use for large-scale tests. An assessment will be meaningful if the results can be used to improve the quality of the learning process. An assessment will be meaningful if the results can be used to improve the quality of the learning process (McMillan, 2005).

The existence of the midterm and final semester exams in the Real Analysis course is to evaluate the ability of students. Some theories and models that can be used to analyze test items are the ones with the Rasch Model.

In this study, Rasch model was employed to analyze test items. According to Imaroh, Susongko, and Isnani (2017), the items parameter does not depend on the sample. Further, Ningsih and Isnani (2010) revealed the different reliability levels of essay test items analyzed using Item Response Theory model (1PL, 2PL, 3PL) and Rasch model.

The concept of objective measurement in the social sciences and the assessment of education, according to Wright and Mok (2004), must have five criteria, namely: (1) producing linear measurements with equal intervals, (2) exact estimation process, (3) identifying inaccurate (misfits) or uncommon items (outliers), (4) able to handle missing data, (5) produce measurements that are independent of the parameters studied. Of the five conditions, so far only the Rasch model can fulfill these five conditions. The quality of measurements in the assessment of education carried out with the Rasch model will have the same quality as the measurements made in the physical dimension in the field of physics (Sumintono & Widhiarso, 2015). In measuring modern test theory, the Rasch model is seen as the most objective measurement model. The use of the Rasch model in measuring education has advantages in specific objectivity and the stability of high grain parameter estimates (Wu & Adams, 2007).

The main characteristic of the Rasch Model is that this model considers all responses of a test taker regardless of the sequence in solving the problems. It means that the level of difficulty of each test item is not necessarily in consecutive order. The main advantage of the Rasch model is that the mental process used by participants in solving the problems is more accurate. Moreover, compared to other models (particularly classical test theory) this model has the ability to predict the missing data based on a systematic response pattern. This model has been applied to mathematics and reading tests, e.g., at the National Assessment of Educational Progress (NAEP) (Susongko, 2014). This model is also suitable for analyzing personality scale responses that have a multi-point scale.

Unlike the Rasch model which includes all responses without considering the sequence in solving the problems, the Gradation model requires sequential responses of the test takers from a low to a high category. In the Gradation model, the level of difficulty of each test item is arranged in sequence, while in classical test theory, the pattern of students' answers is not considered as classical test theory merely considers correct and incorrect answers. Gradation model is suitable for a course that requires regularities or sequential responses of each test item, such as mathematics, physics, and chemistry.

According to Lababa (2008), one of the oldest test theories about behavioral assessment is classical true-score theory. Classical test theory has an easy application. Moreover, it is a practical model to describe how measurement errors can affect the observed score.

Quantitative item analysis emphasizes the analysis of internal test characteristics through empirically obtained data. Internal characteristics include test item parameters which are the level of difficulty and discrimination power of a test.

Rasch model is a dichotomous scoring model that merely has two categories, namely the correct answer with a score of 1 and the incorrect answer with a score of 0. Currently, it has been developed more extensively in polytomous scoring. According to Retnawati (2014, p. 32), the polytomous scoring model is an item response model that has more than two scoring categories. In the Rasch model, it is assumed that all items have the same discrimination index (Isgiyanto, 2011).

To deal with polytomous data with various ranks, a new type of analysis of the Rasch model is developed, namely the Partial Credit Model. However, the main purpose of the Rasch model is to create a scale measurement at equal intervals. Meanwhile, as the raw scores are not shown in interval form, the scores cannot be used directly to interpret the students' ability. Rasch model requires both per person score data and per item score data. These two scores become the basis for estimating true scores that indicate the level of individual ability as well as the degree of difficulty of the test.

Rasch modeling uses both per person score data and per item score data. These two scores become the basis for estimating true scores that indicate the level of individual ability as well as the degree of difficulty of the test. The advantage of the Rasch Model compares to other models, particularly classical test theory, is the ability to predict the missing data, based on a systematic response pattern.

Some studies had been carried out related to the use of the Rasch Model in analyzing test items. A study by Kurniawan and Mardapi (2015) showed that the Rasch model provides complete information about test i-tems, including its difficulty level. This study is aimed at estimating the difficulty level of the essay test on the first Real Analysis course by using the Rasch Model and describing the estimation of students' ability in Real Analysis course by using the Rasch Model, QUEST program, and R 3.0.3 package eRM program.

Method

This research is an explorative descriptive study of data sets of items and responses of participants in the semester's final examination of the real analysis subject in the academic year 2016/2017. This research is a post-hoc diagnosis that is described as a retrofitting approach (Gierl, 2007). The retrofitting approach is carried out through analysis of the items and item response data in the final semester exam in the real Analysis 2016/2017 academic year.

Some studies have implemented the Rasch model by involving 30 to 300 students as the sample (Bond & Fox, 2007; Keeves & Masters, 1999). The subject of this present study was 82 students of Mathematics Education Department of Universitas Pancasakti Tegal in the academic year 2016/2017 who took the first Real Analysis course.

The sampling technique used in this study is purposive sampling. It is one of the non-random sampling techniques where the researcher determines sampling by specifying specific characteristics suitable with the objectives of the study so that it is expected to answer the research problems. Based on the explanation of the purposive sampling, there are two things that are very important in using the sampling technique, namely non-random sampling and setting specific characteristics according to the research objectives by the researchers themselves.

The instrument used in this study was the final exam test on the first Real Analysis course. The test items include the introduction material, Real Numbers, Sequences and Series, and Limit (Bartle & Sherbert, 2000).

Rasch model was applied to analyze the collected data. This analysis resulted in a description of the difficulty level of the test items. By using the eRm package in R Program version 3.0.3, the analysis generated the estimation of item parameters on the exam of Real Analysis.

Measurement modeling explains the procedure of how to organize raw scores into more meaningful information. Moreover, it can utilize a mathematical model that can interpret raw scores into a score that provides more valid and accurate information. The analysis of raw scores leads to a new finding: the opportunity for students to correctly answer an item is the same as the comparison of students' ability and the difficulty level of the test items.

$$P_{i1}(\theta) = \frac{P_{i1}(\theta)}{P_{i0}(\theta) + P_{i1}(\theta)}$$
$$= \frac{exp(\theta_n - \delta_{i1})}{1 + exp(\theta_n - \delta_{i1})} \quad . \; . \; . \quad (1)$$

(Bryan, 2004)

OCFs (Ogive Curve Function) become a prototype of Rasch model development for polytomous items. If i is a polytomous item with score category = 0, 1, 2,. . . , mi, then the probability of participant n with score x on item i is later described in Category Response Function (CRF), which is illustrated in the following equation (Glas & Verhelst, 1989):

$$P_{ix}(\theta) = \frac{exp\left[\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right]}{\sum_{r=0}^{m_i}\left[exp\left[\sum_{j=0}^{x}(\theta_n - \delta_{ij})\right]\right]} \quad ...(2)$$

Equation (2) can be elaborated by the number of categories in the test items. For example, if a scale has three categories of the score of 0, 1, and 2, then there will be a category (j) as many as three individual probability equations for each category. Probability in category 0 is:

$$P_{i0}(\theta) = \frac{1}{1 + \exp(\theta_n - \delta_{i1}) + \exp[(\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2})]} \quad ...(3)$$

Probability in category 1 is:

$$P_{i1}(\theta) = \frac{\exp(\theta_n - \delta_{i1})}{1 + \exp(\theta_n - \delta_{i1}) + \exp[(\theta_n - \delta_{i1}) + (\theta_n - \delta_{i2})]} \quad ...(4)$$

Probability in category 2 is:

$$P_{i2}(\theta) = \frac{\exp(\theta_n - \delta_{i1}) + exp(\theta_n - \ddot{a}_{i2})}{1 + \exp(\grave{e}_n - \ddot{a}_{i1}) + exp[(\grave{e}_n - \ddot{a}_{i1}) + (\grave{e}_n - \ddot{a}_{i2})]} \quad .. \quad (5)$$

In the probability of category 0, there is a number 1 in the numerator since Rasch Model requires the following equation:

$$\sum_{j=0}^{0}(\grave{e} - \ddot{a}_{ij}) = 1 \quad . \; . \quad (6)$$

(Glas & Verhelst, 1989)

Findings and Discussion

The parameter of the difficulty level of test items has the same value interval as the parameter of participants' ability (θ), which is $b_{i\,j}$ = θ. The $b_{i\,j}$ value ranges from -∞ to +∞. However, the values which are practically (or rationally) used are only between -4.0 to +4.0. It means that the more negative the difficulty level of an item or close to -4, the easier the problem. On the other hand, the more positive the difficulty level or approaching +4, the more difficult the problem (Naga, 2003, p. 224).

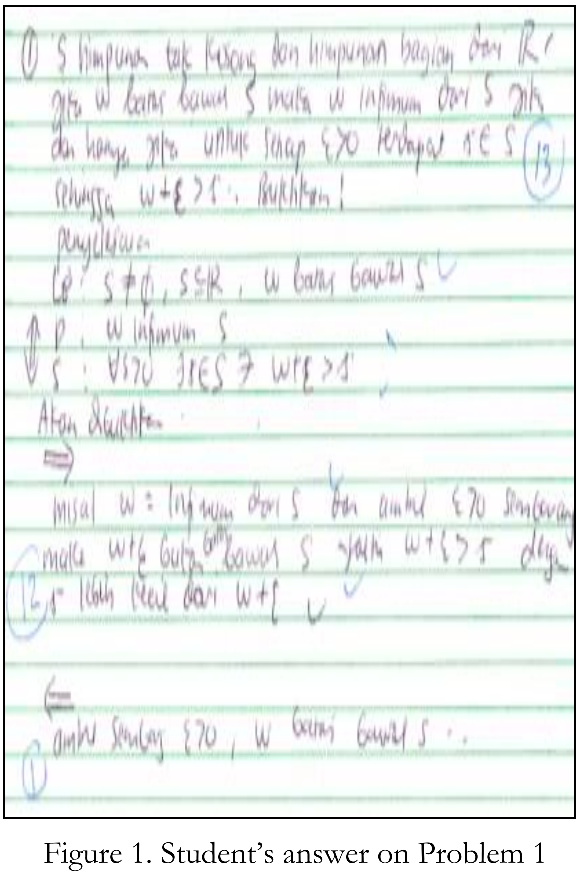
In case the parameter of the difficulty level of a test item meets $b_j \leq -2$, the item is then categorized as a very easy item. If it meets $-2 \leq b_j \leq 0$, the item is then categorized as an easy item. Furthermore, if it meets $0 < b_j \leq 2$ and $b_j \geq 2$, the item is then categorized as a difficult and very difficult item, consecutively (Hambleton, Swaminathan, & Rogers, 1991).

The analysis of the question number 1 showed that δ_{11} = 0.861, δ_{12} = 0.374, and δ_{13} = 0.45. It implies that the difficulty level of the first, second, and third steps is included in the difficult category. In question number 2, the difficulty level of the first step is included in the difficult category (δ_{21}=1.731), while the difficulty level of the second step is identified as very difficult (δ_{22}=2.787). In question number 3, the results obtained were δ_{31}=1.149 and δ_{32}= 1.796, which suggest that the difficulty level of the first and second steps can be included in the difficult category. The analysis of question number 4 resulted δ_{41}=-0.363 and δ_{42}=-0.963. It indicates that the difficulty level in both steps is in included in the easy category.

The results showed that there are three categories (δ_{12}, δ_{21}, δ_{41}) which are identified as easy, one category (δ_{11}) is identified very easy, and six categories (δ_{22}, δ_{31}, δ_{32}, δ_{42}, b_{51}, and *b* δ_{52}) are categorized as difficult. In general, the score of difficulty level of those items was 0.594, thus the four test items were identified as difficult.

It can be inferred from the aforementioned results that the final exam items of Real Analysis course are categorized as difficult for the participants, even though all topics in the questions had been discussed during the course. The value of the difficulty level of item varies (typically) from about -2.0 to +2.0. Item number 1 with sub-topic of the Completeness of Real Numbers was identified as a difficult item. Likewise, item number 2 and item number 3 with sub-topic of the Limit of a Sequence and the Theorems of Limit of a Sequence, respectively, were categorized as difficult items. On the contrary, item number 4 with sub-topic of the Theorems of Limit of a Sequence was identified as an easy item. To make it clearer, Figure 1, Figure 2, and Figure 3 present the questions in the test and the sample of student's answers.

From the students' answers which are presented in Figure 1, Figure 2, and Figure 3, it can be foreseen that the student was incapable to solve the problems number 1, 2, and 3 systematically, because of the incapacity in understanding some theorems and definetions which are related to the problems. The students could not recognize and analyze the relation between the theorems and definitions.



① S himpunan tak kosong dan himpunan bagian dari $\mathbb{R}$, jika w batas bawah S maka w infimum dari S jika dan hanya jika untuk setiap $\varepsilon > 0$ terdapat $s' \in S$ sehingga $w + \varepsilon > s'$. Buktikan!

13

Penyelesaian

Dik: $S \neq \emptyset$, $S \subseteq \mathbb{R}$, w batas bawah S

$\Updownarrow$ P: w infimum S

S: $\forall \varepsilon > 0 \; \exists s' \in S \ni w + \varepsilon > s'$

Akan dibuktikan

$\Rightarrow$

misal w = infimum dari S dan ambil $\varepsilon > 0$ sembarang maka $w + \varepsilon$ bukan batas bawah S yaitu $w + \varepsilon > s'$ dengan s' lebih kecil dari $w + \varepsilon$

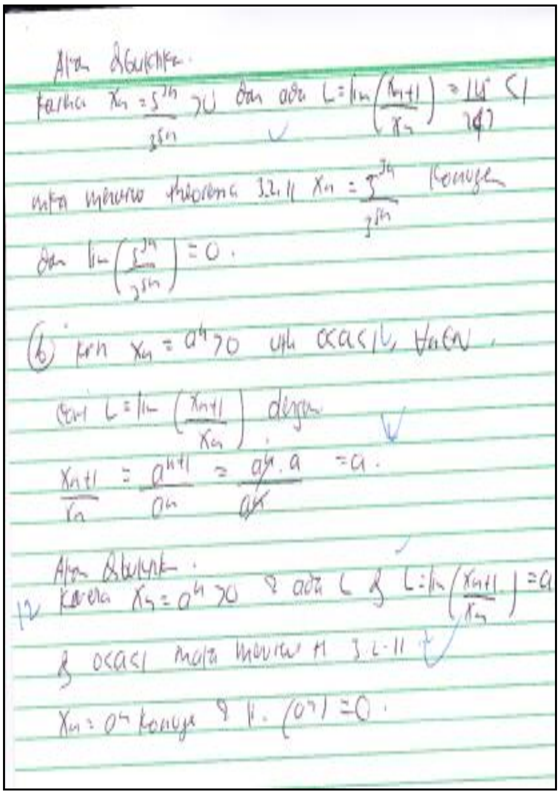
12

$\Leftarrow$

ambil sembarang $\varepsilon > 0$, w batas bawah S ..

1

Figure 1. Student's answer on Problem 1



Akan ditunjukkan.

Karena $x_n = \frac{5^{3n}}{3^{5n}} > 0$ dan ada $L = \lim\left(\frac{x_{n+1}}{x_n}\right) = \frac{125}{243} < 1$ ✓

maka menurut theorema 3.2.11 $x_n = \frac{5^{3n}}{3^{5n}}$ konvergen

dan $\lim\left(\frac{5^{3n}}{3^{5n}}\right) = 0$.

(b) Karena $x_n = a^n > 0$ untuk $0<a<1$, $\forall n$.

Cari $L = \lim\left(\frac{x_{n+1}}{x_n}\right)$ dengan ✓

$\frac{x_{n+1}}{x_n} = \frac{a^{n+1}}{a^n} = \frac{\cancel{a^n} . a}{\cancel{a^n}} = a$.

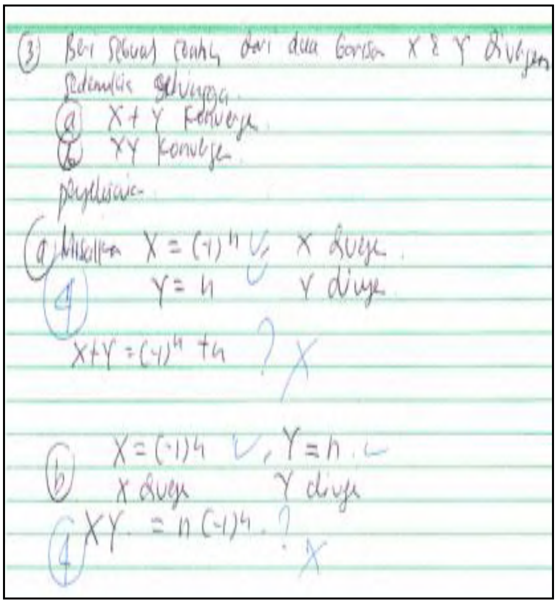
Akan ditunjukkan:

12 Karena $x_n = a^n > 0$ & ada L & $L = \lim\left(\frac{x_{n+1}}{x_n}\right) = a$

& $0<a<1$ maka menurut th 3.2.11 ✓

$x_n = a^n$ konvergen & $\lim(a^n) = 0$.

Figure 2. Student's answer on Problem 2



(3) Beri sebuah contoh dari dua barisan X & Y divergen sedemikian sehingga

(a) X + Y konvergen

(b) XY konvergen

penyelesaian

(a) Misalkan $X = (-1)^n$ ✓ X divergen

4 $Y = n$ ✓ Y divergen

$X+Y = (-1)^n + n$? ✗

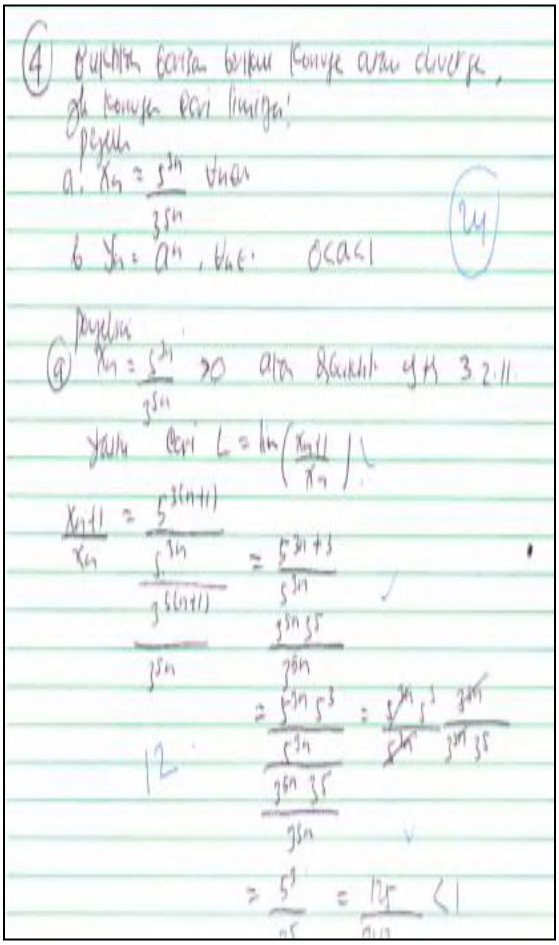
(b) $X = (-1)^n$ ✓, $Y = n$ ✓

X divergen Y divergen

4 $XY = n(-1)^n$? ✗

Figure 3. Student's answer on Problem 3

It is presented in Figure 4 that in the fourth problem, the student seemed to comprehend the topic. The theorems related to sequences and series were analyzed before the implementation for solving a problem. It can be seen from the sample in which the student could use the theorems systematically as suggested in solving the problem.



(4) Tunjukkan barisan berikut konvergen atau divergen, jika konvergen cari limitnya!

penyel

a. $x_n = \frac{5^{3n}}{3^{5n}}$ ∀n∈N

24

b. $x_n = a^n$, ∀n∈N $0<a<1$

Penyelesaian

(a) $x_n = \frac{5^{3n}}{3^{5n}} > 0$ akan ditunjukkan y th 3.2.11.

yaitu cari $L = \lim\left(\frac{x_{n+1}}{x_n}\right)$ ✓

$\frac{x_{n+1}}{x_n} = \frac{\frac{5^{3(n+1)}}{3^{5(n+1)}}}{\frac{5^{3n}}{3^{5n}}} = \frac{\frac{5^{3n+3}}{3^{5n}3^5}}{\frac{5^{3n}}{3^{5n}}}$ ✓

$= \frac{\frac{5^{3n}5^3}{3^{5n}3^5}}{\frac{5^{3n}}{3^{5n}}} = \frac{\cancel{5^{3n}}5^3}{\cancel{5^{3n}}} \frac{\cancel{3^{5n}}}{\cancel{3^{5n}}3^5}$

12

$= \frac{5^3}{3^5} = \frac{125}{243} < 1$

Figure 4. Student's answer on Problem 4

The result of the analysis showed that the ability of the test participants was quite diverse. In fact, merely a small number of students can solve questions number 1, 2, and 3 correctly. Most of the students could not determine specific theorems and definitions to solve the problems, especially in the second and third problems. In contrast, most of the students already understand the theorems used to solve the fourth problem, which are the sequences and series theorems, even though they faced a difficulty to analyze the theorems.

The estimation of the students' ability is presented in the interval scale (-3, +3). The category score in Rasch Model shows the number of the required steps to solve an item correctly. A high score indicates a good ability category. On the contrary, a low score indicates a low category of ability as well. The output of the estimation of ability parameter obtained from QUEST program and the package eRM with partial credit modeling or

Rasch Model is used to illustrate the comparison between the students' ability estimated using the Joint Maximum Likelihood (JML) approach with the package eRM and those estimated using the Conditional Maximum Likelihood (CML) approach with the QUEST program.

In JML approach, the students' ability could not be expressed in score 0 and score 100. Meanwhile, in CML approach, the students' ability can be expressed in score 0 (approximately a value of -3.09) and score 100 (as approximately a value of 85). Therefore, it can be inferred that Rasch Model using CML approach is more suitable than Rasch Model using JML approach to estimate the students' ability in understanding the subject-matter.

The result of analysis meets the OutfitMSQ criteria if the value is 0.035 < OutfitMSQ < 3.239. The analysis resulted a value of 0.5 < OutfitMSQ < 1.5, thus it fulfills the range of OutfitMSQ. The criteria of INFIT MNSQ is 0.5 < MNSQ <1.5. According to the mean value and the standard deviation of Rasch model, the CML approach with the package eRM is eligible since the mean and the standard deviation meets the criteria. On the contrary, the JML approach with Quest program is less appropriate as indicated by the mean and the standard deviation that do not meet the criteria.

In conclusion, the result of analysis on the estimation of students' ability reveals that the estimation of students' ability using Rasch model with CML approach and eRm program is more accurate than the estimation of students' ability using Rasch model with JML approach and QUEST program. Similarly, based on OutfitMSQ, Rasch model using CML approach with eRm program has better performance than Rasch model using JML approach with Quest program.

Conclusion

Based on the results and discussions, it can be concluded that the essay test items on the first Real Analysis course that have been tested to the students of Mathematics Education Department, Universitas Pancasakti Tegal can be classified as a good test. Besides, the students' ability can be estimated precisely by using Rasch Model with CML approach and eRm package. The estimation of participants' ability was quite diverse. A small number of students can solve questions number 1, 2, and 3 correctly despite these questions were classified difficult. Meanwhile, most of students already understand the theorems used to solve the fourth problem. The students are capable to apply the theorems systematically to solve the fourth problem.

References

Bartle, R. G., & Sherbert, D. R. (2000). *Introduction to real analysis*. New York, NY: John Wiley & Sons.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Buckley, K. E., Winkel, R. E., & Leary, M. R. (2004). Reactions to acceptance and rejection: Effects of level and sequence of relational evaluation. *Journal of Experimental Social Psychology*, *40*(1), 14–28. https://doi.org/10.1016/S0022-1031(03)00064-7

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the Rule-Space model and Attribute Hierarchy method. *Journal of Educational Measurement*, *44*(4), 325–340. https://doi.org/10.1111/j.1745-3984.2007.00042.x

Glas, C. A. W., & Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika*, *54*(4), 635–659. https://doi.org/10.1007/BF02296401

Griffin, P., & Nix, P. (1991). *Educational assessment and reporting: A new approach*. Sydney: Harcourt Jovanovich.

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

Imaroh, N., Susongko, P., & Isnani, I. (2017). Uji validitas tes ulangan akhir semester gasal mata pelajaran matematika (Studi deskriptif analisis dokumenter di SMP

Negeri Slawi tahun pelajaran 2016/2017). *JPMP (Jurnal Pendidikan MIPA Pancasakti)*, *1*(1), 80–89. https://doi.org/10.24905/jpmp.v1i1.792

Isgiyanto, A. (2011). Analisis data ujian nasional matematika berdasarkan penskoran model Rasch dan model Partial Credit. *Prosiding Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA*, 43–52. Retrieved from https://eprints.uny.ac.id/7172/1/PM-7 - Awal Isgiyanto.pdf

Keeves, J. P., & Masters, G. N. (1999). Introduction. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon-Elsevier Science.

Kurniawan, D. D., & Mardapi, D. (2015). Penyetaraan vertikal tes matematika SMP dengan teori respons butir model Rasch. *Jurnal Evaluasi Pendidikan*, *3*(1), 12–25. Retrieved from http://journal.student.uny.ac.id/ojs/index.php/jep/article/view/1221/1093

Lababa, D. (2008). Analisis butir soal dengan teori tes klasik: Sebuah pengantar. *Iqra'*, *5*, 29–37. Retrieved from https://jurnaliqro.files.wordpress.com/2008/08/03-jun-29-36.pdf

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

McMillan, J. H. (2005). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34–43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x

Naga, D. S. (2003). *Teori pengukuran*. Retrieved from http://dali.staff.gunadarma.ac.id/Downloads/folder/0.1

Ningsih, L. D., & Isnani, I. (2010). Studi komparatif tingkat reliabilitas tes prestasi hasil belajar matematika pada tes bentuk uraian dengan model penskoran GPCM (Generalized Partial Credit Model) dan Penskoran GRM (Graded Response Model). *Cakrawala: Jurnal Pendidikan*, *4*(8). https://doi.org/10.24905/cakrawala.v4i8.176

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

Schwartz, S. L. (2005). *Teaching young children mathematics*. London: Praeger.

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

Susongko, P. (2014). *Pengantar metodologi penelitian pendidikan*. Tegal: Universitas Pancasakti Tegal.

Tyler, R. (1950). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Wright, B., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 1–24). Maple Grove, MN: JAM Press.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement.

REiD (Research and Evaluation in Education), 5(2), 2019, 103-119
Available online at: http://journal.uny.ac.id/index.php/reid

An authentic assessment model to assess kindergarten students' character

***1Umi Faizah; 2Darmiyati Zuchdi; 3Yasir Alsamiri**
1Department of Islamic Early Childhood Education, Sekolah Tinggi Pendidikan Islam Bina Insan Mulia Yogyakarta
Jl. Jembatan Merah No. 116K, Prayan, Depok, Sleman, Yogyakarta 55283, Indonesia
2Department of Social Sciences Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
3College of Education, University of Hail
P.O. Box 2440, Hail, 81481, Kingdom of Saudi Arabia
*Corresponding Author. E-mail: umifaizah74@gmail.com

Submitted: 25 April 2019 | Revised: 19 August 2019 | Accepted: 3 September 2019

Abstract

The aim of character development is essential to know, especially for kindergarten pupils. It might help teachers in developing students and offering helps and services. Thus, an assessment model is needed to identify children's character development easily and accurately. This study aims to (1) develop an assessment model for evaluating kindergarten pupils' character which is considered valid, reliable, and fulfill the criteria of goodness of fit statistic, and (2) know the characteristics of an authentic assessment model instrument to assess the achievements of early childhood characters in kindergarten. This research used Plomp's Research and Development Model. Data were collected using a questionnaire, documentation, interview, observation, and Focus Group Discussion. The validation was proven by the expert judgment with the Aiken's V formula and the reliability was estimated with Cronbach's Alpha. The validation construct and reliability were examined using Exploratory Factor Analysis, followed by Confirmatory Factor Analysis to ensure the result. Furthermore, the results of this research show that (1) the assessment model developed is ASOKA. This model is considered valid and reliable as it meets the criteria of goodness and fits the statistic; (2) the characteristics of an authentic assessment model instrument to assess the achievements of early childhood characters in kindergarten are: (a) with the index of Aiken's V analysis results of 0.901, the content validity is considered high; (b) the instrument construct validity has fulfilled the criteria of goodness of fit statistic; (c) as seen from Alpha Cronbach value coefficient 0.914, its instrument reliability is considered good enough.

Keywords: *ASOKA model, authentic assessment, pre-school student*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.24588

Introduction

Education in Indonesia is still encountering an unfavorable situation, particularly when it is being related to the low quality of the educational process and outcomes as well as the nation's weak characters (Abidin, 2012). Many issues reflect weak characters, one of which is the brawl among students which is unexpectedly worrying. Another worse example is the *nglithih* (viciously hurting someone by using sharp things) which may cause death. Those various conflicts arising today are not only triggered by the economic crisis, but also by the moral crisis. Given these circumstances, the educational institution is the first institution to be questioned. One of the reasons that can be put forward is that educational institutions are the most effective means of strengthening the character of the nation. Besides, the character is also a benchmark of educational success. Based on the historical research of all countries in the world, education has two main purposes: to guide young people to be intelligent and to have virtuous behavior (Lickona, 1991, p. 5). The educational purposes are not limited solely to intelligence. Character is another important purpose of education. The existing education is supposed to be able to create highly intelligent students with the best character.

In addition, education is again the most essential aspect of the quality improvement of Indonesian people. Referring to its basic role, education is a path of human quality improvement that emphasizes the formation of basic quality, such as faith, piety, personality, intelligence, and so forth (Naim, 2012, p. 25). On the other hand, education has a very strategic value in improving the quality of the nation. Thus, breakthroughs to develop the character of the nation through character education programs need to be reformulated.

As might have known, character education is not a new thing in the national education system, since the objectives of the national education, as stated in all laws, substantively contain character education, despite the different formulas. The characters which are developed within the system are not just a horizontal relationship between individuals with other individuals, yet it also deals with the vertical relationships between individuals and Allah *Swt.* In this case, faith becomes the core of a man while he is controlled by his belief/faith (Majid & Andayani, 2011, p. 65). Thus, in Islamic educational institutions, faith becomes the target of education. Similarly, in the formulated character education references, religious values become the main target that must be developed for each learner. Conceptual character education manuscripts have been designed in such a way to produce the character of human beings. It is reflected in the 2005-2025 National Long-Term Development Plan which places character education as the first mission to realize the vision of national development. Hence, character building is meant to make particular groups realize the well-mannered society with noble characters.

Concerning the effort to realize character education as mandated in the National Long-Term Development Plan, the character building has been stipulated in the function and objectives of national education, as in the Law of Republic of Indonesia No. 20 of 2003 on National Education System. Thus, the National Long-Term Development Plan and the Law on National Education System are solid foundations for implementing the operation of cultural education and nation character as the priority program of National Education Ministry 2010-2014, as outlined in the National Character Education Action Plan.

Based on the data obtained on implementation of the character education in Kindergarten or *Raudhatul Athfal*, character education had not been implemented optimally. It is in line with the findings of Zuchdi (2006, pp. 92–93) which concluded that the context of school education has not fully supported the implementation of character education, especially the achievement in implementing character education at the kindergarten level, and only four skills were managed to be developed, namely greeting, being friendly, helping others, and asking for help politely. In achieving the mission of developing people's characters, each related institution, starting from families, educational institutions, and communities should take their roles based on their own capacities. Concerning character development in kindergartens as the initial formal education, schools and teachers attempt to integrate values in the character education through lessons and school environment, and there should be assessment systems that can monitor the character development of early learners in kindergartens.

The learners' cognitive development has been the focus of almost all teaching strategies. In this context, it is not easy to develop the affective and psychomotor aspects of learning (Suyadi, 2013, p. 189). The affective learning strategy will be useful to improve learners' attitudes during learning activities (Hamruni, 2009, p. 20). It is developed based on the behavioral psychology with stimulus-response (s-r) concept to form new behavior (attitude). Thus, the affective strategy aims to develop values in character education. In other words, the affective aspect will strongly influence the learners' feelings or positive emotions, so teachers will view the learning as 'a process to become' rather than 'the results' (Suyadi, 2013, p. 190). Thus, an assessment model that can be used to track the character building in a learning process is needed. To serve the purpose, the authentic assessment is the best assessment that can be used.

The authentic assessment is based on the real-life context and requires multiple approaches to solve a problem. This assessment involves performance measurement which reflects the learners' competencies as observed in their learning, achievement, motivation, and attitude (O'Malley & Pierce, 1996, p. 4). The competencies to achieve are attitude, knowledge, and skill, based on the recommended national standards (Law of Republic of Indonesia No. 20 of 2003 section 35 verse 1). In other words, this type of assessment monitors and measures learners' competencies in multiple problem-solving situations in a real-life context. It is expected that this authentic assessment model can be a solution to provide an excellent assessing system and the instruments can be easily used by teachers and accurately measure learners' performance in kindergartens (*Taman Kanak-Kanak* or *TK*) and *Raudlatul Athfal* (*RA*).

During the preliminary study, a survey and interview were conducted to 21 participants from *TK* and *RA* principals and teachers from June to September 2014 in several kindergartens in Sleman regency, Yogyakarta. The result shows that 70% of teachers had not implemented a proper assessment of learners' learning, 40% of teachers revealed that they regarded the assessment only as burdensome administrative duties, and 80% of teachers and *TK* and *RA* did not have proper assessment instruments. The assessment in character education should be carried out integratedly and continuously, that is by observation, task completion, conversation, and task submission to provide a conclusion for the achievement of indicators for character values. The preliminary study also revealed that the majority (65%) of *TK/RA* teachers determined the level of children's character achievement based on the teachers' personal perception, knowledge, and interpretation, and not based on daily recorded data. It then resulted in the teachers' unawareness in learning the achievement level of children's character development confidently; particularly on whether it has developed optimally or not.

Besides, it can be seen that although character values have been integrated into the daily activity plans, teachers remain depicting hesitance in determining the decision on the achievement level of children's character development because the indicators employed in the children's character development assessment tend to be too broad and not specified in details. Thus, it is necessary to have an assessment model that can be applied specifically to identify children's character development easily and accurately, so that teachers can provide assistance and service to students to develop their potentials and characters according to each child's developmental needs. With appropriate assessment options, teachers will be able to detect each child's developmental achievements appropriately.

Based on those descriptions, the model developed in this study is an authentic assessment model to assess the character achievement level, including: (1) the theoretical formation of *TK/RA* children's character dimensions in the construct of the assessment model; (2) the development of an authentic assessment model instruments from the construct of an authentic character assessment model.

Authentic Assessment in Early Childhood Learning in *TK/RA*

Early age children, including those in *TK* and *RA*, are in the period of growth and development. Early childhood comprises of various activities of motions, games, and habituations. Thus, the assessment of early age children is done by observing their growth and developmental stages, then comparing them with the indicators. Previous studies (Edgington, 2004, p. 149; Suyanto, 2005, p. 194) state that an assessment of early age and kindergarten children is a process of observing, recording, and documenting the children's performance and work (Jamaris, 2004, p. 119), skills attitudes, and performance.

An early childhood assessment aims not necessarily at measuring the achievement and achieving scholastic success, but rather observing the level of the developmental progress and abilities that children have made in their various actions, attitudes, performance, and appearance. Thus, assessing early age children's characters in kindergartens does not serve as a purpose to compare the children, but to see and comprehend the development of one child and the other. Jamaris (2004, p. 134) suggests that an observation must focus on the child's behaviors which are then compared with their age. The assessment should be sustainable and holistic, authentic, individual, natural, multi-sources, and multi-context (Suyanto, 2005, pp. 195–196). The use of authentic assessment to measure character can be done by observing children's performances and comparing them with the children's developmental level during the observation. To gain accurate data on children's development, the observation may be done in a school setting during in- or out-class activities.

Assessment of early childhood is conducted in an authentic manner with real, functional, and natural activities (Suyanto, 2005, p. 196). It is done to get an overview of the real development of children's abilities, by presenting valid and comprehensive data through record-keeping of children's creativity in detail about their strengths and weaknesses, as well as significant events in their lives (Edgington, 2004, p. 147; Jamaris, 2004, p. 119).

An authentic assessment in learning is a process or formal effort to collect information on the important variables of learning as evaluation materials and decision making by teachers to improve the process and students' learning outcomes (Herman, Aschbacher, & Winters, 1992, p. 95; Popham, 1995, p. 5). In this description, it is understood that an authentic assessment involves learners in purposeful and meaningful authentic assessment.

The term 'authentic assessment' was first introduced by Wiggins in 1988 in the journal *Phi Delta Kappa* entitled 'Authentic Assessment' (Zaenul, 2001, p. 4). The assessment is also known as an alternative assessment, in contrast with its more widely known traditional counterpart of the traditional assessment in the form of a paper and pencil test. Due to this case, the idea of the alternative assessment raises more serious attention and becomes the turning point of the widespread discussion of authentic assessment.

An authentic assessment is considered as an effort to integrate the learning achievement measurement with the overall learning process. It must be noted that the assessment itself is a part of the learning process as a whole. Therefore, *TK/RA* teachers should learn about the purpose of the authentic assessment and be able to apply it in the learning process to make it more effective.

Sometimes, the term 'authentic assessment' is interchangeable with other terms. That is an alternative assessment referring to the process of assessing students' behavior performance on a multidimensional basis in real situations. In other words, alternative assessment can be defined by using non-traditional approaches to measure students' performances and learning outcomes.

An authentic assessment typically involves a task for learners to display and assessment criteria or a rubric to assess the task performance. Arends (1997, p. 284) defines an authentic assessment as a process to assess students' performance in carrying out certain tasks in real situations.

From those statements, it is concluded that the assessment to assess young learners' characters in *TK/RA* is categorized as an authentic and classroom-based assessment. A classroom-based assessment will be able to reveal the learners' real conditions in the classroom (Stiggins, 1991, p. 8). This authentic and classroom-based model is appropriate for assessing the child's character achievement consisting of 14 characters. An authentic assessment is a comprehensive assessment process (encompassing all aspects of learning), continuous and inseparable from the learning process, aiming to determine the progress and achievement of students and to improve the planning, process, and learning achievement.

Character Values Developed in *TK/RA*

The character values developed in TK/RA are based on the opinions of experts (Bar-On, 2000, 2005; Gardner, 1996; Thorndike, Hagen, & Sattler, 1986). Further, through the FGD consisting of two experts in character education and five other experts, three aspects were selected, namely spiritual, personal, and social aspects. In the development of character values which had been validated by expert judgment, the three aspects were developed into 14 characters values: faithful, worshipping ritual, humane, honest, patient and modest, brave and confident, disciplined, creative, independent, caring/empathy, tolerant, cooperative, and polite and humble. Each of these characters is explained as follows.

Faithful

Being faithful is the first character instilled in every Muslim child. It can be seen from the way Muslims welcoming their newborn by reciting *adzan* in the baby's right ear and *iqomah* in the left ear. It is evidence that the first and foremost value developed by Muslims is believing in God (Allah) (Marzuki, 2015, p. 32). Being faithful in this context is linked with *rukun iman* (the pillars of faith).

Worshipping Ritual (Hablun minallah)

Worshipping ritual is a part of *sharia* (Islamic law). The Prophet Muhammad *Saw.* taught that after *tauhid* (believing that Allah is the One and Only God) comes *sharia* in the form of worship and *muammalah* (humanity) (Marzuki, 2015, p. 45). Worshipping can be defined as rules regulating the direct relationship (ritual) between human beings and Allah (Ash-Shiddieqy, 2009). In other words, the character-building through worshipping rituals is simplified by implementing six pillars of faith and five pillars of Islam.

In this study, the discussion of worship and *muamalah* are separated into the character values that must be instilled in students of *TK/RA*. Based on Islamic teaching, everything conducted by Muslims can all be counted as worship, when it is intended as a form of devotion to Allah *Swt.* Thus, the term worshipping ritual is used. Worshipping ritual is associated with the implementation of the Five Pillars (*rukun Islam).*

Humanity (Hablum minannas)

The Islamic character is divided into two parts: the character of *Khalik* (Allah *Swt.*) and the character of beings (other than Allah) (Marzuki, 2015, p. 32). The character of beings can be broken down into several types, one of which is the character of fellow humans (*hablun minannas*).

Muamalah means treatment or action towards others, the relationship of interests (Munawwir, 1997), and *muamalah* means action between humans and other than humans. The actual activity is difficult to distinguish from the character of the social aspect, but in this study, performing *muamalah* (*hablun minannas*) is doing an activity which have something to do with other people is related to behavior in Islamic teachings, namely, those contained in the Qur'an and/ or *al-hadits.*

Honesty

Honesty literally means straight heart, not lying, not cheating. Honesty is an important value that must be owned by everyone. Honesty is not only shown verbally, but it is reflected in everyday behavior (Naim, 2012, p. 132). Honest character must be instilled from an early age by using various approaches and setting exemplary behaviors.

Patience and Modesty

Being patient means being able to refrain from anger. Patience is a positive character that must be instilled in children early on. It is the ability to control oneself. Armed with patience, a child will be able to resist the inner impulse and think before acting. It will guide the child to do the right things and less likely to take actions ending in bad results (Naim, 2012, p. 56). By instilling patience, it is expected that children will be able to wait patiently without easily getting upset, and they are willing to wait their turn orderly (queuing).

Modesty is a way of life that is not excessive. Modesty is the inner attitude of a person who fully believes that God enlarges the sustenance of his servants, so he becomes a servant of God who is impervious and satisfied with what has been earned so far (Munawar-Rachman, 2015, p. 373).

Bravery and Confidence

Being brave and confident are important characters that should be instilled early on. Being courageous is often associated with self-confidence because it is believed that courage grows from a positive self-image. A child with a positive self-image will be courageous to try tackling difficult things or challenges. The child has the confidence that he is able and, therefore, he is willing to try. Confidence is an attitude of believing in one's ability (self-ability). Confidence removes worry in conducting one's action. Confidence fosters a sense of freedom to do as he desires and at the same time fosters a sense of responsibility for his actions. It also fosters a sense of achievement and fosters the ability to recognize the advantages and disadvantages of oneself. Louster (2002, p. 4) describes a person with self-esteem as a selfless person who does not need the encouragement of others, always being optimistic, and happy. According to Schiller and Bryant (2012, pp. 76–77), self-confidence is someone's ability to weigh choices and make the decision to choose one choice freely and consciously.

Self-Discipline

Self-discipline is shown when students respect and do a system requiring people to obey provisions, orders, and regulations in pursuance. In other words, self-discipline is an attitude of obeying established regulations and provisions without any intention/expectation of rewards (Naim, 2012, pp. 142–143).

Self-discipline is an intended influence to help children deal with their environment. Self-discipline is developed from the need to maintain the stabilization between individuals' tendencies and intentions to achieve their goals in regard to environmental expectations (Semiawan, 2008, pp. 27–28). Self-discipline stated in this study is children's willingness to correctly adhere to the rules.

Creativity

Creativity is one of the most essential character values. By possessing creative character, a child may experience a dynamic life. His mind is constantly developed and he constantly conducts activities to explore valuable things (Naim, 2012, p. 152). Creativity is an attitude and action reflected in innovation in many aspects of problem-solving. Thus, a creative person can always find better new ways with diverse results compared to the previous ones (Suyadi, 2013, p. 8). Creativity becomes an essential character for early-aged children.

Self-Independence

Self-independence is the ability to be independent without relying on other people (Marzuki, 2015, p. 98). It is an essential character that needs to be instilled from an early age. By instilling a sense of independence, it is hoped that children will be initiated in doing things they should be able to do by themselves. As a result, they will be skillful in life.

Caring/Empathy

Caring is an attitude that shows tendencies on problems, situations, and conditions occuring in the children's surroundings by being involved in them. Being emphatic is reflected in the way children intend to be treated. It is in line with Schopenhauer (1997, p. 190) who states that caring is based on a principle: 'Treat others the way you want to be treated'. Children who care are those who move to do something to inspire, change, and do good deeds to surroundings. Caring usually comes from loving. To develop a sense of caring, as in other moral values, learning involving approaches of developing three characters aspects i.e. knowing, feeling, and acting needs to be done (Lickona, 1991, p. 312).

Tolerance

Tolerance is one of the essential characters to be instilled from an early age. By instilling tolerance at an early age, children are expected to accept diversity and believe that God creates a variety of humans with various perfections and lacks. Tolerance is a permissive attitude toward disagreement or the ability to consider different opinions, attitudes, or ways of life (Naim, 2012, p. 138). The development of tolerance is due to willingness and awareness to respect differences.

Responsibility

Responsibility is the main essential character which need to be instilled at an early age. The concept of responsibility in this study is the effort performed when completing tasks. The tasks should be done at best, and anybody performs the effort should take any possible risk. Moreover, he should be able to solve any problem. In the literal meaning, it is the ability to respond to something. A people-oriented attitude is implied in the definition, and it shows the attitude of actively responding to other's needs (Lickona, 2004, p. 44).

Responsibility is defined by maximizing one's capabilities in an attempt of doing something (Munawar-Rachman, 2015, p. 345). With the responsibility character, it is expected that children can maximize their effort in doing their tasks.

Cooperation

Cooperation stated in this research includes mutual cooperation and active participation in a work assigned in groups. The sense of cooperation is shown in how children do group tasks well, help others who have not finished their tasks, and ask others to play around together. Cooperation is developed by the principle of mutual respect and affection. Cooperations also mean helping each other in kindness and devotion as suggested by Islamic provisions: *ta'awun 'ala al birr wa al taqwa* (Munawar-Rachman, 2015, p. 259).

Politeness and Humbleness

Politeness and humbleness are in one of the main characters that should be instilled in early childhood. The concepts of politeness and humbleness stated in this paper include the character of prioritizing values in behaving and respecting others in the ways children speak and act (Miskawaih, 1994, p. 81).

Humbleness teaches children to take others' knowledge, strength, and other qualities into account. Humbleness is politeness and not arrogant (Munawar-Rachman, 2015, p. 230). Human glory is measured by the quality of their humbleness in life.

Method

This research and development employ a set of stages suggested by Plomp (1997, p. 5) covering five stages: (1) the preliminary investigational phase, (2) the design phase, (3) the realization/construction phase, (4) the test, evaluation, and revision phase, and (5) the implementation phase. There are two main stages in this research: the research/pre-research stage and the development stage. The pre-research activities include investigation, design, and construction, while the development stage covers the activities of testing, evaluating, and revising. The R&D model by Plomp includes five stages: preliminary investigation, assessment model planning, and designing, assessment instrument development, assessing, assessment of evaluation items, item assessment, and assessment implementation. The development procedure of ASOKA (Authentic Character Assessment) can be seen in Figure 1.

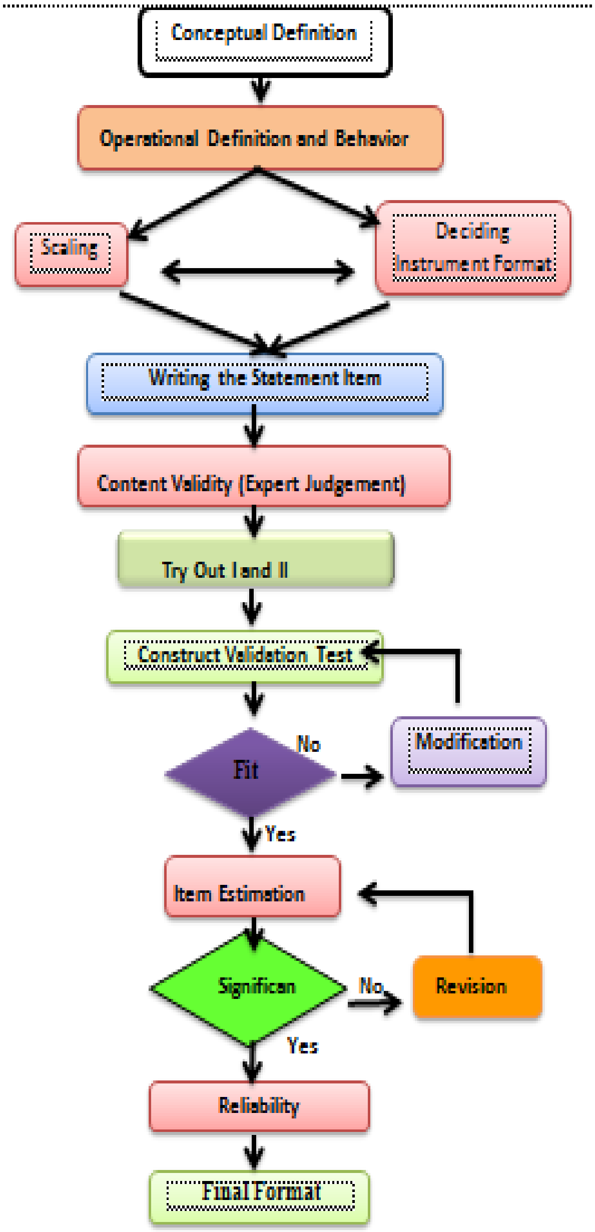

Conceptual Definition
Operational Definition and Behavior
Scaling
Deciding Instrument Format
Writing the Statement Item
Content Validity (Expert Judgement)
Try Out I and II
Construct Validation Test
Fit
No
Modification
Yes
Item Estimation
Significan
No
Revision
Yes
Reliability
Final Format

Figure 1. Development procedure of ASOKA measure

Findings and Discussion

Findings

The Theoretical Validation Result of ASOKA Model

In the evaluation phase of the first model of ASOKA, to prove the validity, the content of the instrument was analyzed by the experts which consist of assessment experts, character education experts, *PAUD/TK/RA* experts, and Islamic Education experts to gather the expert judgment. The experts on children's development psychology were not included in this phase since the researchers consider the assumption that they are represented by the experts on *PAUD/TK/RA*. The readability test was administered to teachers of *TK/RA* and some *TK/RA* principals.

The validation test was done to know whether the assessment measure could measure the early age children's character development, especially those in *TK/RA* level. The explanation of the product evaluations from *expert judgment* is described as follows.

Content Validity in the ASOKA Model from Experts

The first product validation of the ASOKA model was carried out through expert judgment using the Delphi technique. Validation is intended to make sure that the developed ASOKA model instruments can be used to detect the attainment level of TK/RA children. This Delphi technique was chosen with considerations easier to do, more in-depth input, and focused on the problem under study.

Content validation analysis from experts/experts is done by using the formula of Aiken's V. The results of the analysis show that the ASOKA instrument has a good representation related to the accuracy of the indicators of the aspects and accuracy of the items on the indicator. For the criteria for the accuracy of indicators on the aspects assessed, the Aiken validity index identifies that there is 1 (one) indicator that has a lower index than the other indicators (<0.76), while the accuracy of the items on the indicator, the Aiken validity index identifies six items, namely items 1, 2, 3, 4, 10 and 22, have lower indices than other items (<0.76). The six items were then annulled and replaced with new items after a long discussion with SMEs. Furthermore, the new items were reassessed by the SMEs and the content validity index value (V) ≥ 0.76 was obtained, so it can be concluded that all the items in the ASOKA instrument which amounted to 65 items met the content validity.

The results of the discussion and input from experts, as well as the final index obtained, were then consulted with the promoter and co-promoter. Some changes after consultation with experts through the Delphi method are as follows. (1) Changes to the spiritual aspect, initially consisting of three indicators, namely faith, worship, and honesty, then changed places and new characters emerged, namely, the changes were perfected into faith, worship ritual (*hablun minallah*), and performing *muamalah* (*hablun minannas*), the addition of indicators on the spiritual aspect in the form of *muamalah*, actually will be a little confusing with the social aspects, but in *muamalah,* these special indicators are made that are different from the social aspects. (2) Changes in indicators consist of adding, combining, adding a lot given to the character of the faith by including six pillars of faith as a whole, the addition of indicators is also on creative characters, who previously could only use three indicators, into five indicators. (3) Amendment to the number of item statements, previously 56 items, after summarizing all entries from the seven experts to 65 items. (4) Changes to the choice of terms of achievement of child characters, previously with the term Not Developing (*Belum Berkembang* or BB), Starting to Grow (*Mulai Berkembang* or MB), Developing Expectations (*Berkembang Sesuai Harapan* or BSH), and Developing Very Good (*Berkembang Sangat Baik* or BSB), changing to Unappeared (*Belum Muncul* or BM), Appearing with Stimulation (*Muncul dengan Stimulus* or MS), Emerging Not Consistent (*Muncul Belum Konsisten* or MBK), and Emerging Consistently (*Muncul Konsisten* or MK).

The full explanation of the three aspects of the character developed is as follows: (1) spiritual aspects, including the character of faith, ritual worship (*hablun minallah*) and *muamalah* (*hablun minannas*). (2) Personal aspects, including honest, patient and simplicity, brave and confident, disciplined, creative, and independent. (3) Social aspects, including care/patience, tolerance, responsibility, cooperation, courtesy, and humility.

Readability Test Results by Practitioners

The readability test of the ASOKA instrument aims to ensure (1) the clarity of clues, the scope of ASOKA components, the language used, and the writing procedure and appearance of the assessment in general from the instrument, as a whole are understood by prospective users, (2) the clarity of aspects of character, (3) the clarity of character indicators, (4) the formulation of communicative statements, (5) the use of easily understood sentences and words, (6) the clarity of assessment rubrics, and (7) the written procedures related to letter form, font size, format or instrument layout.

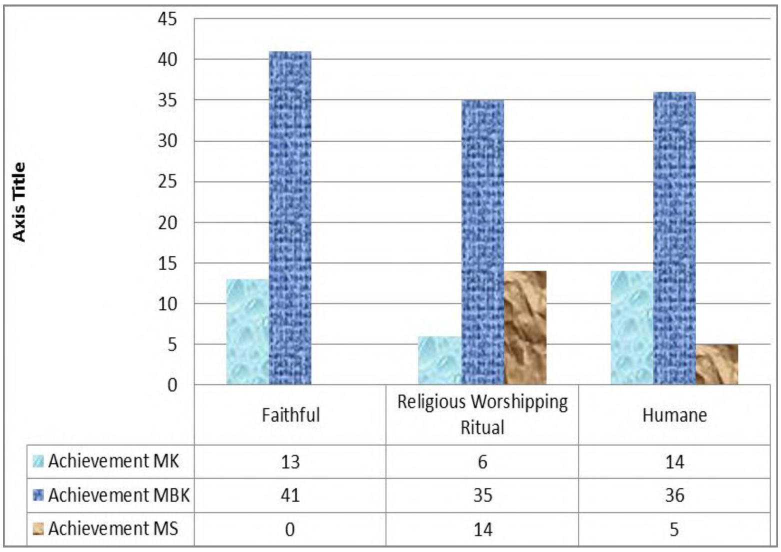
This readability test activity involves practitioners from the elements associated with prospective users, namely teachers, heads of TK/RA, assessment of readability using a modified Likert scale with four choices, namely a minimum score of 1 (cannot be used), score 2 (can be used with little improvement), score 3 (can be used without improvement, and score 4 (ideal used). When consulted with the guidelines the feasibility categorization of the model is included in either classification or it gives an indication that the level of readability of this developed instrument can be classified as good or feasible to use.

Results of ASOKA Model Try-Out I

Try-Out I involves 106 learners of group B from six TK/RA in Yogyakarta, including: RA DWP UIN Sunan Kalijaga, RA Nurul Dzikri, TK Islam Tunas Melati, TK Aisyiyah Bustanul Athfal Taruna Alquran, RA Masyitoh Ngeposari Gunung Kidul, TKIP Salsabila Pandowoharjo-Sleman. The sample of the limited test is determined by proportional random sampling. The results of the first try-out of the ASOKA model are explained as follows.

Spiritual Aspects

The achievement of the spiritual aspect of TK/RA is represented by three TK/RA with different characteristics, namely RA Nurul Dzikri (RA-ND), TK Islam Tunas Melati (TK-TM), TK Plus (full day) Salsabila (TKP-SB). The distribution of assessment results on the achievement of children's character in the three institutions RA-ND, TK-TM, and TKP-SB as a representation of TK/RA in Islamic education institutions can be seen in the histogram in Figure 2.



	Faithful	Religious Worshipping Ritual	Humane
Achievement MK	13	6	14
Achievement MBK	41	35	36
Achievement MS	0	14	5

Figure 2. Histogram of distribution of assessment of spiritual character achievement for each component

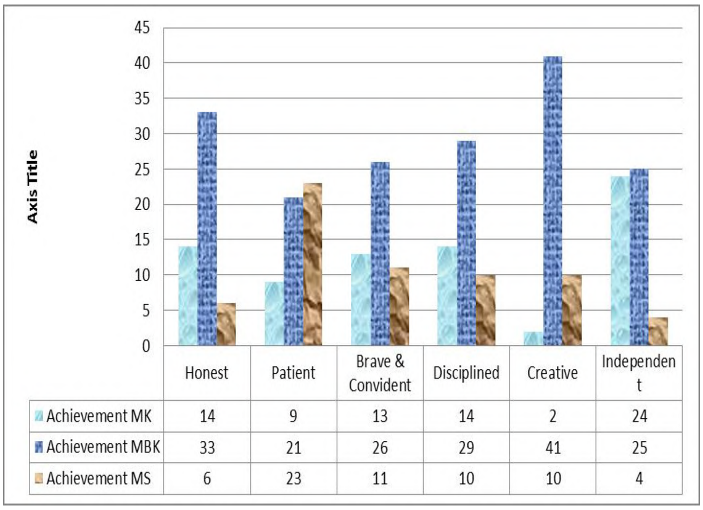
In Figure 2, it can be seen clearly that the average achievement of the children's character in the spiritual aspect is dominated by the achievement of MBK (appearing inconsistent), which is 65% -75%, whereas the attainment of the MK (consistent emergence) /culture is only about 10%. It proves that the spirituality of the child is still in the developing stage and its appearance has not been consistent, for this reason, it needs to be continually improved with various strategies to instill it into the main character possessed by each student.

Personal Aspects

The assessment results for the achievement of the children's character on the personal aspects can be seen in the distribution in Figure 3. In Figure 3, it can be seen clearly that the average achievement of the children's character in this personal aspect is dominated by the achievement of MBK (*Muncul Belum Konsisten* or Emerging Inconsistently), which is 75%, while the lowest in the MK (*Muncul Konsisten* or Consistent Appearance)/creative character is only around 4%, while the highest achievement is independent character, which reaches 45%. It is a challenge for teachers to grow the creative character of each student through a variety of planned stimulation in learning.

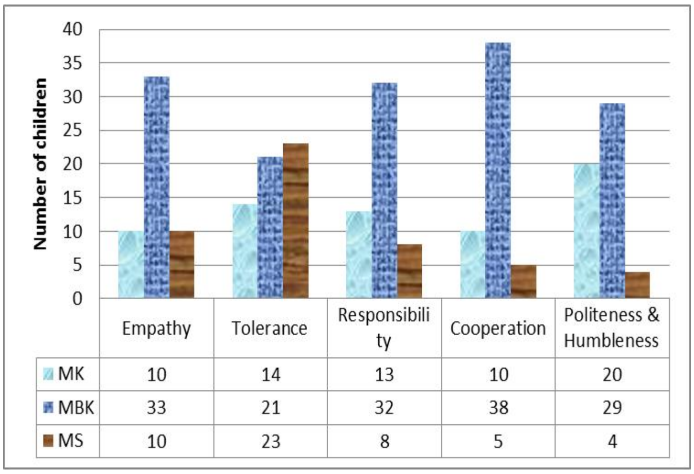
Social Aspects

The assessment results for the achievement of the children's character on the social aspects can be seen in the distribution in Figure 4.



	Honest	Patient	Brave & Convident	Disciplined	Creative	Independen t
▨ Achievement MK	14	9	13	14	2	24
▨ Achievement MBK	33	21	26	29	41	25
▨ Achievement MS	6	23	11	10	10	4

Figure 3. Histogram of personal character aspect achievement for each component



	Empathy	Tolerance	Responsibili ty	Cooperation	Politeness & Humbleness
▨ MK	10	14	13	10	20
▨ MBK	33	21	32	38	29
■ MS	10	23	8	5	4

Figure 4. Histogram of distribution of assessment results achievement of child characters on social aspects

Results of Analysis of Validity and Reliability of ASOKA Instruments

Contract validity testing is a test of validity related to the level of the scale that reflects and acts as a concept being measured (Hair, Black, Babin, & Anderson, 2010, p. 710). Analysis of construct validity on character dimensions was carried out using Explanatory Factor Analysis (EFA) analysis. This analysis serves as a pointer to factors that can explain the correlation between variables.

Each variable has a value of loading factor that represents it. The value of loading factors in EFA can be determined based on the number of samples in the study (Hair et al., 2010, p. 117). The adequacy of the number of observations of data can be identified through the Kaiser-Meyer-Olkin (KMO) parameter with a KMO value of > 0.5.

Correlations between multivariate variables can be identified by Bartlett's Test of Sphericity parameter which must have significance with p-value <0.05. The magnitude of the correlation between multivariate variables can be seen from the value of Measure of Sampling Adequacy (MSA) with the value of MSA > 0.5. The value of communal items has acceptable limits which are above 0.30 (Mooi & Sarstedt, 2011, p. 212). The results of the correlation test between variables are presented in the output of KMO and Bartlett's Test in Table 1.

Table 1. The KMO and Barlett's test instrument of ASOKA on Try-out I

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.868
Bartlett's Test Sphericity	Approx. Chi-Square	1525.141
		91
		.000

In Table 1, KMO MSA (Kaiser-Meyer-Olkin Measure of Sampling Adequacy) has a value of 0.868. The MSA KMO value is good since it is greater than 0.5 (KMO> 0.50). It indicates that all character dimensions have met the adequacy requirements of the number of observations (data). Based on the *Bartlett's Test of Sphericity*, it is obtained a Chi-Square value of 1525.141 at the degrees of freedom of 91 with the significance of less than 0.001 (<0.001). Based on the Anti-Image correlation (AIC), the item having MSA value which is less than 0.50 (<0.50) is not found, as shown in Table 2.

Table 2. Values of AIC

Spiritual Aspect		**Personal Aspect**		**Social Aspect**	
S1	0. 774	P4	0. 830	Sos11	0. 863
S2	0. 742	P5	0. 845	Sos12	0. 915
S3	0. 867	P6	0. 921	Sos13	0. 923
		P7	0. 889	Sos14	0. 916
		P8	0. 895		
		P9	0. 877		
		P1	0. 800		

Table 3. Total Variance Explained Value

Component	**Initial Eigenvalues**			**Rotation Sums of Squared Loading**		
	Total	**% of Var.**	**Cum. %**	**Total**	**% of Var.**	**Cum. %**
1	6.510	46.541	46.541	4.172	29.700	29.799
2	2.421	17.202	63.834	3.634	25.954	55.753
3	1.165	8.322	72.156	2.296	16.403	72.156
4	.695	4.962	77.188			
5	.571	4.079	81.197			
6	.463	3.3.4	84.502			
7	.437	3.122	87.623			
8	.377	2.692	90.315			
9	.287	2.047	92.362			
19	.273	1.950	94.312			
11	.256	1.830	96.143			
12	.216	1.557	97.699			
13	.177	1.267	98.966			
14	.145	1.034	100.000			

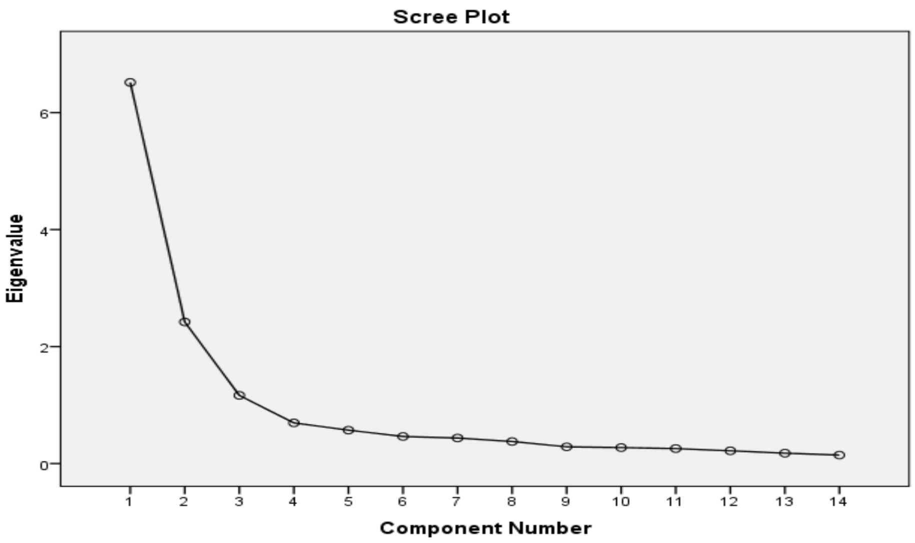


Figure 4. Scree Plot of the ASOKA instrument in the Tryout I

Based on Table 2, no items have MSA value below 0.50 (<0.50), so that in the next process, all ASOKA instrument items are included. Furthermore, to determine the number of possible factors formed can be seen in the Total Variance Explained table. The Total Variance Explained Value is summarized in Table 3.

In Table 3, Total Variance Explained values can be seen, and the variance that can be explained by factor 1, factor 2, and factor 3. The total of these three factors will be able to explain the variable at 72.156%. Thus, because eigenvalues are set 1, then the total value taken is > 1, namely components 1, 2, and 3. In the initial eigenvalues column of the cumulative sub column, it can be seen that the reduction of 14 items analyzed, obtained characteristic values (eigenvalue) as many as three factors. Of the three factors obtained KMO MSA value of 0.868 (> 0.07), it means fulfilling the requirements to continue. Eigenvalues with values above 1 (> 1) have three factors. It shows that there are three factors in achieving the character of early childhood in TK/RA according to the estimated indicators. Thus, it can be said that the ASOKA model instrument is said to be valid in terms of the validity of the construct.

The percentage of the loading factor variance that can explain the variance of the early childhood character achievement in *TK/RA* is the first loading factor of 46.541%, the second loading factor of 17.292%, and the third loading factor of 8.322%. Cumulatively, the three factors comprise of 72.156%. Besides, the scree plot which explains the total variance is illustrated in Figure 4.

Figure 4 shows the tendency of Eigen (eigenvalue) decrease used to determine subjectively the number of factors formed. From Figure 4, it can be seen that the scree plot shows the tendency of the Eigen (eigenvalue) decrease indicates that the formed factor leads to three characters dimensions.

Overall Model Fit

The goodness of the fittest of the measurement model with field data was done using the second-order Confirmatory Factor Analysis (CFA) technique. The second-order CFA analysis aims to determine the validity of indicators developed by the researchers. The existing indicators are said to be valid if the result of the loading factor value is higher than 0.3. The construct validity with the second-order CFA technique is used to test the fitness of the characteristic achievement assessment (Ghozali & Fuad, 2008, p. 137; Joreskog & Sorbom, 1999, p. 115). This approach means that the analysis is done directly in two stages, i.e. from the variable to the indicator, then from the indicator to the item. In addition, the second-order CFA also tests whether the data fit with the model that was formed previously or not.

Based on the standardized second order CFA test output, the statistical values used as the criteria in the good of fit fitness statistics are as follows: df = 69; *p*-value of 0.06624; Chi-square (*p*-value) > 0.05; RMSEA value of 0.05 (RMSEA < 0.08), CFI value of 0.99 (CFI ≥ 0.9); NFI value = 0.9; CFI value = 0.99; and RMR value = 0.018. The results of the character measurement model are shown in Table 4.

Table 4 shows the good of the fit test results of the model. Judged from the loading factor value, the indicators are all above 0.3. It indicates that all indicators constructing the authentic assessment components of early childhood characters in *TK/RA* are valid. The character constructs measurement model has met the goodness of fit statistics so that the character construct measurement model is stated as a good measurement model.

The ASOKA instrument test is done directly, and the SPSS display result shows KMO MSA > 0.05. Thus, it can be explained that all ASOKA dimensions have fulfilled the requirements of a sufficient amount of observation (data). In addition, Barlett's Test of Sphericity shows the significance of a *p*-value less than 0.05 (*p*-value < 0.05), indicating a significant correlation between observed variables of all dimensions. It can be concluded that the data of observations of character dimensions of *TK/RA* students have been qualified for the confirmatory factor analysis.

The overall model goodness of fit evaluation of each character dimension based on the second-order CFA (2nd CFA) test shows the fitness of the model with the data. The main criterion of the model matches with the field data if at least three requirements of seven commonly used measures are fulfilled, namely (1) Chi-square (*p*-value), (2) Goodness of Fit Index (GFI), and (3) Root Mean Square Error of Approximation (RMSEA). The model is said to be fit if Chi-square has a significance level (*p*-value) ≥ 0.005; Goodness of Fit Index (GFI) ≥ 0.90; and Root Mean Square Error of Approximation (RMSEA) is 0.05 < RMSEA ≤ 0.08 (Browne & Cudeck, 1993). Thus, it can be interpreted that the indicators specified to measure each dimension (personal, personal, social) together measure things accordingly. It is further seen that all sizes of GOF (goodness of fit) show a good model matching with the field data, so it can be concluded that overall, the ASOKA model on all character dimensions is fit.

An instrument for assessing students' character had been developed and proved to be valid and reliable, and then the feasibility of the model was tested by trying it out based on the model usage guideline. The ASOKA model usage guideline is in the form of a handbook completed with the assessment instruments and procedures as well as guidelines on writing the assessment reports. This handbook was tested to 15 teachers/potential users from 15 *TK/RA* in Yogyakarta.

Discussion

In developing this character assessment model, the researchers adopt Plomp's (1997, p. 5) approach that is modified into four phases: (1) the initial investigation stage, (2) design, (3) prototype construction, and (4) development. This development process produced a construction of the ASOKA model consisting of three aspects, namely: the spiritual aspect, the personal aspect, and the social aspect. The spiritual aspect consists of three characters: (1) faithfulness, (2) worshipping ritual (*hablum minallah*), and (3) humanity

Table 4. Goodness of fit test results of ASOKA model on Tryout I

No	GOF Effect Size	Fitness Level Target	Estimated Value	Fitness Level
1.	Chi-square(X^2)	< 2df	87.44	good fit
2.	*p*-value	> 0.05	0.06	good fit
3.	RMSEA	≤ 0.05	0.05	good fit
4.	NFI	≥ 0.90	0.96	good fit
5.	NNFI	≥ 0.90	0.99	good fit
6.	CFI	≥ 0.92	0.99	good fit
7.	RMR	< 0.05	0.018	good fit

(*hablum minannas*). The personal aspect consists of six characters, namely: (1) honesty, (2) patience and modesty, (3) bravery and confidence, (4) self-discipline, (5) creativity, and (6) self-independence. The social aspect consists of five characters: (1) caring/empathy, (2) tolerance, (3) responsibility, (4) cooperation, and (5) politeness and humbleness. These characters are the results of the development of the assessment model in its early stages to assess the character development of young learners in *TK/RA*.

The method used in the ASOKA assessment model was an authentic assessment employing observations. It is in accordance with Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett (2001, p. 75) who assert that the appropriate method for assessing kindergarten children is through performance inherent in the curriculum or often referred to as an authentic assessment. The assessment was done gradually and continuously so that the progress towards children's character development can be measured. In line with this fact, Suyanto (2005, p. 189) has suggested that an assessment is done through real, functional, and natural activities starting from the time the students get to school until they go home. The observation method employed in this assessment model has also been appropriate, as Azwar (2015, p. 90) states the assessment of attitudes (characters) can be done through behavioral observations. When a child shows repeated or consistent behaviors, it can be said that they already have those behaviors as their characters. A child's behaviors that appear repeatedly (consistently) show that the child has characters. For example, there is a child who always prays before doing any activities such as eating, drinking, and even playing; then it can be said that the child has a spiritual character. It is said so because when the child always prays before doing anything, he/she shows his/her belief in the existence of God/Allah.

In addition, validity was proven and the reliability of the ASOKA model instruments was estimated. The content validity of the instrument was obtained from expert judgment through the Delphi method and continued with analysis using the Aiken's formula.

Based on the analysis, the overall results of the instruments' indicators and items had Aikens index from 0.714 to 1.000, meaning that the proposed indicators and items were valid. The criterion which was used to determine the validity level was that of Retnawati (2016): Aikens's agreement index of 0.4-0.8 shows medium validity; an index of more than 0.8 shows high validity. In conclusion, all of the proposed indicators can be used to develop an authentic assessment instrument to assess the characters of young learners in *TK/RA*. The Aiken index is chosen because it is considered accurate to measure the content validity of an instrument. The instrument is valid if it measures what it should measure based on the raters' agreement.

The construct validity is the result of testing relating to the scale level that reflects and acts like the concept being measured (Hair et al., 2010, p. 710). The analysis of construct validity to character dimensions was performed by using Exploratory Factor Analysis (EFA). This analysis resulted in KMO value of 0.868 (KMO> 0.50); Chi-Square of 1525.141 with 91 degrees of freedom and significance less than 0.001 (<0.001); Anti Image Correlation (AIC) values of more than 0.50 (AIC > 0.50); communality value of 0.611-0.791 (communality > 0.03) which already fulfills the prerequisite criteria (Mooi & Sarstedt, 2011, p. 212).

Meanwhile, the instrument reliability was calculated by using Cronbach's Alpha approach. The Cronbach's Alpha coefficient for the reliability of the ASOKA model was 0.914. This value is higher than 0.70 (>0.70). This requirement refers to instrument reliability criteria (Mardapi, 2017; Nunnally, 1981, p. 115) which state that an instrument is considered reliable if the alpha reliability is 0.70 or higher. ASOKA model instrument's validity and reliability have been estimated and the results show that this instrument is valid and reliable. The assessment instrument product obtained is then used to assess the achievement of early childhood character development in *TK/RA*. This assessment is in the form of a checklist containing character assessment indicators including 14 characters described in 65 assessment items.

Further product development in this study is a user manual of the ASOKA model. It serves as a guide for the users, that is, kindergarten teachers, in applying the ASOKA instruments. The results of the assessment are used as the basis for the feasibility of using the ASOKA model in *TK/RA*. Assessment results from *TK/RA* teachers show that 72% assume that this model is good, meaning that it can be used without any revision. Fourteen percent thought that the manual is good enough, meaning it could be used with little improvement, while the other 14% would judge that the manual is excellent, ideal to be used as an example for character assessment.

Some weaknesses in ASOKA's user manual have been improved. It is done to improve the quality of the function of the ASOKA model as a valid and reliable assessment model that is able to measure the achievement of early childhood character in *TK/RA*. The use of this ASOKA model is started from the making of the RPPH (Daily Learning Program Plan) and ended by assessing the expected characters. Thus, all the points of behavior that become indicators of the achievement of the child's character can be appraised properly.

In assessing the achievement of the children's characters, it is necessary to diversify the system and the method of the assessment because the use of a varied method will better guarantee the quality of the assessment result. Thus, the construct of the ASOKA model developed by the researchers is one of the important assessment models used to help facilitate *TK/RA* teachers in performing their duty to do the assessment, which is inseparable from their two other tasks namely planning and conducting the learning process effectively.

After passing the experiment with a wide sample, the results of the ASOKA model development fulfill the validity and reliability criteria and it is also considered as the correct measurement model because it fulfills the criteria of goodness-of-fit model. Thus it can be stated that the developed instrument has a feasibility standard as an instrument to detect the level of achievement of early childhood character in *TK/RA*.

Conclusion

Based on the findings and discussion, the conclusion of this study can be formulated as follows. (1) ASOKA is an authentic assessment model developed to assess the achievement of early childhood character in TK/RA. This model consists of an instrument and manual of the assessment model to assess the character of early childhood in TK/RA. The ASOKA model is effective because all the indicators used to measure the spiritual, personal, and social aspects of early childhood character constructs in TK/RA mostly have a loading factor value greater than 0.30, while the reliability of the character constructs is proved from the value of the construct reliability coefficient (CR)> 0.70, i.e. CR = 0.72 on the spiritual aspect; CR = 0.79 on the personal aspect; CR = 0.86 on the social aspect. The ASOKA model also meets the criteria of the goodness-of-fit statistic, so the ASOKA model is considered as an assessment model that can be used to detect the achievement of early childhood character in TK/RA. (2) The characteristics of the authentic assessment model instruments to assess early childhood character outcomes in TK/RA are as follows: (a) the content validity of the ASOKA (Authentic Character Assessment) instrument is high. Based on the results of Formula Aiken's V analysis, the overall results of the indicators have an Aiken index of 0.714 to 1.000, the average index is 0.901; (b) The validity of the ASOKA model instrument construct (Authentic Assessment for Character) using the second-order CFA approach was obtained by a fit model to assess the character of early childhood in TK/RA. It means that the developed model of ASOKA meets the criteria of goodness-of-fit statistics; (c) the reliability of the developed instrument is high with the Cronbach Alpha value coefficient 0.914.

References

Abidin, M. Z. (2012). Tingkat pendidikan di Indonesia. *Seminar Pendidikan Karakter*. Bali: Universitas Udayana.

Arends, R. I. (1997). *Classroom instruction and management*. New York, NY: McGraw-Hill.

Ash-Shiddieqy, T. M. H. (2009). *Sejarah dan pengantar ilmu Hadits*. Semarang: Pustaka Rizki Putra.

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

Bar-On, Reuven. (2000). Emotional and social intelligence: Insights from the emotional quotient inventory. In R. Bar-On & J. D. A. Parker (Eds.), *The handbook of emotional intelligence: Theory, development, assessment, and application at home, school, and in the workplace* (pp. 363–388). San Francisco, CA: Jossey-Bass.

Bar-On, Reuven. (2005). The impact of emotional intelligence on subjective well-being. *Perspectives in Education*, *23*(2), 1–22. Retrieved from https://hdl.handle.net/10520/EJC87316

Browne, M. W., & Cudeck, R. (1993). Alternative way of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Model*. New York, NY: SAGE Publication.

Edgington, M. (2004). *The foundation stage teacher in action: Teaching 3, 4, and 5 years olds*. London: PCP Press.

Gardner, H. (1996). *Intelligence: Multiple perspectives*. Fort Worth, TX: Harcourt Brace College.

Ghozali, I., & Fuad, F. (2008). *Structural: equation modeling*. Semarang: UNDIP Press.

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

Hamruni, H. (2009). *Strategi dan model-model pembelajaran aktif menyenangkan*. Yogyakarta: Fakultas Tarbiyah UIN Sunan Kalijaga.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Jamaris, M. (2004). Assesmen pendidikan anak usia dini. *Seminar Dan Lokakarya Nasional Pendidikan Anak Usia Dini*. Jakarta.

Joreskog, K. G., & Sorbom, D. (1999). *Lisrel 8: User's reference guide*. Chicago, IL: Scientific Software International.

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

Lickona, T. (1991). *Educating for character: How our schools can teach respect and responsibility*. New York, NY: State University of New York.

Lickona, T. (2004). *Character matters: How to help our children develop good judgment, integrity, and other essential virtues*. New York, NY: Simon & Schuster.

Louster, P. (2002). *Tes kepribadian* (C. G. Sumeksto, trans.). Yogyakarta: Kanisius.

Majid, A., & Andayani, D. (2011). *Pendidikan karakter perspektif Islam*. Bandung: Remaja Rosdakarya.

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

Marzuki, M. (2015). *Pendidikan karakter Islam*. Jakarta: Amzah.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal*, *38*(1), 73–95. https://doi.org/10.3102/00028312038001073

Miskawaih, A. A. A. I. (1994). *Menuju kesempurnaan akhlak* (H. Hidayat & E. Hasan, trans.). Bandung: Mizan.

Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research*. Berlin: Springer-Verlag Berlin Heidelberg.

Munawar-Rachman, B. (Ed.). (2015). *Pendidikan karakter: Pendidikan menghidupkan nilai untuk pesantren, madrasah dan sekolah*. Jakarta: LSAF dan ALIVE Indonesia.

Munawwir, A. W. (1997). *Kamus Arab-Indonesia* (14th ed.). Surabaya: Pustaka Progressif.

Naim, N. (2012). *Character building: Optimalisasi peran pendidikan dalam pengembangan ilmu dan pembentukan karakter bangsa*. Yogyakarta: Ar-Ruzz Media.

Nunnally, J. C. (1981). *Psychometric theory*. New York, NY: Mc-Graw Hill.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learning: Practical approaches for teachers*. New York, NY: Addison-Wesley.

Plomp, T. (1997). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research*. Enschede: Faculty of Educational Science and Technology, University of Twente.

Popham, W. J. (1995). *Classroom assessment: What teachers need to know*. Boston, MA: Allyn and Bacon.

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

Schiller, P., & Bryant, T. (2012). *16 moral dasar bagi anak* (S. Sensusi, trans.). Jakarta: PT Elex Media Komputindo.

Schopenhauer, A. (1997). Menembus selubung sang maya. In F. M. Suseno (Ed.), *13 Model pendekatan etika*. Yogyakarta: Kanisius.

Semiawan, C. R. (Ed.). (2008). *Penerapan pembelajaran pada anak*. Jakarta: Indeks.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, *10*(1), 7–12. https://doi.org/10.1111/j.1745-3992.1991.tb00171.x

Suyadi, S. (2013). *Strategi pembelajaran pendidikan karakter*. Bandung: Remaja Rosdakarya.

Suyanto, S. (2005). *Konsep dasar pendidikan anak usia dini*. Jakarta: Departemen Pendidikan Nasional.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet intelligence scale* (4th ed.). Chicago, IL: Riverside.

Zaenul, A. (2001). *Alternative assessment*. Jakarta: Direktorat Jenderal Pendidikan Tinggi, Departemen Pendidikan Nasional.

Zuchdi, D. (2006). Pendidikan karakter melalui pengembangan keterampilan hidup (life skills development) dalam kurikulum persekolahan. In *Laporan penelitian hibah pascasarjana*. Yogyakarta.

REiD (Research and Evaluation in Education), 5(2), 2019, 120-129
Available online at: http://journal.uny.ac.id/index.php/reid

Psychometric characteristic of positive affect scale within the academic setting

***1Kartika Nur Fathiyah; 2Asmadi Alsa; 3Diana Setiyawati**
[1]Faculty of Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
[2,3]Faculty of Psychology, Universitas Gadjah Mada
Jl. Sosio Humaniora Bulaksumur, Karangmalang, Sepok, Sleman, Yogyakarta 55281, Indonesia
*Corresponding Author. E-mail: kartika.fip_uny@yahoo.co.id

Submitted: 9 July 2019 | *Revised: 5 September 2019* | *Accepted: 12 September 2019*

Abstract

This analysis study is one of several stages that must be passed before testing the structural model. This study is initiated due to the limited information related to the measurement of the Positive Affect Scale within the academic settings. The research method used in this study was a quantitative method. It was done in among 724 students of state junior high schools in Sleman, Yogyakarta. The instrument development consisted of guideline arrangement, language feasibility testing, content validation through expert judgments, trials to measure the item discrimination index, item selection based on the item discrimination results, items representation for each indicator, and the construct validity test for the selected items. The testing of the measurement model used the data analysis techniques of Structural Equation Models (SEM) with the assistance of the AMOS program version of 21. The results of the study show that the validity analysis of the Positive Affect Scale within the academic setting was able to produce items that can reveal constructs or latent concepts appropriately.

Keywords: *positive affect scale, validity analysis, academic setting*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.25992

Introduction

Affect plays a significant role in people's life (Nath & Pradhan, 2012) consisting of positive and negative form. Affect usually refers to one's emotion that is recognized and described as pleasantness or unpleasantness (Watson, Clark, & Tellegen, 1988). The negative form provides short-term benefits to facilitate the tendency of specific behaviors in the form of responses, while the positive affect brings long-term benefits (Fredrickson, 1998). The negative form includes tension, hopelessness, fear, and irritation, while the positive form covers spirit, strength, activeness, desire, and stamina (Yik, Russell, & Steiger, 2011).

The positive affect reflects the expansion of high energy, vigor and alert that make an individuals excited, full of concentration, and pleasant feeling. On the other hand, the low positive affect creates sadness and fatigue (Watson et al., 1988). The positive affect means a person's tendency to have a variety of positive emotional experiences (Watson et al., 1988). Related to the trait (the tendency of an individual state to be relatively stable), the positive affect is associated with the more frequent and intense episodes experienced by individuals. Based on the state (the person's condition at a certain time), the positive affect is a beneficial emotional experienced at a particular time (Watson & Tellegen, 1985).

The positive affect is a key component of assessment and effective coping towards stressful situations (Folkman, 2008), and an antidote to negative emotions that can reduce its harmful influence (Fredrickson, Tugade, Waugh, & Larkin, 2003). It develops a mental readiness to grow and step out from unpleasant situations and escalates sources of psychological coping to face stressors (Fredrickson, Mancuso, Branigan, & Tugade, 2000). This affect can also maintain physical and psychological health (Danner, Snowdon, & Friesen, 2001), and build personal resources and well-being (Fredrickson & Joiner, 2002).

Based on Neurobiological perspective, the positive affect occurs due to the release of large amounts of dopamine from temporary phasic to synaptic clefts. The dopamine is then multiplied through the midbrain of the dopaminergic system to the striatum, limbic area, and prefrontal cortex (Ashby, Isen, & Turken, 1999). Several studies have found that positive effects improve performance based on front striatal dopaminergic interactions among healthy individuals (Demanet, Liefooghe, & Verbruggen, 2011).

Studies in various settings have revealed the role of positive affect in improving individual outcomes (Samios, Abel, & Rodzik, 2013; Lyubomirsky, King, & Diener, 2005); Steptoe, Dockray, & Wardle, 2009). In the academic field, the role of positive affect is considered as very meaningful (Schutz & Lanehart, 2002; Goetz, Pekrun, Hall, & Haag, 2006) because it affects teaching and learning (Schutz & Lanehart, 2002), student's subjective well-being, process quality, learning a-chievement, teacher interaction with students, and learning process effectiveness (Goetz et al., 2006). Those roles indicate the importance of positive affect within the academic setting.

In fact, the availability of information on the positive affect in the academic setting is very limited (Linnenbrink-Garcia & Pekrun, 2011) and tends to have little attention from researchers (Pekrun et al., 2010). Therefore, Linnenbrink (2006) and also Seligman, Ernst, Gillham, Reivich, and Linkins (2009) suggest that psychological studies within an academic context should be gained more, especially development of positive affect scale within academic setting, as instrument to measure and support the optimal school functioning.

This research aims at elevating the study on the role of positive affect within the academic setting by developing a proper instrument. Many researches in various settings on the positive affect including those in the academic setting have been using the concept of Watson et al. (1988) considered less specific. As a result, the positive affect cannot be optimally explored based on its context.

In the academic settings, Pekrun (1992) identifies the positive affect or emotions from motivation, learning process, and student performance. He classifies positive affect in the academic setting into positive affect related to assignments and social. Regarding the task, the positive affect comes from (a) the process, as a pleasure when undergoing the academic process; (b) anticipatory joy, as a positive affect that arises before the academic process takes place with happy feeling imagining the results to be achieved and the expectation towards the academic activities; and (c) prospective, a positive affect after the academic process takes place shown by a joy feeling because of the achieved success (joy of success), and satisfaction, pride and relief after undergoing the academic process. Meanwhile, the social concerns on the positive affect that appears because of social interactions during the academic process. The indicators are gratitude, empathy, admiration, and sympathy or love.

The specific measurement model discusses the relation between latent variables (constructs) and measurement indicators, by conducting an instrument construct validity analysis to reveal how well the measurement indicators measure the latent (construct) concept. Construct validity test includes exploratory and confirmatory factor analysis.

Exploratory factor analysis (EFA) is for situations where the relation between observed and latent variables is not known so it requires exploration to determine how and how closely the observed variables relate to the underlying (latent) factors. Conversely, in the confirmatory factor analysis (CFA), the factor structure is assumed to be known (Dachlan, 2014). Because the indicators of this research have been theorized by Pekrun (1992), the analysis was done using CFA. Thus, this paper aims to confirm whether the scale of positive affect within academic setting built already matched between the data obtained with the underlying theory.

Method

The research method used in this study was quantitative method. This study was conducted among the junior high school students in Sleman, Yogyakarta. The subject involved in this study were 724 students, including 359 students in the field trial stage and 365 for the empirical data collection. The data collection at each stage was done to different subjects. The analysis in the field trial used discrimination test, while the validation analysis in this study was the analysis of the empirical data collection in addition to the model testing.

The instrument development consisted of guideline arrangement, language feasibility test, content validation through calculated by Aiken's V formula, discrimination index, item selection based on the item discrimination results, items representation for each indicator, the construct validity test in the selected items and validity and reliability test. The Aiken's formula is described as follows (Aiken, 1985).

$$V = \sum s/[n(c-1)]$$

Notes:
1 = the lowest of validity assessment score (equal to 1)
c = the highest of validity assessment score (equal to 4)
r = the score from the assessor
n = number of assessors = r-1

The testing of positive affect scale in the academic setting employed Structural Equation Models (SEM) with the assistance of AMOS program version 21. To determine the Goodness of Fit Index (GFI) according to Dachlan (2014), it used several criteria: Chi-Square and p-values, CMIN/DF, GFI, AGFI (Adjusted Goodness of Fit Index), CFI (Comparative Fit Index), TLI (Tucker-Lewis Index), and RMSEA (Root Mean Square Error of Approximation).

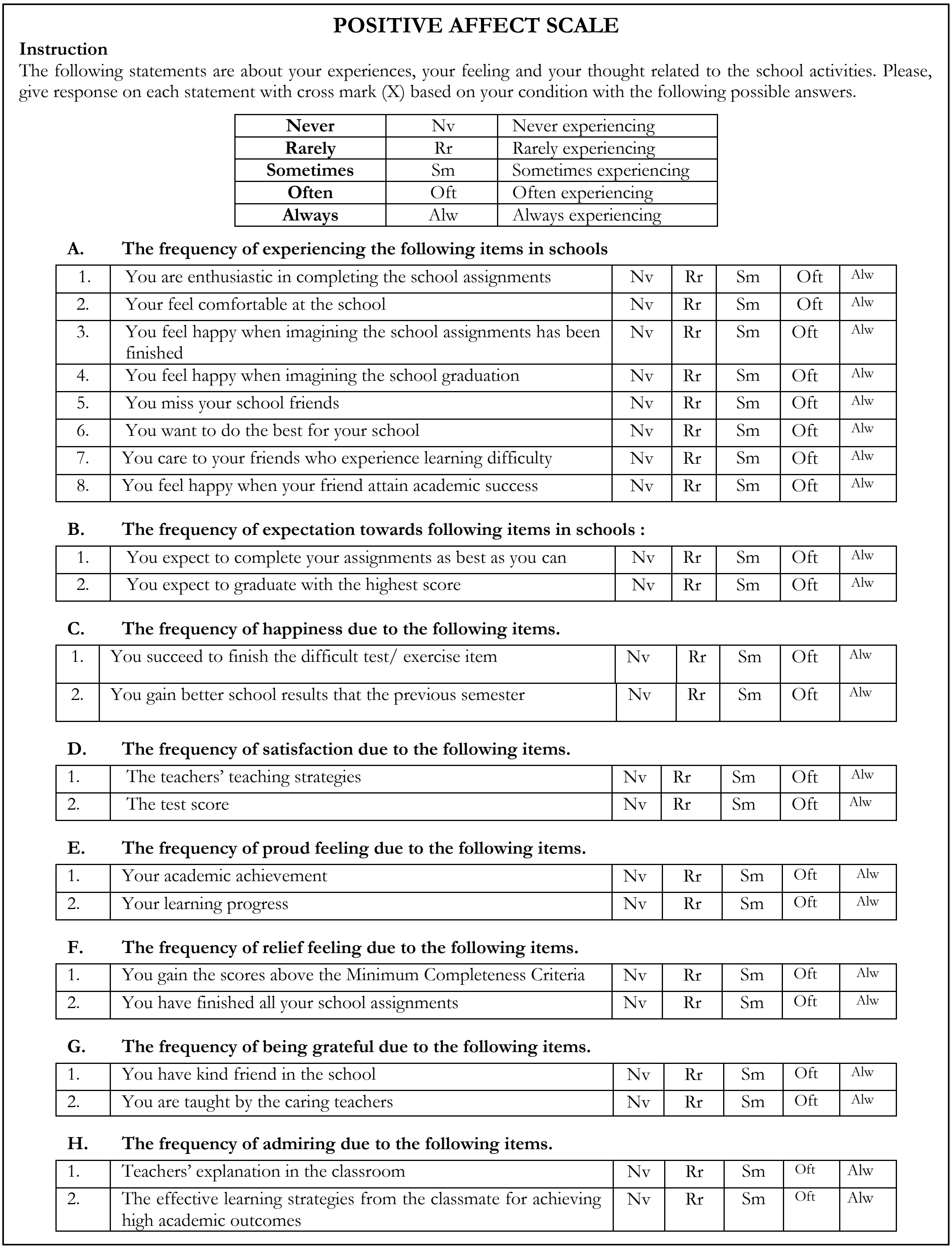
Findings and Discussion

Findings

The initial step of the study in carrying out the validity test of the positive affect scale is to make the guidelines for the instruments. This guideline was arranged referring Pekrun's (1992) theory regarding the general taxonomy of positive emotions relevant to motivation, learning process, and student performance. The scale contains two aspects: (1) task and (2) social aspects. The positive affect scale comprises statements related to school activties. The students were asked to respond each statement based on their experience, feeling, and thought. This scale contains statements that support (favorable) and those that do not support (unfavorable). There were two models of answer choices to respond to the statements. The first model includes the frequency/intensity of 'never', 'rarely', 'sometimes', 'often', and 'always' with the score range from 1 (never) to 5 (always) respectively, while the second model focuses on its appropriateness containing 'very inappropriate', 'inappropriate', 'sometimes', 'appropriate', and 'very appropriate' with the score range from 1 (very inappropriate) to 5 (very appropriate) respectively. The number on the positive affect scale of the trial stage were 30 items. The details of the dimensions, indicators, and number of items is shown in Table 1, while the scale is presented in Figure 1.

Table 1. The guideline of positive affect scale

Aspects	Indicator	Sub-Indicators	Number of Test Item
Tasks	Process	Joy	3
		Anticipatory joy	3
	Prospective	Hope	3
	Retrospective	Joy about success	3
		Satisfaction	3
		Pride	3
Social		Gratitude	3
		Empathy	3
		Admire	3
		Love	3
	Total number of test items		**30**



POSITIVE AFFECT SCALE

Instruction

The following statements are about your experiences, your feeling and your thought related to the school activities. Please, give response on each statement with cross mark (X) based on your condition with the following possible answers.

Never	Nv	Never experiencing
Rarely	Rr	Rarely experiencing
Sometimes	Sm	Sometimes experiencing
Often	Oft	Often experiencing
Always	Alw	Always experiencing

A. The frequency of experiencing the following items in schools

1.	You are enthusiastic in completing the school assignments	Nv	Rr	Sm	Oft	Alw
2.	Your feel comfortable at the school	Nv	Rr	Sm	Oft	Alw
3.	You feel happy when imagining the school assignments has been finished	Nv	Rr	Sm	Oft	Alw
4.	You feel happy when imagining the school graduation	Nv	Rr	Sm	Oft	Alw
5.	You miss your school friends	Nv	Rr	Sm	Oft	Alw
6.	You want to do the best for your school	Nv	Rr	Sm	Oft	Alw
7.	You care to your friends who experience learning difficulty	Nv	Rr	Sm	Oft	Alw
8.	You feel happy when your friend attain academic success	Nv	Rr	Sm	Oft	Alw

B. The frequency of expectation towards following items in schools :

1.	You expect to complete your assignments as best as you can	Nv	Rr	Sm	Oft	Alw
2.	You expect to graduate with the highest score	Nv	Rr	Sm	Oft	Alw

C. The frequency of happiness due to the following items.

1.	You succeed to finish the difficult test/ exercise item	Nv	Rr	Sm	Oft	Alw
2.	You gain better school results that the previous semester	Nv	Rr	Sm	Oft	Alw

D. The frequency of satisfaction due to the following items.

1.	The teachers' teaching strategies	Nv	Rr	Sm	Oft	Alw
2.	The test score	Nv	Rr	Sm	Oft	Alw

E. The frequency of proud feeling due to the following items.

1.	Your academic achievement	Nv	Rr	Sm	Oft	Alw
2.	Your learning progress	Nv	Rr	Sm	Oft	Alw

F. The frequency of relief feeling due to the following items.

1.	You gain the scores above the Minimum Completeness Criteria	Nv	Rr	Sm	Oft	Alw
2.	You have finished all your school assignments	Nv	Rr	Sm	Oft	Alw

G. The frequency of being grateful due to the following items.

1.	You have kind friend in the school	Nv	Rr	Sm	Oft	Alw
2.	You are taught by the caring teachers	Nv	Rr	Sm	Oft	Alw

H. The frequency of admiring due to the following items.

1.	Teachers' explanation in the classroom	Nv	Rr	Sm	Oft	Alw
2.	The effective learning strategies from the classmate for achieving high academic outcomes	Nv	Rr	Sm	Oft	Alw

Figure 1. The positive affect scale statements

After preparing the guidelines, the language feasibility was tested to ensure that the sentence in the scale was understandable by the reader and present the same meaning as the researchers' intention (Azwar, 2016). The respondents of the test were seven junior high school students from various levels (two students from the seventh grade, three from the eighth grade, and two from the ninth grade). They also came from various types of schools:

state, private and Islamic-based schools. Each respondent was asked to examine and provide an assessment on the extent to which the items presented on the scale to be understood.

After making sure with language feasibility, the content validation was following. It was the expert judgment from those who have the relevant scientific capacity to the issue measured, aimed at knowing whether the items were in line with the measured aspects. The assessment was focused on the appropriateness between the item indicators and the measured variables, and the writing procedures, and evaluation for high social desirability (Azwar, 2016). This expert judgment was then calculated using Aiken's V formula to obtain content validity coefficient based on the measured construct (Azwar, 2016).

The obtained scores from the Aiken's V formula calculation ranged from 0 to 1, the bigger number of coefficients indicates that the item shows better content validity (Azwar, 2016). The items in the instrument were assessed by 21 experts with educational background at least Master Degree in Psychology. The suitability level between the item and indicator ranged from 1 to 5 (five points): 1 is 'very inappropriate', 2 is 'inappropriate', 3 is 'moderate', 4 is 'appropriate', and 5 is 'very appropriate'. Based on the coefficient Table of Aiken by taking the value p=0.01 (1% margin of error) from 21 assessors, the score limit to be used so the items can be received was 0.71. The content validation with Aiken coefficient value moved from 0.88-0.96, the mean value of Aiken (V) was 0.91. Thus, the items are suitable with its indicators according to experts which means the positive affect scale is considered to have good content validity.

The next step after the expert judgment was the item discrimination test. This test was done to obtain items with high discrimination index to distinguish individuals or groups of individuals who have and do not have measured attributes. The approach employed total item consistency which showed the suitability between item functions and its scale functions (Azwar, 2016). The score for each item was correlated with the total score. The high correlation values indicated that the item had a high function towards the overall scale function. The items less than 0.30, according to Azwar (2016), can be interpreted to have a low discrimination index (invalid) so it can be deleted. Based on the item discrimination test using Pearson's total item correlation with the assistance of SPSS V.21, the discrimination index of high positive affect scale items moved from 0.330-0.652. Further, these items were selected to be tested in its construct validity through confirmatory factor analysis.

The item selection on positive affect scale was done by selecting two items having the highest discrimination index on each indicator and considering the item representation as the indicator. The selected items in positive affect scale for confirmatory tests can be seen in Table 2. After obtaining the selected item, the construct validity was done by CFA to test the validity of the scale's indicators as the measurement of latent construct. The construct validity provides the belief that the indicators taken from the sample really illustrate the actual scores in the population. Thus, this analysis confirms empirically based on the sample data to provide theoretical truths for latent variables.

Table 2. The selected items in the Positive Affect Scale for confirmatory tests

Aspect	Indicator	Sub Indicator	Old Version Number	Revised Version Number
Task	Process	Joy	A2, A3	A1, A2
	Prospective	Anticipa-tory joy	A4, A5	A3, A4
		Hope	B1, B2	B1, B2
	Retrospective	Joy about success	C1, C2	C1, C2
		Satisfaction	D1, D2	D1, D2
		Pride	E1, E2	E1, E2
		Relief	F1, F3	F1, F2
Social	Gratitude		G2, G3	G1, G2
	Empathy		A7, A8	A7, A8
	Admiration		H1, H3	H1, H2
	Sympathy		A11, A12	A5, A6

Table 3. The Criteria of Goodness-of-Fit (GoF)

Parameter	Critical Scores	Experts
Chi-Square	Closer to 0 is better	Arbuckle (2013), Kline (2011)
Chi-Square/df	< 2	Byrne (2001)
Probability	≥ 0.05	Kline (2011)
GFI	≥ 0.90	Kline (2011), Dachlan (2014), Ghozali (2017)
AGFI	≥ 0.90	Kline (2011), Dachlan (2014), Ghozali (2017),
CFI	≥ 0.90	Kline (2011), Dachan (2014), Ghozali (2017)
TLI	≥ 0.90	Arbuckle (2012), Dachlan (2014), Ghozali (2017)
RMSEA	≤ 0.05	Dachlan (2014)

```
                                 |--------------------
                    175.338      |*
                    191.820      |***
                    208.302      |*******
                    224.784      |************
                    241.266      |*******************
                    257.748      |******************
                    274.230      |*****************
N = 1000            290.712      |***************
Mean = 266.192      307.194      |*********
S. e. = 1.225       323.676      |*******
                    340.158      |****
                    356.639      |**
                    373.121      |*
                    389.603      |*
                    406.085      |*
                                 |--------------------
```

Figure 2. The results of bootstrapping data in Positive Affect Scale

The construct validity can be analyzed from the factor load value (squared multiple correlation) indicators of latent constructs (Ghozali, 2017). To measure the suitability of the model, it was used the measurement of GoF known as the values of CMIN, df, p, GoF, AGFI, TLI, and RMSEA. This GoF standard referred to the opinions of Kline (2011), Arbuckle (2013), and Ghozali (2017). The criteria of GoF can be seen in Table 3.

The factor load value towards the latent construct to maintain the item on positive affect scale was 0.40. It was based on the opinion of Hair, Black, Babin, Anderson, and Tatham (2010) who mention that the determination of the minimum limit of factor load with 200 subjects or more is 0.40. In line with this ides, Hair et al. (2010) and Netemeyer, Bearden, and Sharma (2003) state that the item should have the factor load of 0.40-0.90, while the value less than 0.40 should be disregarded.

The confirmatory analysis of positive affect scale used maximum likelihood (ML) estimation method. The requirement that must be fulfilled by ML method was multivariate normality (Byrne, 2010). The multivariate normality test in Positive Affect Scale showed c.r of 35,069. Because the value of c.r was beyond the range of -2.58 to +2.58, the data were declared abnormal, so it did not meet the assumption of multivariate normality. To overcome the non-normal data, the bootstrap procedure was applied. The visualization of the bootstrapping results on Positive Affect Scale data with the sample of 1000, the percentile confidence level of 95%, and the bias corrected confidence interval of 95% can be seen in Figure 2.

Figure 2 showed that Chi-square distribution value with 1000 bootstrap samples in Positive Affect Scale was 266.192; the cluster values in the multivariate center were normal with 266 because there were several values above and under 266 that were comparable. After fulfilling the normality requirements of the data, the confirmatory analysis was conducted. The preliminary results indicated that Positive Affect Scale measurement model was not in accordance with the model criteria (GoF), as presented in Figure 3.

In Figure 3, it can be seen that Positive Affect Scale did not meet the measurement model of GoF criteria. This was indicated by Chi-square results=1286.932 (relatively high), chi-square/df=209, p=0.00 (critical score p≥ 0.05), GoF, AGFI, TLI, and CFI which was still far below 0.9 (critical value ≥0.9) and RMSEA = 0.119 (critical value ≤0.05).

To achieve GoF criteria to positive affect scale, the items that can be used were those with loading factor of 0.5. Thus, the D2 items were deleted since they did not meet the criteria (loading factor=0.49). The next item selection was by paying attention to the modification suggestions by AMOS program, such as removing items Affect C2, F2, E1, A6, G2, E1, A4, E2, A8, H2, A5, D1, B2, and G1 since it had variance with some other items (cross-loading) with relatively high MI values.

Based on the modifications made, positive affect scale can reach the measurement fit value as shown in Figure 4, i.e. Chi-square of 15.602 with p=0.76; Chi-square/df=1.734; GFI=0.986; AGFI=0.967; TLI=0.978; CFI= 0.987; and RMSEA=0.045 according to the established GoF criteria. In detail, a summary of the analysis of Positive Affect Scale factors based on GoF criteria can be seen in Table 4.

Six items are selected in Table 4: affect A1, A3, B1 (related to the assignment aspects) and affect A7 and H1 (representing the social aspects). After modifications, items in the positive affect scale had been empirically confirmed to the established GoF criteria.

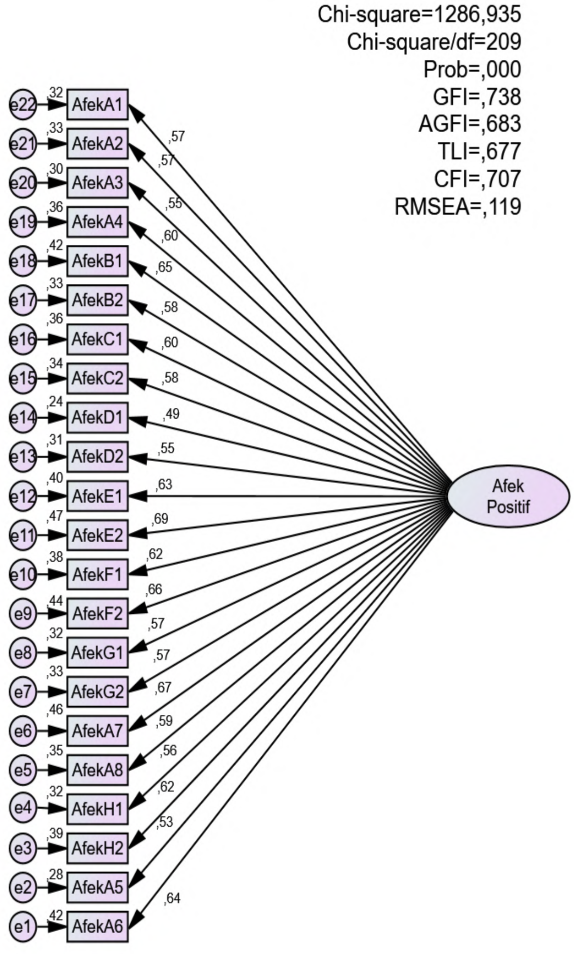


Figure 3. Analysis of confirmatory factors on Positive Affectivity Scale which is not accordance with the criteria of *Goodness of Fit* (GoF)

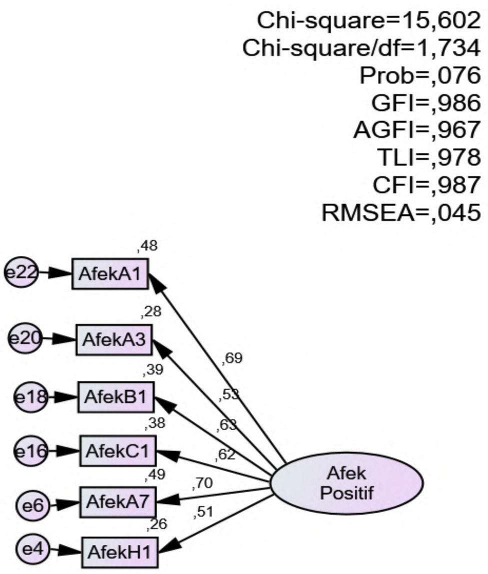


Figure 4. CFA on Positive Affect Scale based on Goodness of Fit (GoF)

Table 4. Summary of CFA on Positive Affect Scale based on GoF

Aspect	Item	Scores of Loading Factors	Significance (p)
Assignment	AffectA1	0.691	significant
	AffectA3	0.531	significant
	AffectB1	0.628	significant
	AffectC1	0.616	significant
Social	AffectA7	0.699	significant
	AffectH1	0.506	significant

Discussion

This research is one of several stages before testing the structural model of positive affect scale as one of the research instruments. The results of the study indicate that the positive affect scale in the academic setting is able to produce items that can reveal the latent constructs or concepts appropriately. There are six selected items in which four items represent the assignment aspects, and two items are related to the social aspects.

Associated with the development of positive affect instruments by Watson et al. (1988), positive affect instruments in the academic domain generated through this research enriched the study of previous positive affect instruments. The positive affect instrument of Watson et al., (1988) was general for all domains, while the positive affect instrument resulting from this study is more specifically revealing the positive affect that develops in academic settings. Thus, the discussion on positive effects in academic settings becomes more detailed and clear according to context.

This study can give beneficial contribution dealing with the limited studies on the affect in the academic setting as stated by Linnenbrink-Garcia and Pekrun (2011) and Pekrun et al. (2002). It is expected that psychological dynamics within the academic context can be investigated comprehensively to build appropriate and efficient solution towards various educational problems.

Conclusion

It is concluded that the validity test of positive affect scale within the academic domain can produce items that can reveal constructs or latent concepts appropriately. By having correct and proper information related to psychological dynamics within the academic context, it can support to create appropriate and efficient solution for various educational problems to improve the quality of education. Further studies are expected to continue its coverage on a wider range of area to see if the research findings can be applied to other unexamined subjects and contexts.

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Arbuckle, J. L. (2013). *IBM® SPSS® AmosTM 22 user's guide*. Chicago, IL: Amos Development Corporation.

Ashby, F. G., Isen, A. M., & Turken, A. U. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, *106*(3), 529–550. https://doi.org/10.1037/0033-295X.106.3.529

Azwar, S. (2016). *Konstruksi tes kemampuan kognitif*. Yogyakarta: Pustaka Pelajar.

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic concept, application, and programming* (2nd ed.). New York, NY: Routledge Taylor & Francis Group.

Dachlan, U. (2014). *Panduan lengkap Struktural Equation Modeling tingkat dasar: Metodologi, konsepsi, aplikasi (dengan Amos)* (1st ed.). Semarang: Lentera Ilmu.

Danner, D. D., Snowdon, D. A., & Friesen, W. V. (2001). Positive emotions in early life and longevity: Findings from the nun study. *Journal of Personality and Social Psychology*, *80*(5), 804–813. https://doi.org/10.1037/0022-3514.80.5.804

Demanet, J., Liefooghe, B., & Verbruggen, F. (2011). Valence, arousal, and cognitive control: A voluntary task-switching study. *Frontiers in Psychology*, *2*(336), 1–9. https://doi.org/10.3389/fpsyg.2011.00336

Folkman, S. (2008). The case for positive emotions in the stress process. *Anxiety, Stress and Coping*, *21*(1), 3–14. https://doi.org/10.1080/10615800701740457

Fredrickson, B. L. (1998). What good are positive emotions? *Review of General Psychology*, *2*(3), 300–319. https://doi.org/10.1037/1089-2680.2.3.300

Fredrickson, B. L., & Joiner, T. (2002). Positive emotions trigger upward spirals toward emotional well-being. *Psychological Science*, *13*(2), 172–175. https://doi.org/10.1111/1467-9280. 00431

Fredrickson, B. L., Mancuso, R. A., Branigan, C., & Tugade, M. M. (2000). The undoing effect of positive emotions. *Motivation and Emotion*, *24*(4), 237–258. https://doi.org/10.1023/a:1010796329158

Fredrickson, B. L., Tugade, M. M., Waugh, C. E., & Larkin, G. R. (2003). What good are positive emotions in crisis? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. *Journal of Personality and Social Psychology*, *84*(2), 365–376. https://doi.org/10.1037/0022-3514.84.2.365

Ghozali, I. (2017). *Model persamaan struktural Kkonsep dan aplikasi dengan program AMOS 24 Update Bayesian SEM* (7th ed.). Semarang: Badan Penerbit Universitas Diponegoro.

Goetz, T., Pekrun, R., Hall, N., & Haag, L. (2006). Academic emotions from a social-cognitive perspective: Antecedents and domain specificity of students' affect in the context of Latin instruction. *British Journal of Educational Psychology*, *76*(2), 289–308. https://doi.org/10.1348/000709905X42860

Hair, G., Black, B., Babin, B., Anderson, R., & Tatham, R. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson.

Kline, R. (2011). *Principle and practice of Sructural Equation Modeling* (3rd ed.). New York, NY: The Guilford Press.

Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement: Introduction to the special issue. *Contemporary Educational Psychology*, *36*(1), 1–3. https://doi.org/10.1016/j.cedpsych.2010.11.004

Linnenbrink, E. A. (2006). Emotion research in education: Theoretical and methodological perspectives on the integration of affect, motivation, and cognition. *Educational Psychology Review*, *18*(4), 307–314. https://doi.org/10.1007/s10648-006-9028-x

Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin*, *131*(6), 803–855. https://doi.org/10.1037/0033-2909.131.6.803

Nath, P., & Pradhan, R. K. (2012). Influence of positive affect on physical health and psychological well-being: Examining the mediating role of psychological resilience. *Journal of Health Management*, *14*(2), 161–174. https://doi.org/10.1177/097206341201400206

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. https://doi.org/10.4135/9781412985772

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology*, *41*(4), 359–376. https://doi.org/10.1111/j.146405 97.1992.tb00712.x

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105. https://doi.org/10.1207/S15326985EP3702_4

Pekrun, R., Goetz, T., Titz, W., Perry, R. P., Pekrun, R., Goetz, T., … Perry, R. P. (2010). *Academic Emotions in Students ' Self-Regulated Learning and Achievement : A Program of Qualitative and Quantitative Research Academic Emotions in Students ' Self-Regulated Learning and Achievement : A Program of Qualitative and Quantitative Research*. (October 2014), 37–41. https://doi.org/10.1207/S15326985EP3702

Samios, C., Abel, L. M., & Rodzik, A. K. (2013). The protective role of compassion satisfaction for therapists who work with sexual violence survivors: An application of the broaden-and-build theory of positive emotions. *Anxiety, Stress & Coping*, *26*(6), 610–623. https://doi.org/10.1080/10615806.2013.784278

Schutz, P. A., & Lanehart, S. L. (2002). Introduction: Emotions in education. *Educational Psychologist*, *37*(2), 67–68. https://doi.org/10.1207/S15326985EP3702_1

Seligman, M. E. P., Ernst, R. M., Gillham, J., Reivich, K., & Linkins, M. (2009). Positive education: Positive psychology and classroom interventions. *Oxford Review of Education*, *35*(3), 293–311. https://doi.org/10.1080/03054980902934563

Steptoe, A., Dockray, S., & Wardle, J. (2009). Positive affect and psychobiological processes relevant to health. *Journal of Personality*, *77*(6), 1747–1776. https://doi.org/10.1111/j.1467-6494.2009.00599.x

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235. https://doi.org/10.1037/0033-2909.98. 2.219

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*(4), 705–731. https://doi.org/10.1037/a0023980

Evaluation of the implementation of Batik-skills training program

***1Hendro Prasetyono; 2Dedeh Kurniasari; 3Laila Desnaranti**
1,3Department of Economics Education, Universitas Indraprasta PGRI
Jl. Nangka No 58C, Tanjung Barat, Jagakarsa, Jakarta Selatan 12530, Indonesia
2Lembaga Pelatihan dan Keterampilan Cosmos
Jl. Kasuari E2/106, Raya Patriot-Jakasampurna, Bekasi, Jawa Barat 17145, Indonesia
*Corresponding Author. E-mail: hen.dro23@yahoo.com

Submitted: 8 March 2019 | *Revised: 3 October 2019* | *Accepted: 29 October 2020*

Abstract

The purpose of this study is to evaluate the implementation of batik skills training program as a recommendation for program improvement. The method used in this research is a qualitative approach using Context, Input, Process, and Product evaluation model. Samples were taken from the Institute of Skills and Training in the areas of Jakarta, Bogor, Depok, Bekasi, and Tangerang. The results of the evaluation components that meet the evaluation criteria are all aspects of the context component, discipline and learning process, while the components of batik teachers' education qualifications, the use of educational facilities and infrastructure standards, curriculum components, program financing, evaluation of learning outcomes, mastery of theoretical competencies, practices, and impacts to program participants have not been met. The batik skills training program needs to be continued with some improvement. It is recommended for the product components, especially on the impact aspects felt by graduates, to be improved.

Keywords: *training program implementation, batik skill, learning outcomes, institute of skills and training*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.23918

Introduction

The development of batik in Indonesia today is impressive. Batik is very popular and growing rapidly in the country since the recognition of batik by UNESCO in 2009 as a world cultural heritage which is originated in Indonesia (Pancapalaga, Bintoro, Pramono, & Triatmojo, 2014). These developments affect and make people realize that batik can create jobs with special skills which are very promising. It is reinforced by Data from Central Java Industry and Trade Department (2001) that there are 11,391 of batik unit productions spreading at 146 production centers with production values of €5.4 million (Hunga, 2011).

Batik is a decorated textile made using the wax-resist method. Wax can be used to create intricate designs using an instrument called *canting*, and batik made using this vessel is called batik *tulis* (Lee, 2016). Batik development in Indonesia experienced various obstacles. One of the inhibiting factors in the development of batik industry using a printing machine. It happens because of advancements in information technology and knowledge which result in instability in all aspects of industrial life. It is necessary to use management theory to characterize and understand the nature of turbulence that occurred (Anderson, Mason, Hibbert, & Rivers, 2017).

One of the management theory implementations in the education field is the development of batik skill practice in the form of non-formal education. Non-formal education as part of the education system has the same task as formal education, which provides the best service to the community. Alternative services programmed outside the education can serve as a replacement, enhancer, and or complement to the formal education system.

According to Field in Hoppers (2006, p. 21), non-formal education is rarely used, the term lifelong learning has increasingly gained currency when referring to the totality of educational activities outside the school system. In the third form of education, the instructor and the learner are not in direct contact with each other. The learning material is provided for students by post and is written in a very simple language so that students can easily understand and comprehend the material. Diagrams and exercises are included in these courses to support and guide the students accordingly. Some materials are communicated through TV channel programs and the internet (Saif, Reba, & Din, 2017).

Non-formal education is usually not documented by a certificate or transcript. It occurs in educational institutions or public organizations, clubs, circles, as well as during individual sessions with teachers or coaches (Ivanova, 2016). Non-formal forms of education are often characterized by the strong participation of volunteers, viewed as an expression of the Protestant teaching of the priesthood of all believers (Schweitzer, 2017).

The targets of non-formal education are increasingly diverse, such as serving the poor, those who have not completed basic education, dropouts and dropping out of formal education, people who do not have access to formal education such as; isolated tribes, rural communities, border areas, and outer island communities and the development of a special skill or skill in the form of training.

The existence of non-formal education such as courses and training institutions helps the community to develop. To improve the quality of education and quality of human resources in the field of education, the government has issued a policy. Law of Republic of Indonesia No. 20 of 2003 on the national education system and the Regulation of the Minister of National Education No. 19 of 2005 on the national education standards, clause 26 verse 5 state that courses and training are held for people who need knowledge, skills, life skills, and attitudes to develop themselves, professions, work, independent business, or continue their education to a higher level.

One of non-formal education form which is widely adopted by the community is training. Mahapatro (2010, p. 252) states that training is the systematic development of the attitude, knowledge, skill pattern required by a person to perform a given task or job adequately. Mahapatro also state that development is 'the growth of the individual in terms of ability, understanding, and awareness.' Training is required to properly motivate and prepare the workers for operating these mechanisms effectively (White, 2004, p. 641).

There are three specific training objectives according to Stredwick (2005, p. 376): (1) to develop the competencies of employees and improve their performance, (2) to help people grow within the organization in order that, as far as possible, in the future, the needs for human resources can be obtained from within the organization, (3) to reduce learning time for employees starting in new jobs on appointment, transfer or promotion, and ensure they become fully competent as quickly and economically as possible. Besides, three components must be prepared in training: academic preparation, pedagogical skills, and teaching practice (Younus & Akbar, 2017).

The training approach can be used in batik development since the training delivery approach, containing a combination of technical and entrepreneurial skills, was relevant in responding to the needs and objectives of adult trainees (Mayombe, 2017). Training can address gap issues between education and industry needs. Perhaps one of the most ambitious challenges currently that instructors of introductory management courses have to deal with is their desire to make their courses more relevant and meaningful for today's learners, including students with little or no work experience (Wright & Gilmore, 2012 in Durant, Carlon, & Downs, 2017).

It is in accordance with Kyriakides and Campbell in Govender, Grobler, and Mestry (2016) that school quality improvement seems supporting a culture of school audits, the philosophy of which notes that governments are now formally placing increased expectations on school leadership/management teams to integrate self-evaluation into both their strategic and operational structures and procedures.

The batik training program is designed to accommodate training participants with knowledge, workability, designing, making, organizing, and packaging batik products needed by the industry. Batik will be preserved and in accordance with the needs of the industrial world. Hence, it is necessary to conduct research to evaluate the success rate of the batik skills training program. It is in accordance with the opinion of Posavac and Carey (1980, p. 8) regarding the six reasons for the need of evaluation: '(1) fulfillment of accreditation requirements; (2) accounting funds; (3) answering requests for information; (4) making administrative decisions; (5) assisting staff in program development; (6) learning about unintended effects'.

An evaluation is a systematic process that gives out information about program achievement (Dewi & Kartowagiran, 2018). Program evaluation is a systematic and ongoing process to collect, describe, interpret and present the information to be used as a basis to make decisions, develop policies, and preparing the next program (Prasetyono, 2016). Meanwhile, evaluation in training according to Torrington, Hall, and Taylor (2005, p. 402) 'Evaluation is straight forward when the output of the training is clear to see, such as reducing the number of dispatch errors in a warehouse or increasing someone's typing speed'. The essence of 'evaluation is a comparison, and surveys are one (but not the only) way to collect information useful for comparing programs or for comparing individual performance' (Langbein & Felbinger, 2006, p. 192).

Evaluation is about a particular initiative. It is generally carried out to assess the initiative, and the results are not generalizable. Evaluations are designed to improve an initiative and to provide information for decision making at the program or policy level; the research aims to prove whether there is a cause and effect relationship between two entities in a controlled situation (Harris, 2010, p. 2).

Evaluation is different from monitoring. According to Singh (2007, p. 54), evaluation is different from monitoring in many ways. Monitoring usually provides information regarding the performance of process indicators, whereas evaluation assesses the performance of impact indicators. Monitoring is an internal process where all concerned project staff devises a monitoring system, while evaluation is usually done by an external agency to assess the project's achievements. Evaluation is a selective exercise that attempts to systematically and objectively assess progress towards the achievement of an outcome.

The purpose of the evaluation according to Edwards, Scott, and Raju (2007, p. 58) is 'the ultimate goal for the evaluation team is to deliver the most useful and accurate information to key stakeholders most cost-effectively and realistically possible'. Program evaluation is a comprehensive approach that involves three general steps: (1) developing program theory, (2) formulating and prioritizing evaluation questions, and (3) answering evaluation questions (Donaldson & Scriven, 2008, pp. 109–110).

The purpose of this research is to evaluate the implementation of the batik skills training program as a recommendation for improving the batik training program in Indonesia based on the program evaluation result with Context, Input, Process, Product (CIPP) model. This CIPP model is suitable for evaluating formal and non-formal education programs, such as training (Madaus & Stufflebeam, 2002) because evaluation can answer the variation of the statement and determine the success in viewing the quality of education (Prasasti & Istiyono, 2018).

Method

The research method used a qualitative approach. Qualitative research is defined as the collection, analysis, and interpretation of comprehensive narrative and visual (i.e., no numerical) data to gain insights into a particular phenomenon of interest (Gay, Mills, & Asian, 2012, p. 7). The program evaluation model selected is CIPP. Corresponding to the letters in acronym CIPP, the model's core concepts are Context, Input, Process, and Product evaluation. Context evaluation assesses needs, problems, and opportunities as a base for defining goals and priorities and judging the significance of outcomes. Input evaluation assesses alternative approaches to meetings need as a means of plans to guide activities and later to help explain outcomes. The evaluative method was used in order to evaluate the process of product development (Istiyani, Zamroni, & Arikunto, 2017). Product evaluation identifies intended and unintended outcomes both to help keep the process on track and determine effectiveness (Stufflebeam, Madaus, & Kellaghan, 2002, p. 279).

The research was conducted on six Skills and Training Institutions (STI) under the *Asosiasi Profesi Batik dan Tenun Nusantara* or Association of Batik and Tenun Nusantara Professions 'Bhuana' (APBTN 'Bhuana'). The APBTN 'Bhuana' innovates in the planning of non-formal education programs for batik courses according to national education standards. APBTN 'Bhuana' under the Directorate of Community Education and Training Course of the Ministry of National Education of the Republic of Indonesia has a great influence and authority in conducting batik training in Indonesia. The six STIs are STI Cosmos (Bekasi City), STI Lesha (Bekasi Regency), STI Tradisiku (Bogor Regency), STI Asri (Depok City), STI Kris (DKI Jakarta), and STI Datik (Tangerang Regency). The total key informants in the interviews were 18 people and the respondents for the questionnaire were 120 people.

The data were collected using interviews, questionnaires, observations, and document studies. The research instrument was developed based on four components in the CIPP evaluation model. The component and aspect of program evaluation with the CIPP model are presented in Table 1.

The program evaluation criteria were divided into points of the interview, questionnaire, observation statements, and document review. The total is 66 items. Expert analysis was used in testing the instrument validity and reliability.

Table 1. Evaluation component and aspect of batik training program

Component	Aspect
1. The purpose of batik skills training	a. Formulation of goals b. Basic formulation c. The foundation of objective formulation deliberation
2. The Program Design	a. Formulation of the problem b. The use of standard learners c. The use of educational standards d. The use of educator and of Education staff standard e. The use of the curriculum program f. The determination of material program g. The design of learners tasks h. Financing
3. Program Implementation	a. The rules for learner discipline b. Calendar of events and learning schedule c. Syllabus and learning process plan d. The learning process e. Program Supervision f. Implementation of learning evaluation g. Supervision of learning evaluation
4. Program Implementation Results	a. Mastery of theoretical competencies b. Mastery of Practice Competencies c. Mastery of Attitudinal Competence d. Graduation certificates e. Opening new entrepreneurs f. Open your own training g. Become an educator h. Being a batik assessor

Data obtained from interviews, document studies, and observations in the analysis using a qualitative approach. Qualitative data analysis is data reduction, display data, and verification (Miles & Huberman, 1992, pp. 16–17). The survey questionnaire consisted of four-point Likert scale items and qualitative items that were developed from and linked to the reviewed literature. The questionnaire was supervised by five teachers to get feedback and make necessary changes; however, their responses were not included in the collected data and they were asked not to take the survey again (Mcminn, Kadbey, & Dickson, 2015). The evaluation study of the unit analysis program implementation focuses on how the process and quality training. According to Dane and Schneider and Dusenbury, et al. in Huff, Preston, and Goldring (2013), their coaches are differentiated in two key dimensions of implementation: dose and the quality of the program delivery.

Findings and Discussion

Findings

The first component is the context component in the evaluation with the CIPP model consisting of three aspects of evaluation. The first aspect is the policy background. Based on the results of interviews and study documents on the Law of Republic of Indonesia No. 20 of 2003, Regulation of the Minister of National Education No. 19 of 2005, and Regulation of the Minister of National Education No. 47 of 2010 on organizing courses, it is found that the batik skills program conducted by STI under the APBTN 'Bhuana' had a policy background in accordance with the law. The second aspect is the formulation of the objectives of the batik skills training program. Based on the results of interviews and document studies regarding the teaching design of each STI, the results of the formulation of program objectives are in line with the law and learning objectives.

The third aspect is the objectives of the Batik Skills Program. Based on interviews and studies of STI administrative documents, the results of the process of formulating the objectives of the program involved various central parties ranging from government elements to representatives from all regions in Indonesia. Besides, the program's objectives are in accordance with community needs, especially in increasing employment and creating national batik business development.

The second component is the Input component in the evaluation with the CIPP model consisting of six aspects of evaluation. The first aspect is the problem formulation of the batik skills program. Based on the results of interviews with managers, teachers, and participants of each STI, the results of the formulation of problems in each STI are adjusted to the conditions in the field and respond to challenges ahead, it's just not specific enough.

The second aspect is the standard of prospective students in the batik skills program. Based on the results of interviews with the study participants' biodata documents, it was found that there was a mismatch between the minimum educational background of program participants on the condition of being a program participant. Thus, it can be concluded that the standard aspects of prospective program participants 50% are met.

The third aspect is the teacher's academic qualifications for the batik skills training program. Based on the results of interviews and studies of training instructors' qualification documents, 67% of the educational qualifications of the trainers meet the educational qualification standards and about 33% have not met the educational standards. All teachers have certificates as a batik instructor issued by the Ministry of Education and Culture also Ministry of Manpower but have no educational background in art majors. There are 80% of teachers who have high school educational background and 65% of teachers have experience as a batik instructor at least for three years.

The fourth aspect is the use of facilities and infrastructure standards. Based on direct observations and photos of new infrastructure facilities, only 50% of STIs meet the standards. The fifth aspect is the batik skills training program curriculum. Based on the results of interviews and study of learning curriculum documents, the curriculum used have used the Indonesian National Qualification Framework or *Kerangka Kualifikasi Nasional Indonesia (*KKNI) up to level 3. The sixth aspect is the financing of the Batik Skills training program. Based on interviews and studies of STI financial report documents, each STI still relied on the money from the training participants as the main source of income.

The third component is the process component in the evaluation with the CIPP model consisting of three aspects of evaluation. The first aspect is the training rules. Based on the results of the distribution of questionnaires to 120 respondents from six STIs, interviews, and document studies, it is found that the results of the activity regulations have been implemented in accordance with activity planning, although not 100% of all STIs have implemented them.

The second aspect is the learning process of the Batik Skills Program. Based on the results of the distribution of questionnaires to 120 respondents from six STIs, interviews, and participants observation, the sub-aspect of teachers provide instructional objectives when starting training gains the achievement rate of 70%. Teaching sub-aspect explained the material with discussion methods with the level of achievement of 87.5%. As for the sub-aspects of the instructor in explaining the material using learning tools, the achievement level is 100%.

The third aspect is the evaluation of learning outcomes in the Batik Skills training program. Based on the results of the distribution of questionnaires to 120 respondents from six STIs, interviews, and participant observation, the highest answer for the sub-aspect of assignment at home is the STI Tradisiku, whereas the lowest number of answers is STI Lesha and STI Datik. The results of the recapitulation of the sub-aspects questionnaires show that the institution which most often achieved the highest number of answers to quizzes are STI Tradisiku, and the lowest number of answers is STI Datik and STI Cosmos.

The fourth component is the product component in the evaluation with the CIPP model consisting of four aspects of evaluation. The first aspect is the mastery of theoretical competence. Based on the results of the recapitulation of questionnaires, document studies, and interviews, the average score of quizzes and theory tests was 78. Meanwhile, for the sub-aspects of mastery of the theory of materials in making batik, 60% was achieved. Sub mastery of material aspects on various techniques in making batik by 42.57% was achieved. Sub mastery of material aspects regarding a variety of tools needed in the new batik process of 88.3% was achieved.

The second aspect is the mastery of practice competencies. Based on the results of a recapitulation of questionnaires, observations, and interviews, all participants get a minimum score of 'Good'.

The third aspect is the proof of graduation. Based on the results of the questionnaire recapitulation, the study of documents, and interviews, all participants get a mark of competency test certificate graduation. The fourth aspect is the impact on program participants. Based on the results of the questionnaire recapitulation, document studies, and interviews, it is found that the sub-aspects of the program graduates were able to open a batik production business with a percentage of the success rate of 20% with the highest number of alumni coming from STI Datik and STI Cosmos and the lowest number of alumni is from STI Kris and STI Asri.

For the sub-aspect of 'the program graduates can open a batik selling business', the percentage of success rate is 50% with the highest number of alumni coming from STI Traditional and the lowest number of alumni is from STI Lesha. For the sub-aspect of 'the program graduates are able to open a batik course', the percentage of success rate is 16.7% with the highest number of alumni coming from STI Asri and the lowest number of alumni from STI Kris and STI Datik. For the sub-aspect of 'the program graduates can open a batik workshop', the percentage of success rate was 23.3% with the highest number of alumni coming from STI Asri and Cosmos while the lowest number of alumni from STI Kris and STI Datik.

For the sub-aspect of 'the program graduates can open a tiered batik training', the percentage of the success rate is 16.7% with the highest number of alumni coming from STI Cosmos and the lowest number of alumni from STI Lesha and STI Tradisiku. For the sub-aspect of 'the program graduates can become batik educators in schools', the percentage of success rate is 86.7% with the highest number of alumni coming from STI Lesha, Traditional, and Datik, while the lowest number of alumni from STI Asri and STI Cosmos.

For the sub-aspect of 'the program graduates are financially capable', only 20% of the total respondents are financially capable and have started the business to open their own batik production, 50% of the total respondents who answered has opened a selling batik business and have their own store or cooperation with third parties; 13.3% of the total respondents who answered are financially capable and had opened a batik course. The majorities are constrained by complex funds and permits; 23.3% of total respondents are financially capable and have opened batik training workshops; 16.7% of the total respondents who answered are financially capable to open tiered batik training. The toughest constraint is the license and standards that must be fulfilled; 86.7% of the total respondents answered they are capable and had become batik educators in schools; 93.3% of the total respondents answered they can and have become batik educators in the general public.

Discussion

The first component in the evaluation of the CIPP model is the context component. The first aspect of the context component is the policy background. According to Ball policies 'create circumstances in which the range of options available in deciding what to do are narrowed or changed' (Ward et al., 2016). The batik skills program has a clear legal basis. Law of Republic of Indonesia No. 20 of 2003 on National Education System, clause 26 verse 5 up to the Regulation of the Minister of National Education No. 19 of 2005, on National Standard of Education, clause 1 verse 18 explain that education evaluation is an activity of controlling, guaranteeing, and determining the quality of education to various education component at every path, ladder, and type of education as a form of accountability of education. It is regulated in the implementation of Permit for Non-Formal Education and Early Childhood, Number: 421.10/1572/Dik/K.018. It is followed by all STIs which have a policy base in accordance with the law and the theory of (Thrupp & Robert, 2003, p. 195).

The second aspect is the basic formulation of program objectives. The basic formulation of program objectives is based on the analysis of learning conducted by non-formal education. It is in accordance with the Law of Republic of Indonesia No. 20 of 2003 on National Education System clause 26 verse 5 that courses and training are held for people who need knowledge, skills, life skills, and attitudes to develop themselves, develop their profession, work, independent business, and or continue to higher education. It is also in line with the definition of objectives in the evaluation programs outlined by Edwards et al. (2007, p. 58) that it is the ultimate goal for accurate information to key stakeholders in the most cost-effective and realistic manner possible. The basic formulation of the program should be able to provide accurate information based on the problems created in society.

The third aspect is the purpose of the program. The process of formulating the goals of the batik skills program involved various parties ranging from the Training Director of Course and Training, Directorate General of Early Childhood Education and Community Education, chairman of the APTBN 'Bhuana', and representative of every STI in Indonesia. The objectives of the program are as follows: (1) preserving batik art, (2) improving the quality and quantity of batik, (3) developing batik as one of the professions in education like a professional teacher, (4) improving the welfare of batik with competency certificates, (5) preparing prospective professsional educators before working for the community, (6) creating entrepreneurship and employment opportunities for people in batik. Context evaluation measures the needs, based on objectives and priorities and also assesses the results significantly (Stufflebeam et al., 2002, p. 279). Context evaluation assesses activities on batik skills which determine the situation and background that affect the types of strategic objectives to be developed in the system. The idea is inappropriate with Frye and Hemmer (2012) that context is a study that identifies and defines program goals and priorities by assessing needs, problems, assets, and opportunities relevant to the program.

The first aspect is problem formulation. The problem formulation on each STI matched with the conditions in the field and is oriented to the future, only less specific. In brief, each STI requires the right strategy and plan. It is in line with the understanding of the strategy by Fidler (2002, p. 10) that strategy is the direction and scope of an organization over the long term goal which achieves advantages for the organization through its configuration of resources within a changing environment, to meet the needs of markets and to fulfill stakeholder expectations. Meanwhile, strategy in management education according to De Kluyver and Pearce II (2009), Gamble, Thompson and Peteraf (2013), Mintzberg, Lampel, Quinn, and Ghoshal (2003), as well as Thompson, Strickland, and Gamble (2010) cited by Albert and Grzeda is described as a strategic framework that includes an analytical process relying on several prescribed tools and expects the student to arrive at a list of strategic options and subsequent recommendations for implementation (Albert & Grzeda, 2015).

This strategy is included in the plan which will be achieved by the training program graduates. The findings are in accordance with the concept of planning said by Yukl (2010, p. 72) that planning is a broadly defined behavior that includes making decisions about objectives, priorities, strategies, organization of the work, assignment of responsibilities, scheduling of activities, and allocation of resources among different activities according to their relative importance.

The second aspect is the standard of prospective learners. Each STI has met the criteria in the standard of prospective learners. The ages of prospective learners vary from the youngest in high school student grade 2 and the oldest is 53 years old. The educational background is diverse, ranging from in with a background of junior high school education to undergraduate (S1). It is in accordance with regulations where trainees are at least 17 years of age with diverse educational backgrounds.

The third aspect is the trainers' educational qualification. The findings, as compared to the prevailing regulations, need to be reanalyzed according to the Regulation of the Minister of National Education No. 19 of 2005 clause 29, verse 4, that educators at senior high school or *Madrasah Aliyah* (Islamic-based senior high school), or other equivalent forms of education have: (a) a minimum education qualification of Diploma-four (D-IV) or Bachelor (S1); (b) a higher education background with an education program appropriate to the subject being taught; and (c) professional teacher certificate for senior high school. Whereas, it includes educators at the course and training skills institutions which consist of teachers, mentors, trainers or instructors, and examiners. Therefore, it means that the faculty of the course and training institutions may come from faculty in higher education institutions. According to Sullivan, Mackie, Massy, and Sinha (2012, p. 24), higher education qualifies graduates for jobs or additional training as well as increasing their knowledge and analytic capacities. These benefits of undergraduate, graduate, and professional education manifest as direct income effects, increased social mobility, and health, as well as other indirect effects.

The fourth aspect is the condition of the facilities and infrastructure. The findings at the Institute of Skills and Education (STI) Cosmos, STI Lesha, and STI Tradisiku have met the minimum standards of educational facilities and infrastructure and the other STI has not met the criteria. When referring to the Regulation of the Minister of National Education No. 19 of 2005 clause 42 verse 1 and 2, the minimum standards of facilities and infrastructure that must be owned by universities as Institute of Teachers' Education or *Lembaga Pendidikan Tenaga Kependidikan* (LPTK) are as follows: (1) each educational unit shall have facilities including furniture, educational equipment, educational media, books, and other learning resources, consumables, and other equipment necessary to support a regular and continuous learning process. (2) Each educational unit is required to have infrastructure covering land, classrooms, head unit room, educator room, administrative room, library room, laboratory space, workshop space, production unit space, canteen, power and service installation, sports venues, places for worship, playgrounds, creative venues, and other space or places needed to support the regular and ongoing learning process.

The findings are in line with the theories of Fry, Ketteridge, and Marshall (2009, p. 308) that 'most educational organizations now work under the pressure of a system in which space in their buildings and infrastructure is measured and accounted for in relation to student numbers and activities'. Based on the aforementioned findings, the researchers conclude that the condition of infrastructure from three of six STIs as the organizer of batik training has been feasible.

The fifth aspect is the batik training program curriculum. All STIs have a Competency-Based Curriculum design. The six STIs have a standard design of batik graduation levels: level 1, 2, and 3. All STIs have skill program material. Based on all these findings, it can be concluded if aspects of the program curriculum have met the criteria.

The sixth aspect is training program financing. This finding is less relevant when compared to the existing provisions. Standard of financing is the standard that regulates the components and the amount of operating unit cost of education applicable for one year (Fry et al., 2009, p. 3). Meanwhile, the financing standard according to the Regulation of the Minister of National Education No. 19 of 2005 clause 62 is elaborated as follows: Educational financing consists of investment cost, operating cost, and personal cost. (1) The investment unit cost of education covers the cost of providing facilities and infrastructure, human resource development, and fixed working capital. (2) The personal cost covers the tuition fees that must be paid by the learners in order to be able to follow the learning process regularly and continuously. The operating cost of the educational unit includes (a) educators and education personnel salaries and all allowances attached to the salary, (b) educational materials or equipment, and (c) indirect education operating costs in the form of electric power, water, telecommunication services, facilities maintenance and infrastructure, overtime pay, transportation, consumption, taxes, insurance, and so forth. (3) The standard operating cost of the education unit shall be stipulated by the Minister of Regulation based on the BSNP proposal.

Every program must have good financial management. Shattock (2003, p. 30) states that financial management emphasizes integrity, frugality, a concern for the pennies rather than the pounds, and a reluctance to borrow, the more it will command internal respect and provide a secure financial base for acting opportunistically and responding quickly to environmental change. Conservative financial control mechanisms, however, can create unnecessary layers of hierarchy and bureaucracy and can choke initiative.

The input component of the evaluation on the batik skills program which consists of the problem formulation and the students' standard has met the criteria. However, the batik teacher's qualification, the use of educational facilities and education standards, the curriculum, and the program financing components have not been fulfilled the criteria.

Evaluation at the stage of the process is to see the program achievements and the obstacles encountered by STI in running the batik program. The evaluation studied the program implementation which emphasizes the process components (Tuytens & Devos, 2014). The further evaluation process focuses on the coaching and professional growth of the teacher during at least two evaluation conferences (one formative and one summative). At the end of this process, an evaluation report is handed to the teacher. In this report, the teacher receives a final conclusion (two possibilities, namely: satisfactory or unsatisfactory) about his or her performance.

The first aspect is the discipline in training. Each STI has discipline regulations for their trainees, but only 80% of the participants who have received the socialization of the contents.

The second aspect is the process of teaching and learning. Teachers provide instructional goals when they start the training. It has a 70% achievement level, while for learning with an explanation with a two-way model, the achievement level is 87.5%. Besides, for the use of learning tools in teaching materials, the achievement level is 100%.

The third aspect is learning evaluation. Not all STIs provide assignments at home. Only STI Tradisiku and STI Asri give a home assignment, and STI Tradisiku is the only one that always gives their trainee quizzes. All STIs conduct formative, summative, and also competency tests. Stufflebeam and Shinkfield (2007, p. 294) said that the evaluation process is an ongoing check on a plan's implementation plus documentation of the process, including changes in the plan and key omissions or poor execution of certain procedures. One of the goals is to provide staff and managers feedback about the extent to which staff is carrying out planned activities on schedule, as planned, and efficiently. This opinion is supported by Patton (1980, p. 60) that the 'process' focus in an evaluation implies an emphasis on looking at how product or outcome is produced rather than looking at the same product itself; that is, it is an analysis of the processes whereby a program produces the result it does. Process evaluation is developmental, descriptive, continuous, flexible, and inductive.

Process components evaluation emphasizes how the product is produced compared to the product itself. The process stage in this research involves the aspects of strategy implementation and the use of facilities or capital or resources in real activities in the field. Thus, it is concluded that the components of the evaluation process from the batik skills training program implementation consist of the order and process of learning activities that have met the criteria, while the evaluation of learning outcomes has not met the evaluation criteria.

Product evaluation is the next stage of the evaluation program in the CIPP model. It is directed to the things that show the changes that occur in the input and what kind of results. Product evaluation serves to interpret success in achieving objectives, assessing the data set, comparing the established criteria with the results obtained in the field, and the considerations associated with the context, inputs, and processes and formulating a rational interpretation. The products from this program are batik professionals so the measured product component is the competence of batik skill training program graduates.

The first aspect is the mastery of the batik theory of program graduates. All the STIs have the average score from the quiz and theoretical exams are above 70. Based on the previously-mentioned findings, the mastery of batik material and various technique theories have not been fulfilled except the mastery of theory about the tools required in batik which has met the criteria.

The second aspect is the Competency Mastery of Program Practice Graduates. All STIs have implemented these criteria and the majority of trainees score above 80%. If there are participants who score below 80, then the STIs held the remedial exam. If the trainees did not pass the remedial exam, then they are required to repeat the training. There are 86.7% of participants from all STIs who can use *canting*, while 100% of participants can use batik *cap*, 100% of participants can prepare the wax, 87.5% of participants can perform the coloring process, and only 82.5% can perform the process of *pelorotan*.

The third aspect is the proof of graduation. All participants get a certificate of the graduate program but not all participants get a certificate of batik competence. If they want to get a certificate of batik competence, they must take the test conducted by Place Competency Test (PCT). It means that after passing the program, a participant will not always be recognized as a batik professional. He/she needs to take another exam recognized by the Ministry of Education and Culture.

The fourth aspect is the impact toward the participant of the batik skill program. It is the final stage of a series of evaluations program in product components. Impact evaluation is directed at focusing on the impact questions felt by the participants after the program. It serves to analyze the influence felt by the participants after the program.

Rossi, Freeman, and Lipsey (1999, p. 25) believes that impact evaluation is directed at focusing on the impact questions felt by the participants after the program. The evaluation questions are about which impact assessment is organized that is related to such matters as whether the desired program outcomes were attained, whether the program was effective in producing a change in the targeted social conditions, and whether the program impact included to unintended side effects. These questions assume a set of operationally defined objectives and criteria for success.

It is also important to know whether the impact measured has a positive or negative impact, the kind of impact, how big the impact is, and whether it can give a positive change for the participants for their future, so it can be analyzed for further program improvement. Stufflebeam et al. (2002, p. 229) said that to assess performance beyond goals, evaluators need to search for unanticipated outcomes, both positive and negative. They might conduct hearings or group interviews to generate hypotheses about the full range of outcomes and follow these up with an effort to confirm or disconfirm the hypotheses.

The findings on the impact aspect which is the most important thing for the individual after a program is to get a better job or income. It is in line with Sullivan et al. (2012, p. 24) that the benefits of undergraduate, graduate and professional education manifest as direct income effects, increased social mobility, health, and also other indirect effects. Measures have been created to monitor changes in these outputs, narrowly defined: numbers of degrees, time to degree, degree mix, and the like. Attempts have also been made to estimate the benefits of education using broader concepts, such as the accumulation of human capital. For estimating the economic returns to education, a starting point is to examine income differentials across educational attainment categories and institution types, attempting to correct for other student characteristics.

All the evaluation stages, starting from the evaluation of context, input evaluation, process evaluation, and product evaluation, are a unity that cannot be separated, but it has to be done together depending on the conditions in the field. The findings are in accordance with Stufflebeam and Shinkfield (2007, p. 294) that the purpose of product evaluation is to measure, interpret, and judge an enterprise's achievements. Its main goal is to ascertain the extent to which the evaluation met the needs of all the rightful beneficiaries.

Based on the findings, the number of stakeholders who have been made as the respondent has met the requirements. According to Bamberger, Rugh, and Mabry (2006, p. 271), program impact and quality cannot be determined without understanding the diverse experience of stakeholders. The perceptions of many must be search out.

The impact aspect on the evaluation of the implementation of batik skills program has given many positive influences to the program participants but has not given certainty about the condition of their destiny in the future, because those who have passed the program have not been able to open their own job or looking for work as batik craftsman. Product components in the evaluation of the program which consist of mastery of theoretical competence has not been fulfilled, mastery of practice competence and pro-gram pass mark has been fulfilled. The impact aspect on the program participants has not provided a promising future when the program participants have completed their training.

Conclusion

Referring to the research findings and discussion, all evaluation component in the implementation of the training program is carried out well, except the results of program implementation. The results that must be improved is the impact that the participants will get after attending the training at the STI. If STI graduates cannot work either opening a batik business or work as batik craftsman, the batik skills training program may not develop. If it does not develop, there is no generation who will continue their batik skills. It is recommended for the Ministry of National Education to collaborate with the Ministry of Cooperatives and Small Medium Enterprises, local governments, and related agencies to provide business opportunities to STI graduates in the form of soft loan assistance, entrepreneurship training, and facilitation in providing business licenses.

Acknowledgments

The researchers would like to thank the Association of Batik and Tenun Nusantara Professions 'Bhuana' (APBTN 'Bhuana') for giving permission and supporting the authors to conduct this research.

References

Albert, S., & Grzeda, M. (2015). Reflection in strategic management education. *Journal of Management Education*, *39*(5), 650–669. https://doi.org/10.1177/1052562914564872

Anderson, L., Mason, K., Hibbert, P., & Rivers, C. (2017). Management education in turbulent times. *Journal of Management Education*, *41*(2), 303–306. https://doi.org/10.1177/1052562916682208

Bamberger, M., Rugh, R., & Mabry, L. (2006). *Real world evaluation*. London: SAGE Publications.

Dewi, L. R., & Kartowagiran, B. (2018). An evaluation of internship program by using Kirkpatrick evaluation model. *REiD (Research and Evaluation in Education)*, *4*(2), 155–163. https://doi.org/10.21831/reid.v4i2.22495

Donaldson, S. I., & Scriven, M. (2008). *Evaluating social programs and problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Durant, R. A., Carlon, D. M., & Downs, A. (2017). The efficiency challenge: Creating a transformative learning experience in a principles of management course. *Journal of Management Education*, *41*(6), 852–872. https://doi.org/10.1177/1052562916682789

Edwards, J. E., Scott, J. C., & Raju, N. S. (2007). *Evaluating human resources programs*. San Francisco, CA: John Wiley & Sons.

Fidler, B. (2002). *Strategic management for school development: Leading your school's improvement strategy*. https://doi.org/10.4135/9781446219614

Fry, H., Ketteridge, S., & Marshall, S. (Eds.). (2009). *A handbook for teaching and learning in higher education: Enhancing academic practice* (3rd ed.). New York, NY: Routledge.

Frye, A. W., & Hemmer, P. A. (2012). Program evaluation models and related theories: AMEE guide no. 67. *Medical Teacher*, *34*(5), e288–e299. https://doi.org/10.3109/0142159X.2012.668637

Gay, L. R., Mills, G. E., & Asian, P. (2012). *Educational research: Competencies for analysis and applications*. Boston, MA: Pearson.

Govender, N., Grobler, B., & Mestry, R. (2016). Internal whole-school evaluation in South Africa. *Educational Management Administration & Leadership*, *44*(6), 996–1020. https://doi.org/10.1177/1741143215595414

Harris, M. J. (2010). *Evaluating public and community health programs*. San Francisco, CA: Jossey-Bass.

Hoppers, W. (2006). *Non-formal education and basic education reform: A conceptual review*. Paris: International Institute for Educational Panning UNESCO.

Huff, J., Preston, C., & Goldring, E. (2013). Implementation of a coaching program for school principals. *Educational Management Administration & Leadership*, *41*(4), 504–526. https://doi.org/10.1177/1741143213485467

Hunga, A. I. R. (2011). Uncover the invisible: Home-workers in micro-small-medium industries based on "putting-out" system (The case study of the Batik and Batik convection industry in a Sragen-Surakarta-Sukoharjo cluster of Indonesia). *The International Journal of Interdisciplinary Social Sciences*, *5*(9), 311–322.

Istiyani, D., Zamroni, Z., & Arikunto, S. (2017). A model of madrasa ibtidaiya quality evaluation. *REiD (Research and Evaluation in Education)*, *3*(1), 28–41. https://doi.org/10.21831/reid.v3i1.13902

Ivanova, I. V. (2016). Non-formal education: Investing in human capital. *Russian Education & Society*, *58*(11), 718–731. https://doi.org/10.1080/10609393.2017.1342195

Langbein, L., & Felbinger, C. L. (2006). *Public program evaluation*. New York, NY: M. E. Sharpe.

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

Lee, T. (2016). Defining the aesthetics of the Nyonyas ' Batik Sarongs in the straits settlements, late nineteenth to early twentieth century. *Asian Studies Review*, *40*(2), 173–191. https://doi.org/10.1080/10357823.2016.1162137

Madaus, G. F., & Stufflebeam, D. L. (2002). *Evaluation in education and human services*. Basel: Springer Nature.

Mahapatro, B. B. (2010). *Human resource management*. New Delhi: New Age International Limited.

Mayombe, C. (2017). Integrated non-formal education and training programs and centre linkages for adult employment in South Africa. *Australian Journal of Adult Learning*, *57*(1), 105–125. Retrieved from https://www.ajal.net.au/integrated-non-formal-education-and-training-programs-and-centre-linkages-for-adult-employment-in-south-africa/

Mcminn, M., Kadbey, H., & Dickson, M. (2015). The impact of beliefs and challenges faced, on the reported practice of private school science teachers in Abu Dhabi. *Journal of Turkish Science Education*, *12*(2), 69–79. https://doi.org/10.12973/tused.10141a

Miles, M. B., & Huberman, A. M. (1992). *Qualitative data analysis*. Beverly Hills, CA: SAGE Publication.

Pancapalaga, W., Bintoro, V. P., Pramono, Y. B., & Triatmojo, S. (2014). The chrome-tanned goat leather for high quality of Batik. *Journal of the Indonesian Tropical Animal Agriculture*, *39*(3), 188–193. https://doi.org/10.14710/jitaa.39.3.188-193

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills, CA: SAGE Publication.

Posavac, E. J., & Carey, R. G. (1980). *Program evaluation: Methods and case study*. Englewood Cliffs, NJ: Prentice-Hall.

Prasasti, I. H., & Istiyono, E. (2018). Developing an instrument of national examination of equivalency education Package C of mathematics subject. *REiD (Research and Evaluation in Education)*, *4*(1), 58–69. https://doi.org/10.21831/reid.v4i1.15556

Prasetyono, H. (2016). Graduate program evaluation in the area leading educational, outlying and backward. *Journal of Education and Practice*, *7*(36), 109–116. Retrieved from https://www.iiste.org/Journals/index.php/JEP/article/view/34641

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

Regulation of the Minister of National Education No. 47 of 2010, on the Competence Standard of Training Graduates. , (2010).

Rossi, P. H., Freeman, H. E., & Lipsey, M. (1999). *Evaluation: A systematic approach* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Saif, P., Reba, A.-, & Din, J. U. (2017). A comparitive study of subject knowledge of B.Ed graduates of formal and non-formal teacher education systems. *Journal of Education and Educational Development*, *4*(2), 270–283. https://doi.org/10.22555/joeed.v4i2.1354

Schweitzer, F. (2017). Researching non-formal religious education: The example of the European study on confirmation work. *HTS Teologiese Studies / Theological Studies*, *73*(4), 1–9. https://doi.org/10.4102/hts.v73i4.4613

Shattock, M. (2003). *Managing successful universities*. London: Open University Press.

Singh, K. (2007). *Quantitative social research methods*. New Delhi: SAGE Publications.

Stredwick, J. (2005). *An introduction to human resource management*. Oxford: Elsevier.

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2002). *Evaluation models: Viewpoints on education and human services evaluation* (2nd ed.). Boston, MA: Kluwer Academic Publisher.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models and applications*. San Francisco, CA: Jossey-Bass.

Sullivan, T. A., Mackie, C., Massy, W. F., & Sinha, E. (Eds.). (2012). *Improving measurement of productivity in higher education*. https://doi.org/10.17226/13417

Thrupp, M., & Robert, W. (2003). *Education management in managerialist times: Beyond the textual apologists*. Philadelphia, PA: Open University Press.

Torrington, D., Hall, L., & Taylor, S. (2005). *Human resource management Sixth Edition* (6th ed.). London: Pearson Education.

Tuytens, M., & Devos, G. (2014). The problematic implementation of teacher evaluation policy. *Educational Management Administration & Leadership*, *42*(4_suppl), 155–174. https://doi.org/10.1177/1741143213502188

Ward, S., Bagley, C., Lumby, J., Hamilton, T., Woods, P., & Roberts, A. (2016). What is 'policy' and what is 'policy response'? An illustrative study of the implementation of the Leadership Standards for Social Justice in Scotland. *Educational Management Administration & Leadership*, *44*(1), 43–56. https://doi.org/10.1177/1741143214558580

White, C. (2004). *Strategic management*. New York, NY: Palgrave Macmillan.

Younus, F., & Akbar, R. A. (2017). Comparison of evaluation methods of teaching practice of formal and non-formal teacher education institutions of Punjab. *Bulletin of Education and Research*, *39*(1), 159–173. Retrieved from http://pu.edu.pk/images/journal/ier/PDF-FILES/12_39_1_17.pdf

Yukl, G. A. (2010). *Leadership in organizations*. Upper Saddle River, NJ: Prentice Hall.

REiD (Research and Evaluation in Education), 5(2), 2019, 144-151
Available online at: http://journal.uny.ac.id/index.php/reid

The respondent factors on the digital questionnaire responses

***1Muhardis; 2Burhanuddin Tola; 3Herwindo Hariwibowo**
1,2,3Department of Educational Research and Evaluation, Universitas Negeri Jakarta
Jl. R. Mangun Muka, Rawamangun, Pulo Gadung, Jakarta Timur, DKI Jakarta 13220, Indonesia
*Corresponding Author. E-mail: adi_perdana2000@yahoo.com

Submitted: 2 September 2019 | Revised: 19 November 2019 | Accepted: 27 November 2019

Abstract

Progress in the field of technology often facilitates human work. One of them is progress in the development of questionnaire modes. Currently, existing questionnaires have been based on a digital platform, which makes evaluators easy to design, disseminate, and conduct scoring. All are computer-based, making them reachable by the respondents no matter how far the location of the respondent is, as long as they are connected to the internet. However, any progress is accompanied by several obstacles. For example, the respondents experienced an error in responding to having the intent to respond 'Yes' option but pressing the 'No' button instead. It is very different from filling in paper and pencil based questionnaires in which they are sure to put a checkmark using a pencil on the answer choices. This problem is what the researchers found when distributing digital questionnaires to participants of the National Questions Writing Program based on the 'SIAP' (Sistem Inovatif Aplikasi Penilaian) application. On conditional questions (if you choose 'No', please stop), some respondents who have chosen 'No' answers still respond to the next questions. It causes the data obtained are unreliable. After conducting a more in-depth analysis, the researchers found that respondents' factors as psychological factors are the cause, such as the new experience of accessing applications, understanding of applications, stress, and personal health. Uniquely, the respondents who have problems are those in the context of productive age, i.e 30 to 39 years old, more than five years of teaching experience, postgraduate level, and female.

Keywords: *computer-based assessment, digital questionnaire, respondent factors, SIAP application*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.26943

Introduction

Progress in the field of assessment is increasingly rapid. Assessment is no longer carried out in conventional ways, such as paper-based test (PBT) but has been in the form of a computer-based test (CBT). Many experts argue that well-designed computer-based assessments are considered more comprehensive, more accurate, and able to describe the profile of each test participant (Nezami & Butcher, 2000). Computer-based assessment is also the first step in the test model for the future (Yanxia, 2017) by not ignoring other aspects of the application that can cause the computer to make an error in processing the response given (McArthur & Choppin, 1984).

Errors are predictable and deserve more attention than the application of CBT, not only errors from the application side but also those from the respondent's side. Web-based questionnaires, as one of the other forms of application of CBT, as they have been used in Australia (Nulty, 2008), also need to pay attention to these errors. Before being applied, it is better for prospective respondents to be given socialization, training, and how to use the application (Sevillano-García & Vázquez-Cano, 2015).

Web-based questionnaires, also known as digital questionnaires, need to pay attention to some aspects such as questions asked and visual design so respondents are more aware in answering so that there are no mistakes when answering. Excessive use of media, such as quality-check reminders, will take a long time which can make the respondents frustrated and leave the questionnaire prematurely (Reja, Manfreda, Hlebec, & Vehovar, 2003).

It also happened to the web-based questionnaire used by the Center for Educational Assessment or *Pusat Penilaian Pendidikan (Puspendik)* as an instrument for measuring the level of satisfaction with the implementation of the national question writing application based on SIAP. As a new program that has only been operating since 2017, SIAP needs to be evaluated to see its effectiveness.

Evaluation of the participants' satisfaction level was done using a web integrated with the SIAP application. On the questionnaire's initial page, it was stated that the questionnaire was arranged to increase the capability of SIAP in providing information technology services to users, especially writers, reviewers, and material administrators. There are two types of responses provided in this questionnaire: Yes/No answers and gradation of five Likert scales approval levels (SIAP Application Development Questionnaire, n.d.).

It is interesting to find based on the response that there were respondents who still answered follow-up questions that they should not answer if they chose the response 'No'. An example can be seen in Table 1.

Table 1. Communication results with the person in charge of the subject matter

Choice of Answer	Number of Respondents (n=295)	Percentage (%)
Yes	200	47.96
No	92	21.82
Abstain	3	30.21

Table 1 is the result of question data regarding communication with the person in charge (PIC) of the subject matter. Of the 200 respondents, 92 people answered that they did not communicate with the PIC of the subject matter. However, of the 92 people, 30 of them (32.60%) still answered the follow-up questions, namely regarding clarity of indicators, improvements, and rejections.

It shows that web-based questionnaires require the ability of 'extensive reading' rather than 'speed reading'. In fact, more than 60% of respondents are postgraduate students. Thus, the habit of delaying the task of reading during education (Onwuegbuzie et al., 2004) or the habit of skimming reading (Cheng & Tsai, 2017) can be the cause. The respondents who graduated from the *Bahasa* Indonesia major did not show a much different pattern. In addition, most of them are male. Indeed, the results of the study by Parquette (1952) show that men find it difficult to understand a text in the first chance of reading.

Data that are duly not filled, but still get the response, as mentioned earlier, is said to be data outliers or outlier data. These data cause the shrinking parameter to be zero so the model used becomes inappropriate, or there is a bias in the parameters (Martel, 2015). Further, these outliers can also influence subsequent observations which have implications for the resulting parameter estimators (McQuarrie & Tsai, 2003). It may cause fatal errors if the next statistical analysis cannot be performed optimally only because of the data outliers.

This study focuses on the context 'attached' to the outliers: explaining the context behind the outliers' emergence. The context, in this case, refers to the things attached to the respondent, such as gender, age, education level, and length of teaching time. In contrast to studies conducted by psychometrists, such studies on outliers only focused on the outliers location in the data (Ali, 1994), statistical behavior tests (Prescott, 1978), and outliers' patterns in the contingency tables (Rapallo, 2012), this study looks at outliers from a qualitative perspective, which uses statistical data, that are, nonparametric scale.

Method

Data on the outliers and respondents' identity (context) were obtained from secondary data (Harris, 2001) from the web questionnaire (SIAP Application Development Questionnaire, n.d.) in the SIAP application. The respondents are 295 people from national question writers from national selection results in 2016 and 2017. The number of respondents sampled was all respondents who answered 'No', but still answered follow-up questions for three questions, 106 respondents for questions took a long time accessing the SIAP application, 92 respondents for communication questions with the subject matter PIC, and four people for communication questions with the SIAP administrator.

The questionnaire used as the data source is in the form of a digital questionnaire consisting of five parts, namely, the first part containing the identity of the respondents (name, subject, post, city, province, age, gender, the experience of accessing SIAP applications, length of teaching, level of education, suitability for subjects taught), the second part in the form of questions about the time needed to access the SIAP application, the third part in the form of questions about communication with the PIC of the subject matter, the fourth part which is a question regarding communication with the SIAP administrator, and the last part in the form of suggestion. The stages of research are guided by evaluative methods (Ross & Cronbach, 1976) and (Fíncher, 1981), namely by reading-intensive results of the response, especially responses related to questions that have further questions; recording into an identity data card (in this case the context in the form of gender, recent education, and length of teaching time) of respondents who are outliers; doing coding based on context; grouping the data; giving meaning to the context of outliers; drawing a conclusion; and compiling reports.

Findings and Discussion

The analysis was done on questionnaires distributed to national question writers about the satisfaction level with the SIAP application, communication with the subject matter PIC, and communication with the administrator. The questions on the SIAP application include the need to access the SIAP application for a long time, interactive features, and overall satisfaction with the SIAP application. The level of satisfaction with regard to communication with the PIC of the subject matter contains questions about the explanation of the indicators, remedial comments, and explanations of rejection. Communication with the administrator contains questions related to the call center, WhatsApp, and email.

Table 2. Response to the SIAP access time, communication results with the PIC of the subject matter and the SIAP administrator

	SIAP Access Time (n=295)	**Communication with the PIC of the Subject Matter (n=295)**	**Communication with SIAP Admin (n= 295)**
Number of respondent answered *Yes*	189	200	266
Number of respondent answered *No*	106	92	4
Number of abstain respondent answered follow-up questions	86	30	3
a. Gender			
Male	44	13	1
Female	42	17	2
b. Level of education			
Graduate	46	17	-
Post graduate	40	13	3
c. Age			
25 to 29 years old	6	-	-
30 to 39 years old	37	14	1
40 to 49 years old	36	11	1
50 years old or more	6	4	1
d. Length of teaching time			
1 year	19	6	1
2 years	10	3	-
3 years	9	4	-
4 years	2	2	-
5 years or more	45	15	2

Based on the results of the analysis, information was obtained on the percentage of respondents who answered 'No' but still answered the follow-up questions for the three types of questions, respectively 81.13% (for questions that took a long time accessing the SIAP application), 32.6% (for communication questions with the PIC of the subject matter, and 75% (for communication questions with the SIAP administrator). The details can be seen in Table 2.

Taking a Long Time to Access the SIAP Application

According to research conducted by several experts about computer-based applications, the average access needed to open applications is 4.7 seconds (Hastomo & Yuhana, 2013), 0.93 seconds (Shubhi, Yuliana, & Winarno, 2011), and 6.959 kB/s (Mulyana & Sholekan, 2010). However, each application used certainly has different access times. Software specifications used, when the situation opens access (morning, noon, or night) (Mulyana & Sholekan, 2010), availability of Wifi (Pangesti, 2017), as well as the software used are some of the factors that cause the speed of loading or data access.

Surely, it also applies to the SIAP application that is accessed by the writing participants the question that became the respondent of this study. Of the 295 respondents who participated, 189 respondents answered 'Yes' and 106 respondents answered 'No'. It means that most respondents need a long time to access SIAP. There were 86 people from all respondents who answered 'No' (45.50%) respond to the next questions. This number 86 consists of 42 female and 44 male respondents. It seems to support the results of the study by Parquette (1952) that men tend to be difficult to understand reading at once reading. Although only two points adrift, it can be said that men need time to understand reading compared to women. In fact, in the research conducted by Qian, Buchmann, and Zhang (2018), women tend to have good educational adaptability.

However, if they hold the opinions of respondents number 31 and 82 (both women of the same age), in the part of the question regarding the experience of accessing the SIAP application, they wrote that the problem they experienced when accessing was network. They will not have trouble as long as the signal when they open the application is in a strong position or from the side of the SIAP application itself, which is in maintenance status. Indeed, there are respondents (number 71, male) who say it takes two minutes at the beginning of the login, but it becomes smooth during the process of writing questions and saving and submitting. For this reason, he did not consider that it took a long time to access the SIAP application. Sometimes, their understanding of applications that are still relatively new or they are still unfamiliar with using applications makes them stutter technology (based on respondents' answers number 37, 59, and 154; all three are male and are in the same age group, namely 30 to 39 years old). It is undeniable that most people need a long time to understand or learn something new. The same is the case with the results of a research by Onwuegbuzie et al. (2004) conducted on undergraduate graduates in America.

Communication with the Person in Charge of the Subject Matter

SIAP application provides an opportunity for the writer of the questions to communicate with the PIC of the material through the Chat feature. However, some participants did not use the facility. Of the 295 respondents, 200 people (68%) answered 'Yes', 92 people (31%) answered 'No', and three people (1%) did not answer. Of the 92 respondents who answered 'No', 30 respondents continued to answer questions. Of the total 92 respondents who answered 'No', 30 respondents still answered further questions. Male respondents numbered 13 people (43%) and female respondents numbered 17 people (57%). It indicates that more females are not willing to communicate with the PIC of the subject matter.

This condition is in contrary to consistency in continuing to answer questions, namely, in this case, women can be said to be inconsistent, which can be caused by psychological factors (Hornung, 1977), such as stress due to communicating with the PIC of the subject matter that must be socially elevated. In addition to stress, personal health conditions or family can also become the trigger. For example, the condition that occurred in respondent number 264 (female, aged 40 to 49 years old) said that she could not focus because she was caring for his mother who was hospitalized. The same is true for the answer of respondent number 164 (female, aged 50 years old or more) who said that she accessed the SIAP application when she was not ready because she was ill.

It is possible that female respondents did communicate with the PIC of the subject matter, but they received a rejection (it could be in the form of neglecting chat, having to make improvements to the questions submitted, even more extreme: the rejection of the question). It seems that in this case, the stimulus factors and responses (Hovland, 1948) are the determinants of communication that can be established well or not. The reason for fear of being rejected for the questions they wrote seems to be justified when reading the respondent's answers in the question section on the experience of accessing the SIAP application. Respondent number 238, female, postgraduate education, with five years teaching experience (respondents included in the sample who answered follow-up questions) wrote that she was happy and worried about the questions she wrote and submitted. She was afraid that the problem would be rejected so she was very aware of writing questions and submitting them. Errors in running an application can also cause unauthorized written questions. These errors include mistakes in placing stimulus, choice of answers, and key answers and reasons.

Subconsciously, those female respondents accidentally choose the option 'No'. It is likely the inability of females to suppress expressive language tendencies (Maynard, 1988) which is more dominant than male language behavior. She means to answer 'Yes', but accidentally clicks the choice of 'No'.

Communication with the SIAP Administrator

The results of the studies have provided information that technological advances in the field of communication cause various changes in society, such as changes in the speed of information exchange (Zamroni, 2009). Administrators as parties that facilitate the questions writers with various information, such as, the time of the assignment, indicators that must be done, the number of questions that must be written, must be able to take advantage of technological advances in communicating. To support this, in addition to providing communication facilities with the PIC of the material directly, the SIAP application also provides communication facilities between the question writer and the SIAP administrator.

Of the 295 respondents who responded, 266 people (90.16%) used this facility. The number of male respondents was 112 people and female respondents amounted to 154 people, while the number of respondents who answered 'No' was four people and those who did not answer are 25 people. Of the four respondents who answered 'No', only three people participated in answering further questions. The interesting thing is from the three questions that have follow-up questions, specifically the communication questions with the admin getting the response that did not answer the most, namely 25 people. There were 13 female respondents with undergraduate education and five male students; four females with post-graduate education and three males with post-graduate education.

The biggest possibility of respondents who did not provide answers regarding communication with the SIAP administrator was respondents who had less memorable experiences. For example, the answer of respondent number 285 (female, 50 years old or more, post-graduate education) on the question of the criticism and suggestions that she requested that the SIAP administrator be faster and more responsive in answering the questions submitted by the participants. They need immediate answers because the time given by the system for writing questions is not too long, especially for those who communicate via WhatsApp private lines (respondent number 225, woman, undergraduate).

Another reason identified was the respondents' unpreparedness when they got a call from the SIAP administrator who tells them about assigning questions. Often the administrator suddenly contacts the question writer to be assigned several indicators that must be written for a short period of time (respondent's answers number 66, 198, 202, and 290). Upset? Most people will feel annoyed if they have just gotten information but have been demanded to be able to immediately implement with deadlines that are not so long.

As an administrator, it is appropriate to provide the best service to service recipients. Everyone who serves as an administrator must have high commitment and motivation in carrying out their duties (Moenir, 2010). If the administrator can provide optimal service and meet customer demands, the organization can be regarded as the greatest form of success on the service side. On the contrary, if the service provided is not satisfactory, it is necessary to improve the system and implementation mechanism.

Conclusion

The web-based questionnaire or digital questionnaire still has many disadvantages. The presence of data outliers, for example, cannot be avoided. There are only factors that cause respondents to give unexpected data, such as psychological factors when responding to questionnaires. Actually, it can be anticipated if the digital questionnaire is accompanied by an interview to get in-depth information about the aspects being evaluated. It is okay to use a web-based questionnaire, but in its implementation, it will be more targeted if the respondent is accompanied when inputting data on the media used (such as devices, mobile phones, tablet computers, etc.). It provides an opportunity for respondents to ask if there are questions that are difficult to understand. Interviewers can also supervise carefully, especially taking preventive actions related to conditional questions (in this paper stated in terms of follow-up questions).

Data in the form of outliers do not have to be deleted or treated by means of statistical averaging, but can be used as a data source to conduct in-depth interviews with respondents who gave the response. It can be a new field for constructivist paradigm researchers. They can explain outliers from another point of view. Even for psychometrics, outliers can be used as new data to see the relationships between the variables of non-parametric scale data found in the respondents, for example, to see if there is a relationship between the length of teaching time, gender, and age with communication with the subject matter PIC, and SIAP administrator.

In addition, it is suggested that in the formulation of questionnaires (both paper-based and web-based), it is necessary to consider the respondent's experience of novelty and current, such as responding in a new form (digital questionnaire), layered questions (questions that require further questions), because not all of the respondents were in a focus position to answer, even though they were highly educated, of productive age, and had a longer experience.

References

Ali, M. A. (1994). Identifying the location shift outliers that matter. *Sankhyā: The Indian Journal of Statistics, Series A*, *56*(3), 500–511. Retrieved from https://www.jstor.org/stable/25051015

Cheng, Y. H., & Tsai, C. C. (2017). Online research behaviors of engineering graduate students in Taiwan. *Journal of Educational Technology & Society*, *20*(1), 169–179. Retrieved from https://scholar.lib.ntnu.edu.tw/en/publications/online-research-behaviors-of-engineering-graduate-students-in-tai

Fíncher, C. (1981). AIR between forums: The literature of program evaluation. *Research in Higher Education*, *14*(3), 277–280. Retrieved from https://www.jstor.org/stable/40195416

Harris, H. (2001). Content analysis of secondary data: A study of courage in managerial decision making. *Journal of Business Ethics*, *34*, 191–208. https://doi.org/10.1023/A:1012534014727

Hastomo, F., & Yuhana, U. L. (2013). Perancangan dan pembuatan perangkat lunak aplikasi Android untuk pengolahan data transaksi pada perusahaan telekomunikasi "X" dengan menggunakan Pentaho. *Jurnal Teknik*

POMITS, *2*(1), 77–82. Retrieved from http://ejurnal.its.ac.id/index.php/teknik/article/view/2733

Hornung, C. A. (1977). Social status, status inconsistency and psychological stress. *American Sociological Review*, *42*(4), 623–638. https://doi.org/10.2307/2094560

Hovland, C. I. (1948). Social communication. *Proceedings of the American Philosophical Society*, *92*(5), 371–375. Retrieved from https://www.jstor.org/stable/3143048

Martel, A. R. (2015). The detection of outliers in nondestructive integrations with the Generalized Extreme Studentized Deviate test. *Publications of the Astronomical Society of the Pacific*, *127*, 258–265. Retrieved from https://iopscience.iop.org/article/10.1086/680382/meta

Maynard, D. W. (1988). Language, interaction, and social problems. *Social Problems*, *35*(4), 311–334. https://doi.org/10.2307/800590

McArthur, D. L., & Choppin, B. H. (1984). Computerized diagnostic testing. *Journal of Educational Measurement*, *21*(4), 391–397. https://doi.org/10.1111/j.1745-3984.1984.tb01042.x

McQuarrie, A. D., & Tsai, C.-L. (2003). Outlier detections in autoregressive models. *Journal of Computational and Graphical Statistics*, *12*(2), 450–471. Retrieved from https://www.jstor.org/stable/1391204

Moenir, A. S. (2010). *Manajemen pelayanan umum di Indonesia*. Jakarta: Bumi Aksara.

Mulyana, A., & Sholekan, S. (2010). *Aplikasi mini market online berbasis web*. Bandung: Universitas Telkom.

Nezami, E., & Butcher, J. N. (2000). Objective personality assessment. In G. Goldstein & M. Hersen (Eds.), *Handbook of Psychological Assessment* (3rd ed., pp. 413–435). https://doi.org/10.1016/B978-008043645-6/50094-X

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314. https://doi.org/10.1080/02602930701293231

Onwuegbuzie, A. J., Mayes, E., Arthur, L., Johnson, J., Robinson, V., Ashe, S., … Collins, K. M. T. (2004). Reading comprehension among African American graduate students. *The Journal of Negro Education*, *73*(4), 443–457. https://doi.org/10.2307/4129628

Pangesti, B. N. A. (2017). Analisa kecepatan transfer data pada perancangan hotspot sederhana dengan system single sign on di perkantoran. *Fountain of Informatics Journal*, *2*(1), 1–7. https://doi.org/10.21111/fij.v2i1.814

Parquette, W. S. (1952). Intensive reading. *The English Journal*, *41*(2), 78–82. https://doi.org/10.2307/809208

Prescott, P. (1978). Examination of the behaviour of tests for outliers when more than one outlier is present. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *27*(1), 10–25. Retrieved from https://www.jstor.org/stable/2346221

Qian, Y., Buchmann, C., & Zhang, Z. (2018). Gender differences in educational adaptation of immigrant-origin youth in the United States. *Demographic Research*, *38*, 1155–1188. https://doi.org/10.4054/DemRes.2018.38.39

Rapallo, F. (2012). Outliers and patterns of outliers in contingency tables with algebraic statistics. *Scandinavian Journal of Statistics*, *39*(4), 784–797. https://doi.org/10.1111/j.1467-9469.2012.00790.x

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics / Metodološki Zvezki*, *19*, 159–177.

Ross, L., & Cronbach, L. J. (1976). Handbook of evaluation research. *Educational Researcher*, *5*(10), 9–19. https://doi.org/10.3102/0013189X005010009

Sevillano-García, M. L., & Vázquez-Cano, E. (2015). The impact of digital mobile devices in higher education. *Educational Technology & Society*, *18*(1), 106–118.

Shubhi, M. M. F., Yuliana, M., & Winarno, I. (2011). *Sistem monitoring jaringan menggunakan BREW (Binary Runtime Environment for Wireless)*. Retrieved from http://repo.pens.ac.id/1077/1/Paper_Sidang_TA.pdf

SIAP Application Development Questionnaire. (n.d.). *No Title*. Retrieved from https://docs.google.com/forms/d/e/1FAIpQLSemVx0QSQlom5UecJs55aVE2N5kHMcRJSQZTdXHp4EGcT-N-A/viewform

Yanxia, Y. (2017). Test anxiety analysis of Chinese college students in computer-based spoken English test. *Journal of Educational Technology & Society*, *20*(2), 63–73. Retrieved from https://www.jstor.org/stable/90002164

Zamroni, M. (2009). Perkembangan teknologi komunikasi dan dampaknya terhadap kehidupan. *Jurnal Dakwah: Media Komunikasi Dan Dakwah*, *10*(2), 195–211. https://doi.org/10.14421/jd.2009.10205

REiD (Research and Evaluation in Education), 5(2), 2019, 152-168
Available online at: http://journal.uny.ac.id/index.php/reid

The effectiveness of Game-Based Science Learning (GBSL) to improve students' academic achievement: A meta-analysis of current research from 2010 to 2017

***1Heru Setiawan; 2Shane Phillipson**
[1]Sekolah Menengah Atas Global Mandiri Jakarta
Jl. Raya Cakung Cilincing KM. 5, Cakung Timur, Jakarta Timur, DKI Jakarta 13910, Indonesia
[2]Faculty of Education, Monash University, Australia
29 Ancora Imparo Way, Clayton VIC 3800, Australia
*Corresponding Author. E-mail: heru.setiawan@teacher.globalmandiri.sch.id

Submitted: 10 November 2019 | Revised: 6 December 2019 | Accepted: 12 December 2019

Abstract

This study identifies the effectiveness of game-based science learning (GBSL) for improving students' learning outcomes by conducting a literature review of the current research from 2010 to 2017. This study also explores the correlation between variation in school level and year of publication on GBSL effect size. Data were collected from peer-reviewed journal articles published in educational databases including ERIC (Educational Research Information Centre), Springer Link, ProQuest education journal, and A+ education. Seven inclusion criteria were used to select relevant studies. Comprehensive Meta-Analysis (CMA 2.0) was used to analyze the data. This study finds that (1) GBSL intervention has a statistically significant effect on students' learning outcomes with a higher average on the effect size of the experimental group (41.12) than the control group (37.07). The mean of the reviewed studies' effect size is 0.667 in the medium category. (2) The implementation of GBSL in secondary school has a bigger average effect size than in elementary school. Year of publication and effect size has a low positive correlation with a coefficient of correlation 0.40.

Keywords: *game-based science learning, learning outcomes, meta-analysis*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.28073

Introduction

The young generation who was born in the 21st century is a digital native or the Net generation (Bennett, Maton, & Kervin, 2008). The millennial in this era also can be called a game generation (Prensky, 2001). The trend of digital games' use has been increasing in this era (Corbett, 2010; McGonigal, 2011). Millions of people have been immersed in playing digital games either for entertainment or education (Huang, Hew, & Lo, 2019). Gee (2007) reported in his study that approximately 90% of students' mobile phones connect to digital games. Besides, many teachers use digital games as a medium of instruction in their classroom for engaging students during teaching and learning processes, or it is commonly called digital game-based learning (DGBL) (Papastergiou, 2009; van Eck, 2006). Students also obtain feedbacks such as improvement, and win conditions after completing the goals (Okeke, 2016, p. 1). The DGBL that specifically focus on Science is called Game-Based Science Learning (GBSL).

Since 2006, the number of research investigating the effect of Digital games in education has been increasing (Chorney, 2012). Some literature has been debating the effectiveness of GBSL in the last decade (Hamari & Keronen, 2017; Quandt et al., 2015). The community of science education (physics, biology, chemistry, and general sciences) also concern with the potential of game-based learning. Some researchers investigate the effectiveness of GBSL in some science subject matter such as Newtonian mechanics (Clark et al., 2011), human immunology (Cheng, Su, Huang, & Chen, 2014), and photosynthesis (Culp, Martin, Clements, & Lewis Presser, 2015). They argue that science is challenging for some students because of abstract concepts and invisible objects. In addition, some research illustrated that rote memorization and decontextualized learning have potential drawbacks in the Science context (Honey & Hilton, 2011; Mayo, 2007). This issue has an impact on their learning outcomes which can be defined as skills, knowledge, and values as an outcome of students' experiences (The US Council for Higher Education Accreditation (CHEA) cited in Adam (2004, p. 4). Learning outcomes can be knowledge, skills, or attitude. However, in this context, the learning outcome only refers to students' learning outcomes in academic settings. Thus, GBSL is the proper solution to this issue because digital games are highly engaging and motivating (Huang et al., 2019; Tsay, Kofinas, & Luo, 2018). Several researchers demonstrated empirical evidence of the potential of this educational tool to enhance students' learning outcomes in the various context of science subjects through comparing control and experiment group (such as Bello, Ibi, & Bukar, 2016; Fan, Xiao, & Su, 2015).

However, a small number of sample of studies investigating the effect of GBSL on students' learning outcomes tended to have a more significant mean of effect sizes than studies with larger sample sizes (Cheung & Slavin, 2013). Effect size refers to a quantitative measurement of the difference between the mean score of the control group and the treatment group (Nakagawa & Cuthill, 2007). Meanwhile, the small sample size of the research cannot be used to generalize the effect of GBSL. In order to solve this issue, it needs further investigation of the effectiveness of GBSL in students' achievement in sciences with a meta-analysis study to develop a better estimate of effect magnitude (King & He, 2005). Meta-analysis is the process of converting the effects of several similar research into quantitative data so that these averages of the effect size and an overall determination can be made concerning the cumulative findings of several studies (Glass, McGaw, & Smith, 1981). Meta-analysis is a kind of retrospective observational study in which researchers make data recapitulation without any experimental manipulation (Brockwell & Gordon, 2001).

Several literature reviews of Game-Based Learning have been conducted both in the context of sciences and other subjects such as mathematics, language, history, and physical education. In 2006, Vogel et al. (2006) used meta-analysis of digital games versus traditional teaching methods. The overall result of the meta-analysis was that treatment groups were reported higher learning outcomes and better attitudes toward learning than control groups. The report also analyses some moderator categories. He reported that gender, school level, and user type showed significant statistical results. Meanwhile, learner control, type of activity, and realism do not appear to be influential. In the science context, Li and Tsai (2013), reviewed research articles regarding game-based science learning (GBSL) published from 2000 to 2011. The focus of the review is qualitative outcomes including research purposes and designs, the theoretical foundations, game design, and learning focus. Based on the review, GBSL can provide effective learning in a collaborative problem-solving environment. However, the research only focused on qualitative data without discussing and analyzing the quantitative analysis of GBSL intervention and the effect size.

According to the previous research, gaps in the literature have been identified. Although several studies have explored a review of literature of GBSL, few have tested their relative influence on learning the outcome. There was also a lack of research meta-analysis of GBSL with a quantitative approach. Li and Tsai (2013) who focused their research on the qualitative method suggested that quantitative content analysis of GBSL effectiveness such as students' learning outcomes in Science education should be conducted in future investigations. It is because digital games that can promote students' engagement (Annetta, Minogue, Holmes, & Cheng, 2009; Tsay et al., 2018) might also enhance students' learning outcomes. Other similar studies such as Vogel et al. (2006) also have a limitation. Although he specifically focusses on cognitive aspects in the analysis, the context of the study is in a broad context and did not specifically focus on Science education. Based on this gap, a newly proposed work focusing on a meta-analysis of the effect of the digital game on students' learning outcome in Science education or GBSL need to be conducted. Thus, two central research questions (RQs) were addressed in this study: (1) RQ1: Do Game-Based Science Learning (GBSL) effective to enhance students' learning outcomes compared to traditional method as reported by the current studies from 2010 to 2017? (2) RQ2: Do moderator categories including school level of participants (elementary and secondary school context) and year of publication has any correlation with GBSL effect size?

This research contributes to the literature in this field. First, this study reviewed recent trends in GBSL research, especially for those in the field of science education who are interested in quantitative studies of GBSL for students' learning outcomes. A meta-analysis of GBSL has been conducted by several researchers within a broader context such as mathematics, language, and other subjects (Divjak & Tomić, 2011; Young et al., 2012), but lack of research conducted in science education. Second, the consistency of the result of similar studies for several years will be investigated. Therefore, consistency and inconsistency of findings of similar research will be found, and bias of one or more studies in this field could be detected (Borg & Gall, 1983). Third, a meta-analysis uses a significant amount of data, and applying statistical methods by organizing some information comes from a broad cross-section whose function is to complement other purposes (Glass et al., 1981). By the significant number of participants, the study develops a better estimate of effect magnitude (King & He, 2005). The larger sample size in conducting a meta-analysis could be found in one study that will create greater statistical power and more precise confidence intervals. This is because the study collects several similar studies to be analyzed quantitatively. It concentrates on the effect size of this empirical discovery which is relatively better than the other methods of quantitative approaches including narrative review, descriptive review, and vote counting (Lipsey & Wilson, 2001). Moreover, through the substantial number of participants with different variables, the differences may exist because of differences that exist among the articles such as different subject populations, education level, gender, game type, etc. By using meta-analysis, different moderator variables can be investigated. Vogel et al. (2006) state that analyzing moderator variables would give a clearer overview or more complex picture of reviewed studies.

Method

Research Strategies and Data Collection

The search of the literature was conducted from June to July 2017. Data were collected from journal articles published from educational sources including ProQuest education journal, Springer Link, A+ education, and ERIC (Educational Research Information Centre). The databases provide a high impact and a high-quality journal article. The keywords are 'digital game, sciences, physics, biology, chemistry, secondary, high school, elementary.' The Boolean operator, 'AND' or 'OR', was used to combine all key terms. Following the keywords, the researchers read the abstract and full-text. We use some inclusion and exclusion criteria as the evaluation to choose appropriate journal articles. Seven inclusion and exclusion criteria were applied in screening the eligible article included in this study including publication year, unit, participant, game intervention, research design, participant, outcome type, and language. These details of inclusions and exclusions are explained as follows.

(1) Publication year: All of the articles are peer-reviewed journal articles published in the last seven years from January 2010 to June 2017. (2) Unit: The unit in elementary and secondary education in this study is science

subjects including biology, physics, chemistry, and general sciences. Other units such as technical subjects in vocational high school are excluded. Also, unrelated subject matters that have similar keywords, but they are not related to science subjects such as physical education are excluded. (3) Game/ intervention: Digital games in this study is defined as a digital experience where participant use game of computer software and they receive feedback to achieve the goals in the form of a score, progress and win condition. However, learning intervention that focused on creating a digital game for students is not included. The studies compared digital games in science instruction and traditional methods. (4) Research design: All of the journal articles included in this meta-analysis must use experimental and control groups or game versus non-game conditions. The studies must have a sample size, standard deviation, and mean. However, studies that do not have the data were excluded. The studies included used an experimental method to make sure that the included studies have data compared in the statistical analysis. Studies are considered experimental if individual students are randomly assigned to an instructional condition. (5) Participant: The participants of the research in the included studies are elementary and secondary school students. Students with specific clinical criteria such as disabilities are excluded from this study. (6) Outcome type: The data that will be extracted in this study is only quantitative data (numerical data) specifically students' learning outcome or cognitive aspect. Other research outcomes or qualitative data such as behavior, activity, participation, collaboration, engagement, and motivation are not extracted. (7) Language: The study included is an only article published in English without considering the country in which the studies are conducted.

The full text that is related to the inclusion criteria of the topic was evaluated by annotating each article to extract some necessary information. This step was conducted using note-card contained eligibility criteria evaluation rubric recommended by Mertens (2015) including research question, the design of research, data analysis, results, conclusion, and research evaluation. During the preliminary selection of eligibility occurred in 137 articles were identified. Then, after the articles were screened for eligibility to exclude some non-eligible full text by applying inclusion criteria, 12 journal articles are carefully selected although this amount is a small number relative to some meta-analyses in this field.

The data from the selected studies is then extracted for further analysis. First, the data of the characteristics of the reviewed studies that include the year of publication, country of origin, school level of participants, science domain, game name, and the purpose of the study were noted in Microsoft Excel. The data were extracted through manual searches in each article. The data is important to provide an overview of the characteristics of the reviewed studies. Second, the key information which corresponds to the research questions were also extracted for each study. The information which is needed to answer the research questions is only quantitative data (numerical data) that was used in the statistical analysis. The quantitative data extracted are student's achievement means, standard deviation, the number of participants of the control and treatment group.

Data Analysis Method

Microsoft Excel and Comprehensive Meta-Analysis (CMA 2.0) were used for statistical analysis after the quantitative data were extracted. Formerly, the demographic characteristics of the reviewed studies were analyzed with descriptive statistics using Microsoft Excel which present data such as mean, percentage, and also frequencies. The data would also be presented with visual techniques such as a column, bar chart, and histogram. Lately, CMA 2.0 was also used. Several researchers verified the accuracy of the analysis method (Ones, Viswesvaran, & Schmidt, 1993). CMA 2.0 is used to analyze Hedges' g effect size, the lower limit (LL), the upper limit (UL), p-Value, and the Relative weight of all studies (Borenstein, Hedges, Higgins, & Rothstein, 2005). In order to give a clearer overview of the overall effect size, the forest plot to compare the effect of digital games over traditional methods was used (Sutton, Abrams, Jones, Sheldon, & Song, 2000). Two kinds of effect models in a meta-analysis are fixed effect model and random effect model (Michael Borenstein, Hedges, Higgins, & Rothstein, 2010). The decision to select the effect model to analyze data is an essential factor in the meta-analysis (Hedges & Vevea, 1998). Improper determination of the model will cause inefficient estimation and incorrect conclusions (Nickell, 1981). However, in this study, we use the random effect model because all twelve studies which are used in this research were drawn from different populations, such as different populations in different countries. A similar condition of research is conducted by Sacks, Berrier, Reitman, Ancona-Berk, and Chalmers (1998). Moreover, the studies report varies the effect size (ES). In the random-effects model, the true effect size might differ from one study to another study (Olejnik & Algina, 2000). In addition to the estimation of the primary effect, secondary analyses were conducted to take advantage of the coded study characteristics and test the moderating effects. Specifically, secondary analysis tested the influence of grade level (elementary and secondary school) and year of publication. The data from statistical analysis from CMA 2.0 were used in order to address the research questions with the following method of interpretation.

We address the first research question by comparing the experimental group and the control group. There would be no difference between the control and experimental group when the mean of the sample is equal. However, when the experimental group's means score is higher than the control group, it means that GBSL intervention is more effective through looking at the mean difference between the experiment and the control group. The second research question is answered by investigating the effect of moderator categories including the year and school level, to the GBSL effectiveness, we use descriptive analysis by comparing the mean of effect size in each category. We compare the average effect size at each school level (elementary and secondary school) to determine which school level more effective in the game intervention. Then, to analyze whether or not publication year has any correlation with game effectiveness, we use inferential statistics because it strives to make inferences and predictions (Bryman, 2016). The statistical method would improve the previous research that only looks at the pattern of effect size across the years. The data would be presented as scatterplot to illustrate the relationship between two variables (Cohen, Manion, & Morrison, 2007, p. 507). It would also count the Spearman's rank correlation coefficient (r) because both variables are ordinal to see the linear trend using Microsoft Excel. The interpretation to assess the degree of the correlation coefficient were categorized into very high (0.9 to 1.0), high (0.7 to 0.9), moderate (0.5 to 0.3), low (0.3 to 0.5), and negligible correlation (0 to 0.3).

Detection of Publication Bias

Detection of publication bias of reviewed studies is crucial in meta-analysis study (Rothstein, Sutton, & Borenstein, 2006). Publication bias is the tendency of researchers to screen articles for publication based on the statistical significance of effects than the quality of the study (Rothstein et al., 2006, p. 296). Several pieces of evidence show that some research that has a higher effect size is more likely to be published (Peters, Sutton, Jones, Abrams, & Rushton, 2006). Consequently, it will affect the review process. Therefore, the meta-analysis may be overestimated effect size because it uses a biased sample or target of the population. Hence, to avoid this concern or minimizing this bias in this study, it needed a model to know which study is missing. One of the proper models is the funnel plot (J. A. C. Sterne et al., 2011). In the funnel plot, the effect size is plotted in X-axis, and the number of participants is plotted in Y-axis (Sterne & Egger, 2001). Also, asymmetry easily detected in the funnel plot. The studies will be distributed symmetrically when the publication bias is absent (Schmidt & Hunter, 2014). The next problem is whether the observed overall effect is robust. To solve this issue, some researchers use Rosenthal's Fail-safe N. Orwin (1983) suggested that Rosenthal's Fail-safe N compute the number of studies that should be incorporated in the analysis.

Findings and Discussion

Overview of the Reviewed Studies

The publication years range from 2010 to 2017. The purpose is to know the development of research in this area in the last eight years. The highest number of publications is in 2015 with three publications (Figure 1). Then, the presence of international studies is reflected in the sample. However, 50% of the studies included were conducted within the Asia continent especially in Taiwan, while the others were conducted internationally. There are two countries including Taiwan and Singapore from Asia. Within this international group, Spain is well represented by two studies, while the other research is from the U.S and Nigeria, Africa (Figure 2). Based on the school level, elementary and secondary education has an almost equal number. Eight studies are from elementary school and four studies from high school (Figure 3). Subject areas are also well represented with three in the context of biology, seven general sciences, while each physics and chemistry are only one study (Figure 4).

The studies included are presented in Table 1. Table 1 outlines the characteristics of the included studies meeting all the eligibility criteria.

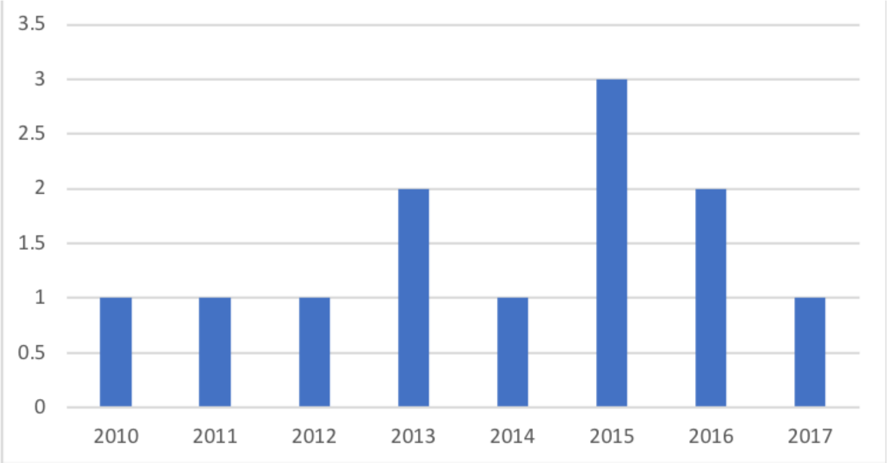


Figure 1. The number of reviewed studies by year of publication

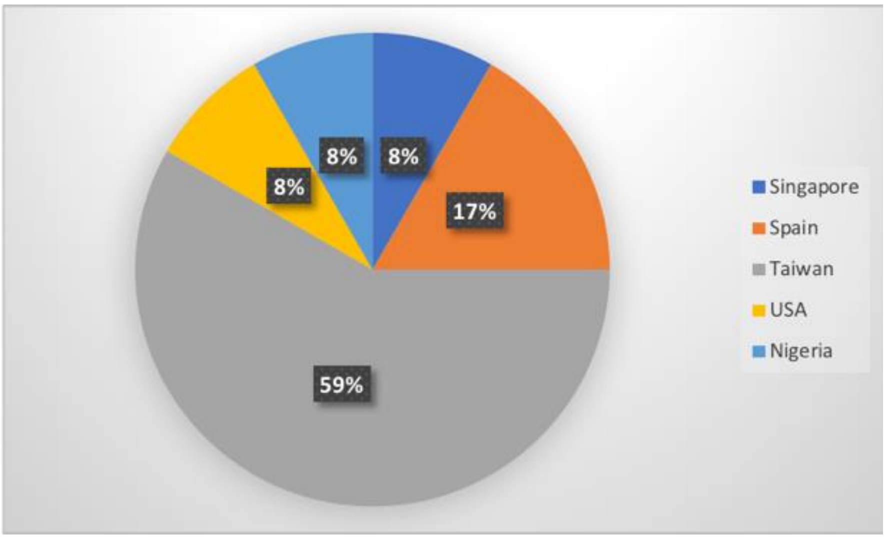


Figure 2. The reviewed studies by country

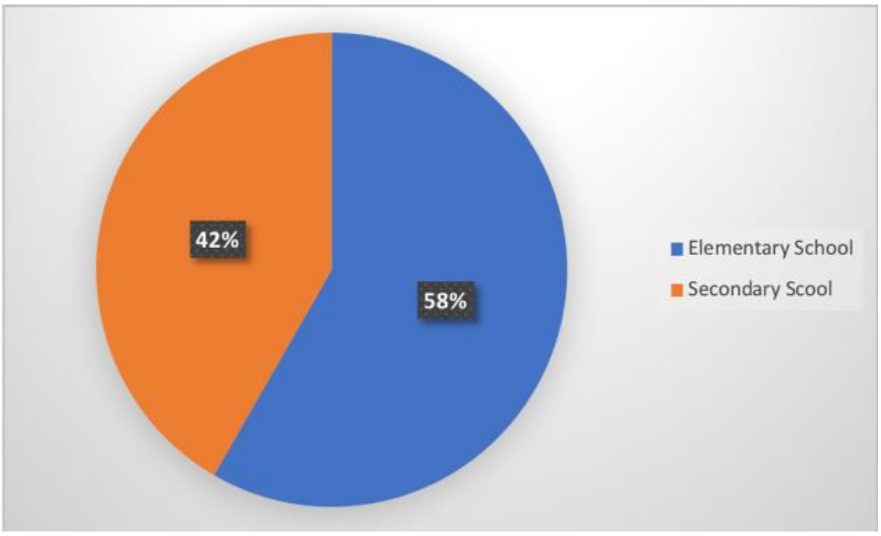


Figure 3. The reviewed studies based on the education level of participants

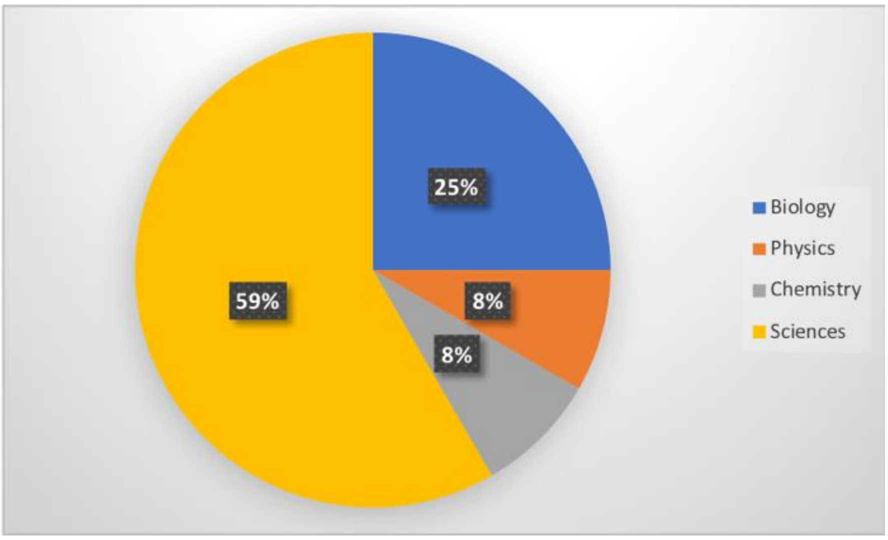


Figure 4. The reviewed studies according to science domain

Table 1. The background information of reviewed articles

Authors	Country	Game Name	School Level	Science Domain
Bello et al. (2016)	Nigeria	n/a	Secondary	Sciences
Chee & Tan (2012)	Singapore	Alkhimia	Secondary	Chemistry
Wrzesien & Raya (2010)	Spain	Supercharged	Elementary	Sciences
Anderson & Barnett (2013)	USA	Supercharged	Secondary	Physics
Sung & Hwang (2013)	Taiwan	Alien Invasion	Elementary	Sciences
Yien, Hung, Hwang, & Lin (2011)	Taiwan	Nutrition Supplement Battle	Elementary	Biology
Chu & Hung (2015)	Taiwan	Kodu	Elementary	Sciences
Su & Cheng (2015)	Taiwan	Find Insect	Elementary	Sciences
Chen & Hwang (2017)	Taiwan	Alien Invasion	Elementary	Sciences
Fan et al. (2015)	Taiwan	The MMBCLS	Secondary	Biology
Furió, Juan, Seguí, & Vivó (2015)	Spain	iPhone game	Elementary	Sciences
Chen, Yeh, & Chang (2016)	Taiwan	Role Play Game (RPG)	Secondary	Biology

How Effective GBSL Does to Enhance Students' Learning Outcomes in Sciences Compared to the Traditional Method as Reported by the Current Studies from 2010 to 2017?

The first research question is answered by comparing the average Mean of the reviewed studies. The result of data extraction is presented in Table 1 which compares the twelve studies with the treatment group and control group. The number of participants in the twelve studies is 954 students. Most of the studies have an equal number of participants in the treatment and control group, although some of them have a slightly higher participant in one group than the other group. There are 489 students in a total of the control group and 465 students from the experimental group. The number of participants in the studies is varied from 38 to 180 students. The standard deviation of all of the studies is also varied from the lowest 0.93 to the highest 23.54. The detail of the data for each study is shown in Table 2.

Based on Table 2, the average learning outcome mean from the overall studies of the experimental group (40.82) is higher than the control group (36.82). The mean difference analysis shows that one study, Chu and Hung (2015), has a negative mean difference between experimental and control group compared to the other ten studies that have a positive mean difference. The highest mean difference between the studies is 19.63, while the lowest mean difference is -15.03. The experimental and control group's standard deviation shows a variation.

The Analysis Result of Standardized Mean Difference Effect Size, Variance, Weight, and Confidence Interval (CI)

The random-effects model was used to know the composite effect size with Comprehensive Meta-Analysis (CMA). The summary of the final analysis for all studies is presented in Table 3. We calculate Hedges's g for each study separately to maintain consistency of measurement. In addition to the individual effects, we also present a 95% confidence interval (lower limit and upper limit) around each study and the relative weight (W). The overall effect size of the twenty studies is g = 0.661, p<.001; with a 95% confidence interval between 0.223 and 1.090. It indicates a moderate overall effect for the synthesized GBSL interventions that is statistically different from a null effect. The largest effect size influencing this study is Bello et al. (2016) of 2.338. In contrast, the study contributing the smallest overall influence is Chu and Hung (2015) with an effect size of -0.637. The comparison of the SMD effect size of all studies is presented in a forest plot in Figure 5.

Table 2. Mean, standard deviation, and sample size of the studies on digital games versus control method

Authors (year)	**Experiment Class**			**Control Class**			**N Total**	**Mean Difference**
	Mean	**SD**	**N**	**Mean**	**SD**	**N**		
Bello et al. (2016)	66.23	7.07	90	46.6	9.48	90	180	19.63
Chee & Tan (2012)	3.28	2.61	40	2	1.71	38	78	1.28
Wrzesien & Raya (2010)	6.33	2.2	24	5.88	1.54	24	48	0.45
Anderson & Barnett (2013)	6.3	1.2	32	5.9	1.27	32	136	0.4
Sung & Hwang (2013)	57.26	16.87	31	43.07	14.24	31	62	14.19
Yien et al. (2011)	16.94	2.38	33	15.09	3.39	33	66	1.85
Chu & Hung (2015)	56	23.54	30	71.03	23.04	29	59	-15.03
Su & Cheng (2015)	82.94	10	34	75.59	9.595	34	68	7.35
Chen & Hwang (2017)	86.78	9.15	27	82.35	12.38	26	53	4.43
Fan et al. (2015)	88	9	23	76	12	23	46	12
Furió et al. (2015)	4.89	1.45	19	4.74	0.93	19	38	0.15
Chen et al. (2016)	18.51	2.71	43	16.63	4	77	120	1.88
	41.12		Σ= 426	37.07		Σ= 456	Σ= 954	

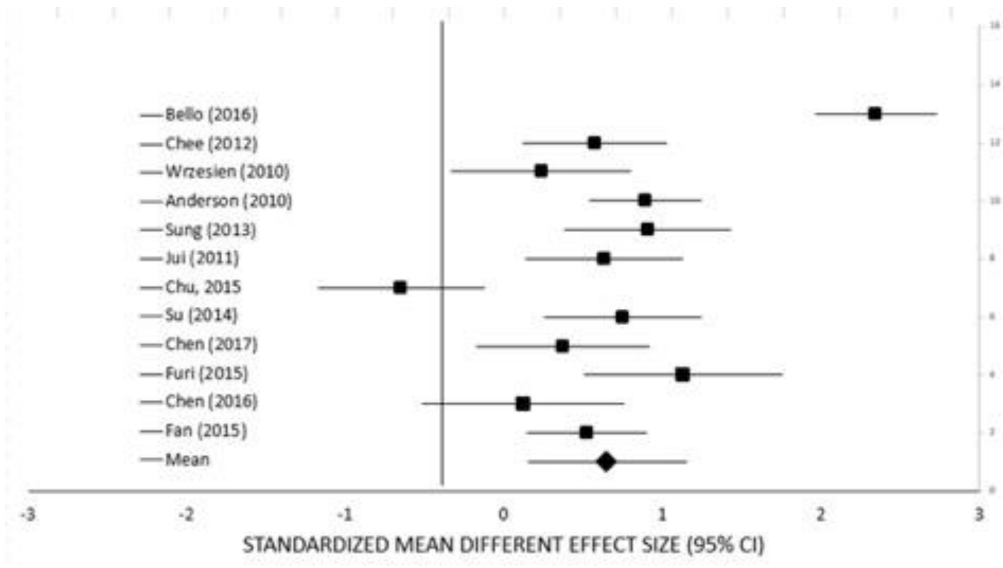


Figure 5. The forest plot's comparison of the SMD effect size of reviewed studies

Table 3. Effect sizes, confidence intervals, and relative weights of reviewed studies

Name of the Study	Hedges's g	Lower Limit	Upper Limit	p-Value	Relative Weight (W)
Bello et al. (2016)	2.338	1.959	2.716	0.00000	8.738
Chee & Tan (2012)	0.571	0.123	1.020	0.01255	8.502
Wrzesien & Raya (2010)	0.233	-0.325	0.792	0.41332	8.089
Anderson & Barnett (2013)	0.886	0.536	1.237	0.00000	8.821
Sung & Hwang (2013)	0.898	0.381	1.414	0.00065	8.253
Yien et al. (2011)	0.624	0.136	1.113	0.01227	8.358
Chu & Hung (2015)	-0.637	-1.153	-0.120	0.01571	8.252
Su & Cheng (2015)	0.741	0.255	1.228	0.00279	8.367
Chen & Hwang (2017)	0.402	-0.134	0.938	0.14151	8.177
Fan et al. (2015)	1.112	0.500	1.724	0.00036	7.873
Furió et al. (2015)	0.121	-0.503	0.744	0.70453	7.826
Chen et al. (2016)	0.520	0.143	0.896	0.00682	8.743
Randon effect model	0.661	0.232	1.090	0.00253	

Table 4. Mean effect size of GBSL based on school level

Moderator	Number of studies	% of study	d	N
Elementary	7	58.33%	1.08	394
Secondary	5	41.67%	0.34	560

Do Moderator Categories Including School Level of Participants (Elementary and Secondary School Context) and Year of Publication Have Any Correlation with GBSL Effect Size?

Based on our analysis of moderating variables as the addition to the overall effect size, subsequent analyses of some moderating variables were run by school level and year of journal article's publication, shown in Table 4.

Firstly, we made two comparisons from the school level including elementary and secondary schools (Table 4). Seven studies are in the context of an elementary school setting with the mean of effect size 1.08. The other five studies tested on secondary school setting with an effect size mean of 0.34. This number shows that the effect size of GBSL on secondary school contexts nearly two and a half times higher than elementary school students sample effect size. Thus, the implementation of GBSL in secondary school tend to have a larger effect size than in elementary school context.

Secondly, we made a comparison of effect size according to the year of publication (Table 5). According to the correlational analysis between the year of publication and effect size, it shows that the variable has a low correlation with the r= 0.40 (r2= 0.16). Figure 6 illustrates a scatter plot that shows the relationship between year of publication (X-axis) and effect size (y-axis). Figure 6 shows that from 2010 the effect size average is 0.23, followed by approximately double to 0.55 in 2011. Five years later, in 2016, the effect size significantly increased again to 2.54.

Analysis for Publication Bias

According to the analysis of Rosenthal's Fail-safe N (Orwin, 1983), among the various methods for assessing bias, Rosenthal's Fail-safe N has the advantage of focusing on the potential impact any unpublished or unidentified studies may have on the current estimated effect size. It provides an estimate for the number of hypothetical missing studies that must be identified in order to bring the calculated overall effect below the level of researcher-imposed substantive significance (Easterbrook, Gopalan, Berlin, & Matthews, 1991). It assumes that those missing studies have negligible effects. Based on the analysis, 307 more studies are needed to make p-value to be alpha (Z for alpha= 1.959). The other method to analyze publication bias is using the Funnel Plot, which has two diagonal lines that represent the 95% confidence interval, and a vertical central line. The x-axis represents the study sample size, and the y-axis represents the effect size. Figure 7 illustrates the Funnel plot of Standard Error (SE) by Hedges' g effect size.

According to Figure 7, the nine studies fall around the two horizontal lines or a confidence interval of 95%. However, three studies fall outside the funnel plot, indicating that these studies were not as significant as the other nine studies.

Table 5. Mean effect size of GBSL based on year of publication

Year of Publication	Number of Studies	% of study	d	N
2010	1	8.33%	0.23	48
2011	1	8.33%	0.55	66
2012	1	8.33%	0.75	78
2013	2	16.67%	0.892	198
2014	1	8.33%	0.77	68
2015	3	25.00%	0.51	143
2016	2	16.67%	2.54	300
2017	1	8.33%	0.36	53

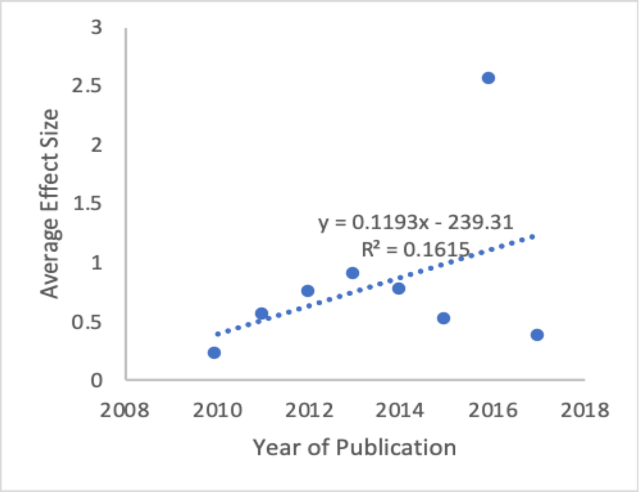


Figure 6. Scatter plot of the relationship between the year of publication and average effect size

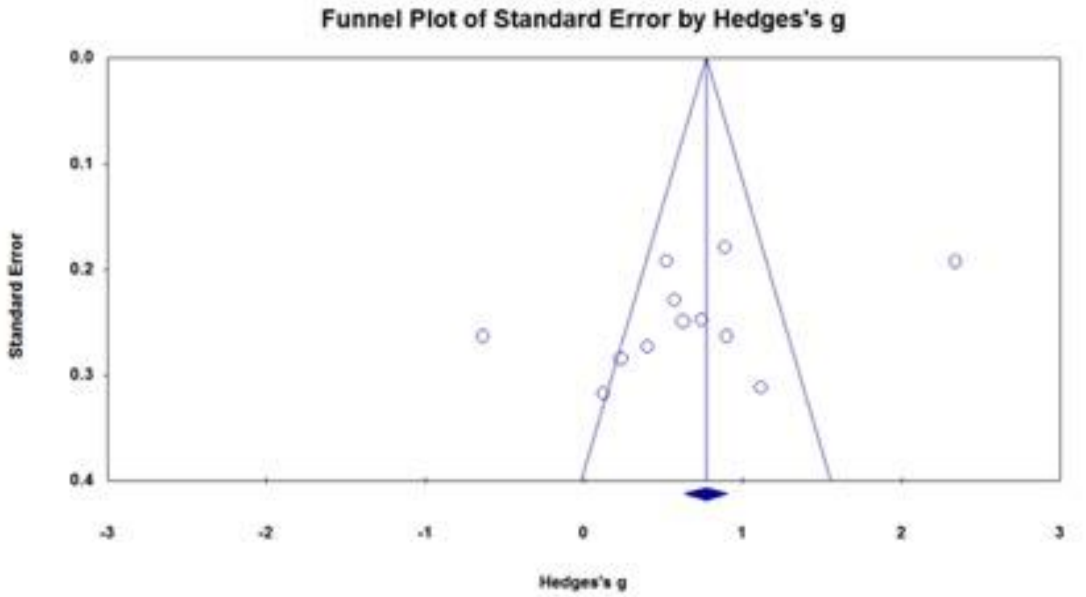


Figure 7. Funnel plot of standard error (SE) by hedges's g effect size of reviewed studies

The Performance of the Result of This Study with Similar Research

The performance of this study aligns with similar studies of literature reviews using meta-analysis on gamification across various context, such as mathematics, language, and also physical education over a decade, which has consistently found that game-based learning outperforms traditional-based learning (Divjak & Tomić, 2011; Vogel et al., 2006; Young et al., 2012). However, some notable differences regarding the statistical analysis are revealed. First, the fail-safe number (*Nfs*) that we found in this research, that is 307 studies, is much lower than the previous meta-analysis. The fail-safe number is only approximately a fifth than the findings of Vogel et al. (2006) with *Nfs 1*465. Second, the number of studies in this meta-analysis is only twelve, which is lower than similar research in this field, such as Divjak and Tomić (2011) with 32 studies, and Young et al. (2012) with more than 300 articles. In addition, the findings of this research support the findings of Li and Tsai (2013) regarding the potential of GBSL to promote students' learning. Li and Tsai (2013) believe that GBSL can promote students' engagement. Therefore, students' engagement and motivation might lead to an improvement in students' learning outcomes in Science.

Conclusion and Recommendation

Conclusion

Based on the result and discussion, some conclusions can be drawn. First, based on the investigated studies conducted from 2010 to 2017, the use of GBSL is statistically significant to improve students' learning outcomes in elementary and secondary school. The learning outcome of the experimental group of the overall studies is higher than the control group, which is 41.12 against 37.07 respectively. The mean of Hedges' g random effect size of the reviewed studies is 0.667, which can be classified into a medium effect size. Second, moderator categories or variation of school level of the study have any correlation on digital game effectiveness on which the implementation of GBSL in secondary school have a greater effect size than in elementary school context. Also, the year of publication and effect size has a low positive correlation with r= 0.40.

Recommendation

The result of this study has implications for future studies. Experimental research of GBSL in Science education across various contexts is still needed. It is supported by the result of detection publication bias which showed that at least 237 studies in this area of research are needed that would bring p-value to be alpha. This research is complex, but the description of the process and result has been presented. Furthermore, we use Comprehensive Meta-Analysis 2.0 as trusted software for quantitative meta-analysis.

However, our study has some limitations. The study only includes a small amount of research. It might be caused by the topic used is too specific where it only includes the effect of GBSL in a subject (Science) and the outcomes only specifically focus on cognitive aspects. There are many potential studies in GBSL in Science education and in the timeframe (2010-2017), but they were not included in this study because they were not eligible in the screening process with the seven inclusion and exclusion criteria which is determined in the research design. Some researches have no complete data to be extracted, or the topic is not suitable for this research. For example, the research use case study which only has an experimental group does not have a control group (Echeverría et al., 2011; Spires, Rowe, Mott, & Lester, 2011). Other studies are not eligible because they focus on other outcomes such as engagement (Annetta et al., 2009), collaboration and problem-solving (Sánchez & Olivares, 2011), and developing serious games (Khalili, Sheridan, Williams, Clark, & Stegman, 2011; Nilsson & Jakobsson, 2011; Ting, 2010).

Therefore, future studies should not only focus on the cognitive or quantitative outcome but also affective or qualitative outcomes such as students' engagement, motivation, self-efficacy, participation, collaboration, communication, and problem-solving skills. The research to review the qualitative outcome can be conducted with a systematic review, narrative review, or descriptive review (For example, Kim, Munson, & McKay, 2012; Li & Tsai, 2013).

The limited number of research identified might also due to the restricted criteria of the year of publication, sources of databases, context, and moderator categories. First, the included studies were conducted from 2010 to 2017. Therefore, the result of this study does not capture the studies outside this period. Second, the review only includes some databases, including ERIC, Springer Link, ProQuest, and A+ Education. Future studies can also be conducted by extending the literature to other educational databases such as ISI Web of Sciences or sources like Google Scholar, conference proceedings, and dissertations. There many articles related to GBSL.

Third, regarding context, investigating the effectiveness in different contexts/country and expanded educational level such as preschool could also be explored in future studies. It is because we found that most of the research included in this meta-analysis was conducted within Asia and educational level in the preschool context has not been explored. The last, for moderator categories, our research only focused on the school level of participants and year of publication of the study. Therefore, future research can explore different moderators such as gender (Tsay et al., 2018; Vogel et al., 2006), game genre (individual, peers, or groups), stream type or typical games (Sjöblom, Törhönen, Hamari, & Macey, 2017), learner control, and type of activity (Vogel et al., 2006).

Acknowledgment

This research is funded by LPDP (*Lembaga Pengelola Dana Pendidikan*). We also acknowledge the lecturer of Research Approaches in Education and Research Project in Education in the Faculty of Education, Monash University Australia who gave many suggestions to this article. Lastly, we thank Monash University Library that provides resources for this article.

References

Adam, S. (2004). A consideration of the nature, role, application, and implications for European education of employing 'learning outcomes' at the local, national, and international levels. *United Kingdom Bologna Seminar*. 1-2 July 2004, Heriot-Watt University (Edinburgh Conference Centre), Edinburgh.

Anderson, J. L., & Barnett, M. (2013). Learning physics with digital game simulations in middle school science. *Journal of Science Education and Technology*, *22*(6), 914–926. https://doi.org/10.1007/s10956-013-9438-8

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M.-T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education*, *53*(1), 74–85. https://doi.org/10.1016/j.compedu.2008.12.020

Bello, S., Ibi, M. B., & Bukar, I. B. (2016). Effect of simulation techniques and lecture method on students' academic performance in mafoni day secondary school Maiduguri, Borno state, Nigeria. *Journal of Education and Practice*, *7*(23), 113–117. Retrieved from https://www.iiste.org/Journals/index.php/JEP/article/view/32584

Bennett, S., Maton, K., & Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, *39*(5), 775–786. https://doi.org/10.1111/j.1467-8535.2007.00793.x

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2*. Englewood Cliffs, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm. 12

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

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*(6), 825–840. https://doi.org/10.1002/sim.650

Chee, Y. S., & Tan, K. C. D. (2012). Becoming chemists through game-based inquiry learning: The case of "legends of alkhimia." *Electronic Journal of E-Learning*, *10*(2), 185–198. Retrieved from http://www.ejel.org/issue/download.html?idArticle=188

Chen, C.-L. D., Yeh, T.-K., & Chang, C.-Y. (2016). The effects of game-based learning and anticipation of a test on the learning outcomes of 10th grade Geology students. *EURASIA Journal of Mathematics, Science and Technology Education*, *12*(5), 1379–1388. https://doi.org/10.12973/eurasia.2016.1519a

Chen, C., & Hwang, G. (2017). Effects of the team competition-based ubiquitous gaming approach on students' interactive patterns, collective efficacy and awareness of collaboration and communication. *Educational Technology & Society*, *20*(1), 87–98. Retrieved from https://www.jstor.org/stable/jeductechsoci.20.1.87

Cheng, M.-T., Su, T., Huang, W.-Y., & Chen, J.-H. (2014). An educational game for learning human immunology: What do students learn and how do they perceive? *British Journal of Educational Technology*, *45*(5), 820–833. https://doi.org/10.1111/bjet.12098

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Chorney, A. I. (2012). Taking the game out Of gamification. *Dalhousie Journal of Interdisciplinary Management*, *8*(1), 1–14. https://doi.org/10.5931/djim.v8i1.242

Chu, H.-C., & Hung, C.-M. (2015). Effects of the digital game-development approach on elementary school students' learning motivation, problem solving, and learning achievement. *International Journal of Distance Education Technologies*, *13*(1), 87–102. https://doi.org/10.4018/ijdet.2015010105

Clark, D. B., Nelson, B. C., Chang, H.-Y., Martinez-Garza, M., Slack, K., &

D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers & Education*, *57*(3), 2178–2195. https://doi.org/10.1016/j.compedu.2011.05.007

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). New York, NY: Routledge.

Corbett, S. (2010, September 15). Learning by playing video games in the classroom. *The New York Times*. Retrieved from https://www.nytimes.com/2010/09/19/magazine/19video-t.html

Culp, K. M., Martin, W., Clements, M., & Lewis Presser, A. (2015). Testing the impact of a pre-instructional digital game on middle-grade students' understanding of photosynthesis. *Technology, Knowledge and Learning*, *20*(1), 5–26. https://doi.org/10.1007/s10758-014-9233-5

Divjak, B., & Tomić, D. (2011). The impact of game-based learning on the achievement of learning goals and motivation for learning mathematics - literature review. *Journal of Information and Organizational Sciences*, *35*(1), 15–30. Retrieved from https://jios.foi.hr/index.php/jios/article/view/182

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Echeverría, A., García-Campo, C., Nussbaum, M., Gil, F., Villalta, M., Améstica, M., & Echeverría, S. (2011). A framework for the design and integration of collaborative classroom games. *Computers & Education*, *57*(1), 1127–1136. https://doi.org/10.1016/j.compedu.2010.12.010

Fan, K.-K., Xiao, P., & Su, C. (2015). The effects of learning styles and meaningful learning on the learning achievement of gamification health education curriculum. *EURASIA Journal of Mathematics, Science and Technology Education*, *11*(5), 1211–1229. https://doi.org/10.12973/eurasia.2015.1413a

Furió, D., Juan, M.-C., Seguí, I., & Vivó, R. (2015). Mobile learning vs. traditional classroom lessons: A comparative study. *Journal of Computer Assisted Learning*, *31*(3), 189–201. https://doi.org/10.1111/jcal.12071

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (Revised an). New York, NY: Palgrave Macmillan.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage Publications.

Hamari, J., & Keronen, L. (2017). Why do people play games? A meta-analysis. *International Journal of Information Management*, *37*(3), 125–141. https://doi.org/10.1016/j.ijinfomgt.2017.01.006

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Honey, M. A., & Hilton, M. (Eds.). (2011). *Learning science through computer games and simulations*. Washington, DC: National Academy of Sciences.

Huang, B., Hew, K. F., & Lo, C. K. (2019). Investigating the effects of gamification-enhanced flipped learning on undergraduate students' behavioral and cognitive engagement. *Interactive Learning Environments*, *27*(8), 1106–1126. https://doi.org/10.1080/10494820.2018.1495653

Khalili, N., Sheridan, K., Williams, A., Clark, K., & Stegman, M. (2011). Students designing video games about immunology: Insights for science learning. *Computers in the Schools*, *28*(3), 228–240. https://doi.org/10.1080/07380569.2011.594988

Kim, H., Munson, M. R., & McKay, M. M. (2012). Engagement in mental health treatment among adolescents and young adults: A systematic review. *Child and Adolescent Social Work Journal*, *29*(3), 241–266. https://doi.org/10.1007/s10560-012-0256-2

King, W. R., & He, J. (2005). Understanding the role and methods of meta-analysis in IS research. *Communications of the Association for Information Systems*, *16*, 665–686. https://doi.org/10.17705/1CAIS.01632

Li, M.-C., & Tsai, C.-C. (2013). Game-based learning in science education: A review of relevant research. *Journal of Science Education and Technology*, *22*(6), 877–898. https://doi.org/10.1007/s10956-013-9436-x

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage Publications.

Mayo, M. J. (2007). Games for science and engineering education. *Communications of the ACM*, *50*(7), 30–35. https://doi.org/10.1145/1272516.1272536

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world.* New York, NY: Penguin.

Mertens, D. M. (2015). *Research and evaluation in education and psychology* (4th ed.). London: SAGE Publication.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*(6), 1417–1426. https://doi.org/10.2307/1911408

Nilsson, E. M., & Jakobsson, A. (2011). Simulated sustainable societies: Students' reflections on creating future cities in computer games. *Journal of Science Education and Technology*, *20*(1), 33–50. https://doi.org/10.1007/s10956-010-9232-9

Okeke, G. N. (2016). *The impact of digital games on high school students' learning outcomes in mathematics education: A meta-analytic investigation.* Doctoral thesis, University of North Texas, Denton, TX.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286. https://doi.org/10.1006/ceps.2000.1040

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679–703. https://doi.org/10.1037/0021-9010.78.4.679

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159. https://doi.org/10.2307/1164923

Papastergiou, M. (2009). Digital game-based learning in high school Computer Science education: Impact on educational effectiveness and student motivation. *Computers & Education*, *52*(1), 1–12. https://doi.org/10.1016/j.compedu.2008.06.004

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, *295*(6), 676. https://doi.org/10.1001/jama.295.6.676

Prensky, M. (2001). Digital natives, digital immigrants Part 1. *On the Horizon*, *9*(5), 1–6. https://doi.org/10.1108/10748120110424816

Quandt, T., Van Looy, J., Vogelgesang, J., Elson, M., Ivory, J. D., Consalvo, M., & Mäyrä, F. (2015). Digital games research: A survey study on an emerging field and its prevalent debates. *Journal of Communication*, *65*(6), 975–996. https://doi.org/10.1111/jcom.12182

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-*

analysis: Prevention, assessment and adjustments. Hoboken, NJ: John Wiley & Sons.

Sacks, H. S., Berrier, J., Reitman, D., Ancona-Berk, V. A., & Chalmers, T. C. (1998). Meta-Analyses and Large Randomized, Controlled Trials. *New England Journal of Medicine*, *338*(1), 59–62. https://doi.org/10.1056/NEJM199801013380112

Sánchez, J., & Olivares, R. (2011). Problem solving and collaboration using mobile serious games. *Computers & Education*, *57*(3), 1943–1952. https://doi.org/10.1016/j.compedu.2011.04.012

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. London: SAGE Publications.

Sjöblom, M., Törhönen, M., Hamari, J., & Macey, J. (2017). Content structure is king: An empirical study on gratifications, game genres and content type on Twitch. *Computers in Human Behavior*, *73*, 161–171. https://doi.org/10.1016/j.chb.2017.03.036

Spires, H. A., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). Problem solving and game-based learning: Effects of middle grade students' hypothesis testing strategies on learning outcomes. *Journal of Educational Computing Research*, *44*(4), 453–472. https://doi.org/10.2190/EC.44.4.e

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., … Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*(jul22 1), d4002–d4002. https://doi.org/10.1136/bmj.d4002

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8

Su, C.-H., & Cheng, C.-H. (2015). A mobile gamification learning system for improving the learning motivation and achievements. *Journal of Computer Assisted Learning*, *31*(3), 268–286. https://doi.org/10.1111/jcal.12088

Sung, H.-Y., & Hwang, G.-J. (2013). A collaborative game-based learning approach to improving students' learning performance in science courses. *Computers & Education*, *63*, 43–51. https://doi.org/10.1016/j.compedu.2012.11.019

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Chichester: John Wiley & Sons.

Ting, Y. L. (2010). Using mainstream game to teach technology through an interest framework. *Journal of Educational Technology & Society*, *13*(2), 141–152. Retrieved from https://scholar.lib.ntnu.edu.tw/en/publications/using-mainstream-game-to-teach-technology-through-an-interest-fra

Tsay, C. H.-H., Kofinas, A., & Luo, J. (2018). Enhancing student learning experience with technology-mediated gamification: An empirical study. *Computers & Education*, *121*, 1–17. https://doi.org/10.1016/j.compedu.2018.01.009

van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE Review*, *41*(2), 16–30. Retrieved from https://er.educause.edu/articles/2006/1/digital-gamebased-learning-its-not-just-the-digital-natives-who-are-restless

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*(3), 229–243. https://doi.org/10.2190/FLHV-K4WA-WPVQ-H0YM

Wrzesien, M., & Raya, M. A. (2010). Learning in serious virtual worlds: Evaluation of learning effectiveness and appeal to students in the E-Junior project.

Computers & Education, *55*(1), 178–187. https://doi.org/10.1016/j.compedu.2010.01.003

Yien, J.-M., Hung, C.-M., Hwang, G.-J., & Lin, Y.-C. (2011). A game-based learning approach to improving students' learning achievements in a nutrition course. *TOJET: The Turkish Online Journal of Educational Technology*, *10*(2). Retrieved from http://www.tojet.net/articles/v10i2/1021.pdf

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., … Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, *82*(1), 61–89. https://doi.org/10.3102/0034654312436980

REiD (Research and Evaluation in Education), 5(2), 2019, 169-182
Available online at: http://journal.uny.ac.id/index.php/reid

Parallel tests viewed from the arrangement of item numbers and alternative answers

*1Badrun Kartowagiran; 2Djemari Mardapi; 3Dian Normalitasari Purnama; 4Kriswantoro
1,2,3,4Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
*Corresponding Author. E-mail: kartowagiran@uny.ac.id

Submitted: 23 February 2019 | Revised: 7 December 2019 | Accepted: 23 December 2019

Abstract

This research aims to prove that a parallel test can be constructed by randomizing the test item numbers and or alternative answers' order. This study used the experimental method with a post-test only non-equivalent control group design, involving junior high schools students in Yogyakarta City with a sample of 320 students of State Junior High School (SMPN) 5 Yogyakarta and 320 students of SMPN 8 Yogyakarta established using the stratified proportional random sampling technique. The instrument used is a mathematics test in the form of an objective test consisting of a five-question package and each package contains 40 items with four alternatives. The test package is randomized in the item numbers' order from the smallest to the largest and vice versa. The options in each item are also randomized from A to D and vice versa. Each item is analyzed using the Classical Test Theory and Item Response Theory approaches, while data analysis is done using the discrimination index with Kruskal-Wallis test technique to see the differences among the five-question packages. The study reveals that the result of item analysis using the Classical Test Theory and Item Response Theory approaches shows no significant difference in the difficulty index among Package 1 until Package 5. Nevertheless, according to the Classical Test Theory, there is a category shift of the difficulty index of Package 2 until Package 5 when compared to Package 1 – the original package – which is, in general, not a good package, because it contains too easy items.

***Keywords**: correct option placement, order of items, parallel test*

Permalink/DOI: https://doi.org/10.21831/reid.v5i2.23721

Introduction

National Examination (NE) is one of the government efforts to improve the quality of education. In addition to its function to measure and evaluate the achievement of school graduate competence in certain subjects, as well as to map out the quality of primary and secondary education, NE also functions as the motivator for related parties to work better in order to achieve a good examination result (Center for Educational Assessment, 2014, p. 1). The education system and teaching quality are two related matters. A good teaching system will result in good quality learning (Mardapi, 2014, p. 12). Furthermore, teaching quality can be seen from the result of the evaluation done by teachers or educators.

According to Law No. 14 of 2005 of Republic of Indonesia, teachers are professsional educators whose main duties are to educate, teach, guide, direct, drill, assess, and evaluate students of formal early-childhood education, primary education, and secondary education, and it can be understood that teachers' role is not only to plan and implement teaching, but also to assess or evaluate. The assessment of students' learning achievement is expected not only to find out whether or not the stated learning objectives have been achieved, but also to reveal whether the objectives are important for students, and how the students achieve the learning objectives. Studies have shown that 87% teachers still find it difficult to perform assessment (Rusilowati in Rohmawati, 2013). The unsocialized procedures for conducting assessment becomes one of the constraints. This indicates that teachers' competence in doing good assessment still needs improvement.

One of the evaluation techniques which can be used to see students' competence is a testing technique. Miller, Linn, and Gronlund (2009, p. 28) define testing as an evaluation given to students for a certain period in their comparatively equal condition. A test usually consists of a set of questions. The aim of a test is to answer the question of how well an individual does something, in comparison to others or compared to the domain of performance work. The result of an assessment is the information about the characteristics of an individual or a group (Rasyid & Mansur, 2008, p. 11). In other words, using the assessment technique, a teacher is able to identify the characteristics of students' competence in a certain subject.

Based on the types of students' answers, tests can be classified into written tests, oral tests, and also performance tests (Sanjaya, 2010, p. 355). Written tests are further classified into essay tests and objective tests. One of the forms of objective tests is multiple-choice tests. Multiple-choice tests require students to choose a correct response out of some choices provided. In some choices, students choose the best choice from a list of alternatives. The choice of this type of tests is due to some consideration in relation to the strengths and weaknesses of multiple-choice tests (Reynolds, Livingston, & Willson, 2009). On the one hand, the strengths of multiple-choice tests include that a relatively large number of multiple-choice items can measure efficiently, objectively, and in a reliable manner. Besides, multiple-choice tests are very good at measuring lower cognitive objectives (for example knowledge, comprehension, and application) and they can minimize construction factors which are not relevant.

On the other hand, multiple-choice tests have weaknesses, including that they are relatively difficult to construct because more time has to be spent on making multiple-choice questions, they cannot measure all educational objectives (e.g. writing skill) although the item with alternative answers is very suitable for measuring higher cognitive domains (i.e. analysis, synthesis, and evaluation) and there are possibilities of random guessing. Identifying the strengths and weaknesses of multiple-choice tests becomes the consideration for choosing a certain type of tests.

The requirement for a good test is that it is valid and reliable (Azwar, 2013). Further, Azwar adds that measurement is said to be highly valid if it results in data that accurately give the description of the measured variable. Being accurate in this case means being exact and precise so that when a test results in data which are irrelevant to the goal of measurement, then it is considered as the measurement that has low validity. In addition to validity, the test reliability also needs to consider. The reliability of a test is said to be good if the test can result in scores/answers which are consistent although it is used by other examiners and at different time with the same condition. There are various ways of finding out the reliability of a test, one of them is the use of a parallel test (Azwar, 2015, p. 55).

Parallel tests are two or more tests whose purpose, difficulty index, and construction are the same but whose items are different. They are needed so that the alternative tests can be administered to different examinees or at different time, and also in the context of reliability estimation for the test given (Kronmüller et al., 2008; Werheid et al., 2002). So far, a parallel test can be constructed by changing the item number and or by changing the order of the alternative answers from one test package into various test packages. This has been an assumption for some teachers, while in reality there has not been a study showing that the strategy is correct, and that by changing the item number and or by changing the order of the alternative answers we can get parallel tests. For this reason, this research aims to find out the effect of randomizing item numbers and or the order of the alternative answers on item difficulty index.

Method

This research is an experiment applying the quantitative approach. It used the post-test only non-equivalent control group design to reveal the effect of randomizing item numbers and or the order of the alternative answers on item difficulty index. In this study, a mathematics test to be administered was randomized in terms of its item numbers and placement of the correct alternative to prove that the tests are really parallel. The sample was established using the stratified proportional random sampling technique. The test was reassembled into five test packages, administered to 320 students of State Junior High School (SMPN) 5 Yogyakarta and 320 students of SMPN 8 Yogyakarta. The result of the analysis shows the difference of the item difficulty index before and after the examinees do the test whose item numbers and alternative answer have undergone changes.

The data were collected using a test. The test used is a grade IX junior high school mathematics test consisting of 40 multiple-choice items with four alternative answers. The test was randomized in terms of the item numbers and alternative answers, resulting in five test packages as shown in Table 1.

Table 1. Test package and type of randomization

Test Package	Types of Randomization
1	Without randomization
2	Number randomization 1 -40
3	Number randomization 1-20 and 21-40
4	Number randomization 1-10. 11-20. 21-30 and 31-40
5	Alternative answer randomization

Table 1 shows the type of randomization of test packages. Package 1 is the original grade IX junior high school mathematics test whose item numbers and alternative answers did not undergo randomization. Package 2 is the test package resulted from the randomization of entire numbers 1–40, by reversing number 40 into number 1, number 39 into number 2, number 38 into number 3 and so on. Package 3 is the test package resulted from the randomization of numbers 1-20 and 21–40, by reversing number 20 into number 1, number 19 into number 2, and so on. Furthermore, number 40 is reversed into number 21, number 39 into number 22, and so on. Package 4 is the test package resulted from the randomization of numbers 1-10, 11-20, 21-30 and 31-40, by reversing number 10 into number 1, number 9 into number 2, and so on. Further, number 20 was reversed into number 11, number 19 into number 2, and so on. Number 30 was reversed into number 21, number 29 into number 22, and so is the case with numbers 31-40. Package 5 is the test package resulted from the randomization of alternative answers, where alternative d becomes alternative a, alternative c becomes alternative b, alternative b becomes alternative c, and alternative a becomes alternative d.

The result of administering the parallel tests to students was analysed to see if the randomization of item numbers and alternative answers really resulted in parallel tests. The data in the form of students' answer sheets were analysed using the Clasical Test Theory and Item Response Theory. The analysis using the Clasical Test Theory was done with the help of the *QUEST* program to see the item difficulty index, and the analysis using the Item Response Theory used the one-parameter logistic approach employing the *QUEST* program to see the item difficulty index and student competence parameter.

In addition to being analysed using the Clasical Test Theory, the data were analysed for their reliability index. The reliability index was seen based on the output of the analysis using the *QUEST* program. The analysis of the effect of the randomization of item numbers on the test's being parallel was conducted using Kruskall-Wallis analysis with the help of the SPSS Program. If the result of the analysis shows Sig < 0.05 then there is an effect of the randomization of item numbers of the parallel tests on the difficulty index of the tests.

Findings and Discussion

Findings

The findings of this study show that the instrument reliability index of the five test packages is good. Mehrens and Lehmann (1973, p. 122) write that the minimum reliability index of an instrument is 0.80. Package 1, Package 2, and Package 3 have the same reliability index of 0.96, while the reliability index of Package 4 and Package 5 is 0.97. This finding shows that the instrument reliability index of the five test packages is in the reliable category. Furthermore, the five packages were analysed using the Classical Test Theory and the Item Response Theory. This analysis is conducted in order to find out the test parallelism based on the test item difficulty.

Parallel Tests Based on the Classical Test Theory Approach

Before scrutinizing whether or not Package 2, Package 3, Package 4, and Package 5 are parallel to Package 1, the test item characteristic funcion was studied based on the Classical Test Theory. According to the Classical Test Theory, whether a test is good or not depends on the value of the difficulty index, discrimination power, and the functioning of the distractors. Allen and Yen (1979) classify item difficulty indices into three cate-

Table 2. Characteristic function of test Package 1 based on the Classical Test Theory

Item Numbers	Difficulty Index	Category	Discrimination Power	Category	Distractor	Conclusion
1	0.945	Easy	-0.07	Poor	All Functioning	Poor
2	0.874	Easy	0.29	Good	All Functioning	Poor
3	0.929	Easy	0.15	Poor	All Functioning	Poor
4	0.134	Difficult	0.15	Poor	All Functioning	Poor
5	0.638	Moderate	0.35	Good	All Functioning	Good
6	0.709	Easy	0.28	Good	All Functioning	Poor
7	0.961	Easy	0.14	Poor	All Functioning	Poor
8	0.724	Easy	0.34	Good	All Functioning	Poor
9	0.937	Easy	0.27	Good	All Functioning	Poor
10	0.701	Easy	0.36	Good	All Functioning	Poor
11	0.748	Easy	0.29	Good	All Functioning	Poor
12	0.480	Moderate	0.47	Good	All Functioning	Good
13	0.827	Easy	0.16	Poor	All Functioning	Poor
14	0.953	Easy	0.22	Good	A and D not functioning	Poor
15	0.449	Moderate	0.24	Good	All Functioning	Good
16	0.740	Easy	0.74	Good	All Functioning	Poor
17	0.913	Easy	0.35	Good	All Functioning	Poor
18	0.827	Easy	0.24	Good	All Functioning	Poor
19	0.646	Moderate	0.26	Good	All Functioning	Good
20	0.748	Easy	0.05	Poor	All Functioning	Poor
21	0.961	Easy	0.11	Poor	A and D not functioning	Poor
22	0.709	Easy	0.31	Good	All Functioning	Poor
23	0.102	Difficult	0.1	Poor	All Functioning	Poor
24	0.307	Moderate	0.21	Good	All Functioning	Good
25	0.268	Difficult	0.35	Good	All Functioning	Poor
26	0.433	Moderate	0.4	Good	All Functioning	Good
27	0.764	Easy	0.34	Good	All Functioning	Poor
28	0.921	Easy	0.34	Good	All Functioning	Poor
29	0.748	Easy	0.29	Good	All Functioning	Poor
30	0.449	Moderate	0.11	Poor	All Functioning	Good
31	0.654	Moderate	0.44	Good	All Functioning	Good
32	0.898	Easy	0.15	Poor	All Functioning	Poor
33	0.535	Moderate	0.52	Good	All Functioning	Good
34	0.654	Moderate	0.35	Good	All Functioning	Good
35	0.732	Easy	0.22	Good	All Functioning	Poor
36	0.882	Easy	0.31	Good	All Functioning	Poor
37	0.591	Moderate	0.04	Poor	All Functioning	Good
38	0.346	Moderate	0.31	Good	All Functioning	Good
39	0.693	Moderate	0.29	Good	All Functioning	Good
40	0.465	Moderate	0.38	Good	All Functioning	Good

gories. An item is in the difficult category if its coefficient is <0.3; it is in the moderate category when its coefficient is between 0.3 and 0.7, and it is in the easy category when the coefficient is >0.7. A good test item has the difficulty index in the moderate category. The item discrimination power was also used as a consideration in deciding if test item is good or poor. Fernandes (1984) states that a good test item is an item with the discrimination power of >0.2. He adds that a distractor is considered functioning when it is chosen by at least 2% of the total examinees.

Table 2 shows the analysis result of the item characteristic function based on the Classical Test Theory. It shows the characteristics of the 40 test items in Test Package 1. In general, the items in Test Package 1 are in a poor category. A test item is said to be good when it meets three categories, i.e. having a moderate difficulty index and good discrimination power, and all of its distractors function well. In Test Package 1, only fourteen items (35%) are in a good category, while 26 items (65%) are in a poor category. Table 2 also shows that 23 items (57.5%) are categorized as easy.

Table 3. Difficulty index of five test packages (Classical Test Theory approach)

Item Number	**Item Difficulty Index**				
	Package 1	**Package 2**	**Package 3**	**Package 4**	**Package 5**
1	0.945	0.952	0.641	0.651	0.954
2	0.874	0.839	0.516	0.914	0.868
3	0.929	0.903	0.730	0.638	0.829
4	0.134	0.129	0.897	0.063	0.033
5	0.638	0.508	0.468	0.638	0.829
6	0.709	0.508	0.754	0.533	0.638
7	0.961	0.903	0.817	0.330	0.967
8	0.724	0.702	0.833	0.829	0.638
9	0.937	0.902	0.683	0.868	0.914
10	0.701	0.637	0.714	0.954	0.661
11	0.748	0.637	0.889	0.586	0.638
12	0.480	0.306	0.873	0.645	0.408
13	0.827	0.742	0.770	0.816	0.697
14	0.953	0.847	0.079	0.875	0.941
15	0.449	0.435	0.460	0.737	0.474
16	0.740	0.750	0.675	0.474	0.737
17	0.913	0.895	0.921	0.941	0.875
18	0.827	0.863	0.619	0.697	0.816
19	0.646	0.694	0.881	0.408	0.645
20	0.748	0.398	0.651	0.638	0.586
21	0.961	0.919	0.714	0.428	0.961
22	0.709	0.661	0.849	0.632	0.632
23	0.102	0.266	0.651	0.743	0.566
24	0.307	0.839	0.675	0.658	0.493
25	0.268	0.782	0.722	0.250	0.349
26	0.433	0.331	0.849	0.349	0.250
27	0.764	0.766	0.611	0.493	0.658
28	0.921	0.935	0.556	0.566	0.743
29	0.748	0.726	0.683	0.632	0.632
30	0.449	0.460	0.206	0.961	0.428
31	0.654	0.734	0.968	0.283	0.684
32	0.898	0.863	0.556	0.664	0.862
33	0.535	0.718	0.278	0.553	0.618
34	0.654	0.480	0.302	0.645	0.724
35	0.732	0.815	0.325	0.822	0.724
36	0.882	0.895	0.325	0.724	0.822
37	0.591	0.629	0.786	0.724	0.645
38	0.346	0.653	0.857	0.618	0.553
39	0.693	0.718	0.754	0.862	0.664
40	0.465	0.323	0.484	0.684	0.283
Average (*b*)	**0.675**	**0.677**	**0.651**	**0.661**	**0.668**

Based on the analysis from Test Package 1, the difficulty index of each item in Test Packages 2, 3, 4, and 5 was analyzed. The item difficulty index analysis using the Classical Test Theory was done with the QUEST program. The analysis result of the parameter of item difficulty index of each test package is shown in Table 3. It shows the difficulty index of each item in the five test packages. Package 1 is the original test package without any randomization, so Package 2, Package 3, Package 4, and Package 5 which had undergone randomization were reconstructed to their former forms with item numbers being rearranged to their original arrangement.

Table 4 shows that after the randomization of item numbers and alternative answers, the difficulty index of the five packages ranged from 0.102 to 0.968. This range is quite large, because, according to the Classical Test Theory, the difficulty index should range from 0 to 1. Further, based on the result of the analysis shown in Table 4, the characteristics of each test items in the five packages was analysed. The result of analysis of the each test item characteristics in terms of difficulty index is shown in Table 4.

Table 4 shows that all five test packages, viewed from the difficulty index, generally show that the test items are in easy and moderate categories. The test packages have undergone randomization and have been reconstructed into their former construction before randomization. It can be seen from the same proportion of the test packages, while the number of the items in the difficult category is only two or three. A deeper look into it reveals that some items have gone through changes in the category of difficulty index. For instance, Item 6 in Package 1 was categorized as an easy item, but after the randomization in Package 2, it was categorized as a moderate item. Another example is Item 25 in Package 1, categorized as a difficult item, but after the randomization of Package 2, in Package 3 it was categorized as an easy item. It shows that seen from the difficulty index category, many items change after the item numbers are randomized. The percentages of the changes or shifts in the item difficulty category is shown in Table 5.

Table 5 shows that the biggest shift in difficulty index is the shift of 24 items (60%) from Package 1 to Package 3, while the smallest shift is the shift from Package 1 to Package 2, i.e. 9 items (22.5). Based on the result of the analysis using the Classical Test Theory approach, Kruskall-Wallis analysis was conducted to see whether there was any significant difference of the item difficulty index of the randomized test packages. The summary of the result of the analysis is in Table 6.

Table 4. Characteristics of item difficulty index based on Classical Test Theory

Category	Package 1 (Item Number)	Package 2 (Item Number)	Package 3 (Item Number)	Package 4 (Item Number)	Package 5 (Item Number)
Easy	1, 2, 3, 6, 7, 8, 9, 10, 11, 13, 14, 16, 17, 18, 20, 21, 22, 27, 28, 29, 32, 35, 36	1, 2, 3, 7, 8, 9, 11, 13, 14, 16, 17, 18, 21, 24, 25, 27, 28, 29, 31, 32, 33, 35, 36, 39	3, 4, 6, 7, 8, 10. 11, 12, 13, 17, 19, 21, 22, 25, 26, 31, 37, 38, 39	2, 8, 9, 10, 13, 14, 15, 17, 23, 20, 35, 39	1, 2, 3, 5, 7, 9, 14, 16, 17, 18, 21, 28, 34, 35, 32, 34, 35, 36
%	57.5%	60%	47.5%	30%	45%
Moderate	5, 12, 15, 19, 24, 26, 30, 31, 33, 34, 37, 38, 39, 40	5, 6, 10, 12, 15, 19, 20, 22, 26, 30, 34, 37, 38, 40	1, 2, 5, 9, 15, 16, 18, 20, 23, 24, 27, 28, 29, 32, 34, 35, 36, 40	1, 3, 5, 6, 7, 11, 12, 16, 18, 19, 20, 21, 22, 24, 26, 27, 28 29, 32, 33, 34, 36, 37, 38, 40	6, 8, 10, 11, 12, 13, 15, 19, 20, 22, 23, 24, 25, 27, 29, 30, 31, 33, 37, 38, 39
%	37.5%	35%	45%	62.5%	52.5%
Difficult	4, 23, 25	4, 23,	14, 30, 33	4, 25, 31	4, 26, 40
%	7.5%	5%	7.5%	7.5%	7.5%

Table 5. Category shift of item difficulty index of five test packages

Packages 1-2	Packages 1-3	Packages 1-4	Packages 1-5
9 items (22.5%)	24 items (60%)	20 items (50%)	15 items (37.5%)

Table 6 shows that the value of Asymp, Sig in all items whose discrimination power is tested among Package 1, Package 2, Package 3, Package 4, and Package 5 is above 0.05. It means that there is no difference in difficulty index of the items in all five test packages, so there is no effect item number randomization on the item difficulty index. After the effect of item number randomization was scrutinized, the effect of the randomization on discrimination index was analysed. The percentages of the good and poor discrimination index is shown in Table 7.

Table 7 shows that the discrimination index of Test Packages 1, 2, 3, 4, and 5 is in a good category (> 60%). Based on the analysis of Test Package 1, after the randomization of Test Packages 2, 3, 4, and 5, there is a shift in the good discrimination index. However, a closer look reveals that the shift is not big enough, occurring to two to four items only.

Parallel Tests Based on Analysis Using Item Response Theory Approach

Before scrutinizing whether Package 2, Package 3, Package 4 and Package 5 are parallel to Package 1 or not, the researchers need to describe the assumption test of the Item Response Theory (IRT), which is the unidimension assumption test (Naga, 1992). The requirement for unidimension is aimed at sustaining invariance in IRT. If a test item measures more than one dimension, then the answer to the item is a combination of different competencies of the examinees. Thus, the contribution of each competency to the answer is unknown.

Unidimension assumption testing is carried out to reveal whether a test measures one trait. The unidimension assumption is tested by the factor analysis and its empirical result. The KMO-MSA value is sufficient if it is above 0.5 (Field, 2009). By looking at the first eigenvalue contribution to test variance, according to Reckase (1979), the formation of eigenvalue factor has to have a value above 1. In the factor analysis, the first eigenvalue has to have the biggest value (dominant) compared to the second, third, and so forth eigenvalues. The result of the analysis of unidimension assumption testing is shown in Table 8.

Table 6. Result of Kruskall-Wallis analysis of the Classical Test Theory

Item Numbers	Asymp, Sig
1-5	0.810
6-10	0.885
11-15	0.819
16-20	0.760
21-25	0.418
26-30	0.882
31-35	0.344
36-40	0.760

Table 7. Category of power discrimination of five test packages

Discrimination Power	Package 1	Package 2	Package 3	Package 4	Package 5
Good	29 items (72.5%)	27 items (67.5%)	33 items (82.5%)	33 items (82.5%)	32 items (80%)
Poor	11 items (27.5%)	13 items (22.5%)	7 items (17.5%)	7 items (17.5%)	8 items (20%)

Table 8. Unidimension assumption test

Test	KMO and Bartlett's Test		Total Variance Explained		Category
Packages	KMO	Sig.	*Eigenvalue* Factor 1	*Eigenvalue* Factor 2	
Package 1	0.469	0.00	3.637	2.831	Multidimension
Package 2	0.513	0.00	3.807	2.223	Multidimension
Package 3	0.608	0.00	5.891	2.367	Unidimension
Package 4	0.571	0.00	5.345	2.483	Unidimension
Package 5	0.580	0.00	5.003	2.446	Unidimension

Table 8 presents that of the five packages whose unidimension assumption was analyzed, three packages are unidimensional (Package 3, Package 4, and Package 5), while two packages are multidimensional (Package 1 and Package 2). The analysis was based on the size of the sample sufficiency value (KMO) and eigenvalue. The second assumption is the local independence assumption and parameter invariance. According to Retnawati (2014, p. 7), this assumption is automatically proved after it is proved with unidimensionality.

After the assumption testing, the test item characteristic was analysed by the IRT. Testing the fitness of each item to the model followed the formula by Sumintono and Widhiarso (2015, p. 81) that an item fits to a model if the value of Outfit MNSQ is between 0.5 and 1.5. An item difficulty index can be known from the most difficult, moderate, and easiest item. An item difficulty index is categorized easy if it has the difficulty index close to -2.00. An item difficulty index is categorized moderate if its difficulty index value ranges from -1.00 to +1.00. An item difficulty index is categorized difficult if its difficulty index is close to +2.00. The result of the analysis of item characteristics based on difficulty index is shown in Table 9.

Table 9. Characteristics of items in Package 1 based on the Item Response Theory

Item Number	Model Fitness	Category	Difficulty index	Category	Category
1	1.77	Not Fit	0.390	Moderate	Poor
2	0.91	Fit	0.270	Moderate	Good
3	0.97	Fit	0.350	Moderate	Good
4	1.26	Fit	0.270	Moderate	Good
5	0.96	Fit	0.190	Moderate	Good
6	1.16	Fit	0.200	Moderate	Good
7	0.69	Fit	0.460	Moderate	Good
8	0.91	Fit	0.210	Moderate	Good
9	0.75	Fit	0.370	Moderate	Good
10	0.82	Fit	0.200	Moderate	Good
11	0.88	Fit	0.210	Moderate	Good
12	0.94	Fit	0.190	Moderate	Good
13	1.24	Fit	0.240	Moderate	Good
14	0.64	Fit	0.420	Moderate	Good
15	1.03	Fit	0.190	Moderate	Good
16	0.87	Fit	0.210	Moderate	Good
17	0.62	Fit	0.320	Moderate	Good
18	0.99	Fit	0.240	Moderate	Good
19	1,01	Fit	0.190	Moderate	Good
20	1.25	Fit	0.210	Moderate	Good
21	0.96	Fit	0.460	Moderate	Good
22	0.91	Fit	0.200	Moderate	Good
23	1.39	Fit	0.300	Moderate	Good
24	1.13	Fit	0.200	Moderate	Good
25	0.97	Fit	0.210	Moderate	Good
26	0.94	Fit	0.190	Moderate	Good
27	0.81	Fit	0.220	Moderate	Good
28	1.07	Fit	0.340	Moderate	Good
29	0.94	Fit	0.210	Moderate	Good
30	1.16	Fit	0.190	Moderate	Good
31	0.80	Fit	0.200	Moderate	Good
32	0.86	Fit	0.300	Moderate	Good
33	0.83	Fit	0.190	Moderate	Good
34	1.00	Fit	0.200	Moderate	Good
35	1.13	Fit	0.210	Moderate	Good
36	0.88	Fit	0.280	Moderate	Good
37	1.28	Fit	0.190	Moderate	Good
38	0.96	Fit	0.200	Moderate	Good
39	1.00	Fit	0.200	Moderate	Good
40	1.04	Fit	0.190	Moderate	Good

Table 9 shows that, in terms of good criteria items, 39 items fit, and one item does not fit to Rasch model because it is outside the stated OUTFIT MNSQ range. Furthermore, in terms of the item difficulty index, all items fall into the moderate category, and therefore it can be concluded that only one of the 40 items is not good. Later, based on the result of the analysis of Package 1, the analysis of the difficulty index of the items in the other test packages was conducted. The item analysis using the Item Response Theory of five test packages resulted in the value of parameter of the difficulty index of each item as shown in Table 10.

Table 10 shows the difficulty value of each test item in five test packages after the item is suited to the items in Package 1. Baker (2001, p. 11) divides difficulty indices of items according to the IRT into five categories: very easy, easy, moderate, difficult, and very difficult. An item is said to be very easy if its difficulty index value is lower than -2.00. An item is categorized easy if it has the difficulty index value close to -2.00. An item is categorized moderate if it has the difficulty index value

Table 10. Difficulty index of five test packages based on the Item Response Theory

Item Number	**Difficulty Index**				
	Package 1	**Package 2**	**Package 3**	**Package 4**	**Package 5**
1	0.390	0.420	0.200	0.180	0.390
2	0.270	0.250	0.190	0.300	0.250
3	0.350	0.310	0.220	0.180	0.220
4	0.270	0.280	0.300	0.460	0.470
5	0.190	0.190	0.190	0.180	0.170
6	0.200	0.190	0.220	0.170	0.180
7	0.460	0.310	0.240	0.470	0.460
8	0.210	0.200	0.250	0.220	0.180
9	0.370	0.310	0.210	0.250	0.300
10	0.200	0.200	0.210	0.390	0.180
11	0.210	0.200	0.290	0.180	0.180
12	0.190	0.200	0.280	0.180	0.180
13	0.240	0.210	0.220	0.220	0.190
14	0.420	0.260	0.360	0.250	0.350
15	0.190	0.190	0.190	0.190	0.170
16	0.210	0.220	0.200	0.170	0.190
17	0.320	0.300	0.340	0.350	0.250
18	0.240	0.270	0.200	0.190	0.220
19	0.190	0.200	0.290	0.180	0.180
20	0.210	0.230	0.200	0.180	0.180
21	0.460	0.340	0.210	0.180	0.420
22	0.200	0.200	0.260	0.180	0.180
23	0.300	0.210	0.200	0.200	0.180
24	0.200	0.250	0.200	0.180	0.170
25	0.210	0.230	0.210	0.200	0.180
26	0.190	0.200	0.260	0.180	0.200
27	0.220	0.220	0.200	0.170	0.180
28	0.340	0.370	0.190	0.180	0.200
29	0.210	0.210	0.210	0.180	0.180
30	0.190	0.190	0.240	0.420	0.180
31	0.200	0.210	0.520	0.190	0.190
32	0.300	0.270	0.190	0.180	0.240
33	0.190	0.210	0.220	0.170	0.180
34	0.200	0.420	0.210	0.180	0.190
35	0.210	0.240	0.210	0.220	0.190
36	0.280	0.300	0.210	0.190	0.220
37	0.190	0.190	0.230	0.190	0.180
38	0.200	0.200	0.270	0.180	0.170
39	0.200	0.210	0.220	0.240	0.180
40	0.190	0.200	0.190	0.190	0.190
Average	**0.250**	**0.245**	**0.236**	**0.222**	**0.222**

ranging from -1.00 to +1.00. An item is categorized difficult if it has the difficulty index value close to +2.00, and categorized as very difficult if the difficulty index value is higher than +2.00. Based on the result of the analysis using the Item Response Theory, all items in Package 1, Package 2, Package 3, Package 4 and Package 5 have the difficulty index in a good category. This is in line with Table 8 which shows that all difficulty indexes of the items range from higher than -1.00 to lower than 1.00, which means that all items have the difficulty index in the moderate category.

In addition to showing item characteristics based on difficulty index according to the IRT, Table 10 also shows the average difficulty index of 40 test items in five test packages. Table 10 shows that the average difficulty index of the test items in Package 1 is 0.250, in Package 2 it is 0.245, in Package 3 it is 0.236, in Package 4 it is 0.222, and in Package 5 sit is 0.222. Table 10 also shows that all items in five packages have the difficulty index which is not very different from each other. Based on the result of the analysis using the Classical Test Theory, a test was done to see the significance of the differences in item difficulty index among the randomized test packages. The test was conducted using Kruskall-Wallis analysis. The summary of the analysis result is presented in Table 11.

Table 11. The result of the test using Kruskall-Wallis of the Classical Test Theory

Item Number	Asymp. Sig
1-5	0.591
6-10	0.795
11-15	0.178
16-20	0.222
21-25	0.063
26-30	0.094
31-35	0.054
36-40	0.110

Table 11 shows the value of Asymp, Sig of all items whose difference among Package 1, Package 2, Package 3, Package 4, and Package 5 is above 0.05. This means that there is no difference in the difficulty index of the five test packages. Therefore, there is no effect of item number randomization on the item difficulty index.

Discussion

Mathematics is one of the school subjects which is tested in junior high school national examination. Hamdi, Kartowagiran, and Haryanto (2018) believe that students' mathematics competence can be used to solve varieties of problems and difficulties they face in learning various sciences, especially natural science. This fact forms the basis for the importance of mathematics, so that it becomes one of the school subjects examined in the national examination. The mathematics test in the national examination consists of a number of parallel test packages. The packages are constructed with the same items but with randomized item numbers and alternative answers in order to distinguish one package from the others. The use of parallel test packages is expected to prevent students from cheating, so that their real mastery can be known. Unparallel tests may result in error of measurement, that is, the result of the test does not show the real competence mastery of the students (Purnama, 2017). This research is conducted by analysing five test packages which are different based on the item randomization in order to prove whether being randomized the test packages are really parallel.

Whether or not a test is of good quality can be seen in the difficulty index of each item. A test item is said to be good if it is neither too difficult nor too easy, or in other words, the difficulty index is moderate. The item difficulty index is usually related to the aim of the test (Mehrens & Lehmann, 1973, p. 195). This research applies the Classical Test Theory and the Item Response Theory approaches in the analysis of test item difficulty index. The Classical Test Theory approach is a very simple approach and easy to understand in analyzing test items empirically (Güler, Uyanik, & Teker, 2014), while the Item Response Theory approach is used to cover the weaknesses of the Classical Test Theory approach.

Before a further analysis was conducted to find out whether a test remained parallel after its item numbers were randomized, the quality/characteristic function of the items in Package 1 was analysed, because Package 1 is the original test package as the reference for the analysis of the other four test packages. Putro (2013) states that good test items have to meet at least three requirements, i.e. item difficulty index, discrimination power, and well-functioned distractors. The result of the analysis using the Classical Test Theory shows that in general Package 1 is in a poor category. This can be seen in the difficulty index, discrimination power, and the functioning of the distractors. Viewed from the value of the difficulty index, it is very obvious that there are still many items in the easy category, and thus the students can answer correctly.

The result of the analysis of the five test packages using the Classical Test Theory approach shows that, in terms of the difficulty index, out of the 40 test items in five test packages, 5% to 7.5% of the items are difficult items, 35% to 62.5% of the items are moderate or good items, and 30% to 60% of the items are easy items. Viewed from the average of the item difficulty index as shown in Table 4, all of the five test packages have the average difficulty index categorized moderate or good. The value of the item difficulty index of the five test packages lies between 0.102 and 0.968. The higher the difficulty index, the easier the test item will be, and vice versa, the lower the item difficulty index, the more difficult the item will be (Bichi, 2016). This is in line with Allen and Yen (1979) who state that in test item measurement, the item difficulty index is related to the percentage of the examinees who can do the test correctly. Difficulty index is the proportion of the number of test takers who answer a particular question correctly, the proportion of all test takers.

Based on the Classical Test Theory, it is known that there has been a shift in the category of the difficulty index of some items in Package 2, Package 3, Package 4, and Package 5 compared to that of the items in Package 1. For example, test item 1 in Package 1 is in the easy category, in Package 3 and Package 4 it is in the moderate category. Another example is that test item 13 in Package 1 is in the easy category, but in Package 2 it is in the moderate category. Overall, the percentage of the shift of the category of the difficulty index of Package 2 is 22.5%, Package 3 is 60%, Package 4 is 50% and Package 5 is 37.5%. This is due to the weakness of the result of the item analysis using the Classical Test Theory approach, i.e. the size of the item characteristics (in this case the difficulty index) depends on the distribution of the competence of the test takers in the sample that is used (Awopeju & Afolabi, 2016). In line with this opinion, Zaman, Kashmiri, Mubarak, and Ali (2008) add that the comparison of test result of different test takers is one of the weaknesses of the Classical Test Theory which is worth noting, because test takers must do the items which are the same or really parallel. It is one of these weaknesses that necessitate the IRT to come into use.

In the IRT, the first thing to see is the assumption test. The unidimension assumption testing of the five test packages must first see the sufficiency of the sample. Research findings show that the value of KMO-MSA of Package 1 is 0.469, Package 2 is 0.513, Package 3 is 0.608, Package 4 is 0.571, and Package 5 is 0.580. According to Field (2009), the value of KMO-MSA is considered sufficient if it is above 0.5. From this result, it can be concluded that four packages have sufficient sample, i.e. Package 2, Package 3, Package 4, and Package 5, because the value of KMO-MSA >0.5. The result of the significance analysis using Barlett's Test of Sphericity shows that each of the five test packages is at the significance level of 0.000. Therefore, the requirement is met because the significance level is below 0.05.

There are a number of ways to interpret the sufficiency of unidimension assumption. One of the ways is by looking at the contribution of the first eigen value to test variance. The result of the above analysis shows that three test packages have dominant factors whose value is more than twice as much as the second factor, i.e. Package 3 with 5.891 which is higher than the *eigenvalue* of the second factor of 2.367. Package 4 with 5.345 higher than the eigenvalue of the second factor of 2.483, and Package 5 with 5.003 345 higher than the eigenvalue of the second factor of 2.446, where the first factor is the most dominant factor. In the factor analysis, the first eigenvalue should have the highest value (dominant) compared to the second, third, and so forth eigenvalue. This is because the size of the variance is directly proportional with the size of eigenvalue (Field, 2009, p. 652; Johnson & Wichern, 2002, p. 441), and therefore, it can be concluded that the first factor in the factor analysis contributes the most compared to the other factors, and thus the unidimensionality assumption is met.

Difficulty index (b) which lies between the range of -2 and 2 is good (Surya & Aman, 2016). The result of the analysis using the IRT approach shows that the five test packages have the difficulty index ranging from 0.170 to 0.470. The value of the difficulty index shows that all of the test items are in the moderate category, which lies between -1.00 and +1.00 (Sumintono & Widhiarso, 2015). It means that based on the result of the analysis using the IRT, all test items in the five test packages have the same characteristics.

The analysis of the characteristics of the item difficulty index was then followed by Kruskall-Wallis analysis to reveal the effect of the randomization of the item numbers and alternative answers on the item difficulty index. The Kruskall-Wallis analysis was conducted using the value of the difficulty index obtained using the Classical Test Theory and Item Response Theory approaches. The result of the analysis using the Classical Test Theory approach shows that the randomization of the item numbers and alternative answers does not affect the item difficulty index as shown by the value of Asymp, Sig above 0.05. This result is in line with the finding of the research by Santoso (2013) which states that the estimation of the competence and length of the test with randomized design is not significantly different from the test which was not randomized. The research finding applying the IRT approach shows that there is a difference in the difficulty index of the test items in the five test packages after the randomization.

In relation to the case of the Classical Test Theory approach, the absence of the effect of the randomization of the item numbers and alternative answers may result from Package 1 which is the original test package not having undergone any randomization. Package 1 has the characteristics which tend to be poor. Viewed from its difficulty index, more than 50% of the items are easy items which make most students, those with high competence and those with low competence, can answer questions correctly. It means that the test cannot distinguish students with high competence from those with low competence. A test that tends to be easy for students will not show any effect of randomization because they will tend to be able to do it. In addition, a test was conducted using Kruskal-Wallis test on the difficulty index using the Classical Test Theory and Item Response Theory. Package 1 which is the original test package is used to find out whether there is a difference in the difficulty index between items 1-10 and items 31-40. The result shows the Assymp, Sig value of 0.082 when using the Classical Test Theory, and the Assymp, Sig value of 0.054 when using the IRT, where the Assymp, Sig value is above 0.05. It means that in the original test package, before randomization, the values of the item difficulty index are not in a wide range. This may be the reason for the absence of the difference in the difficulty index after randomization. Further studies need to be done on the test items which have good characteristics to see whether or not there is an effect of the randomization of the item numbers and alternative answers on item difficulty index.

Conclusion

All of the five test packages have a good reliability index, lying between 0.96 and 0.97. Package 1, Package 2, and Package 3 have the reliability index of 0.96, while Package 4 and Package 5 have the reliability index of 0.97. It can be concluded that based on the value of the reliability index, the five test packages have equal reliability.

Based on the result of the analysis using the Classical Test Theory, viewed from the average value of the difficulty index, all five test packages have the average difficulty index ranging from 0.102 to 0.968. The result of Kruskall-Wallis analysis of the five test packages shows that there is no difference in the difficulty index of the items in Package 1, Package 2, Package 3, Package 4 and Package 5. Thus, the randomization of the item numbers and alternative answers has no effect on the item difficulty index.

The analysis of the test items using the Item Response Theory shows that the average value of difficulty index of the five test packages ranges from 0.170 to 0.470. The result of the analysis of the difficulty index of the items in the five test packages shows that there is no difference in the difficulty felt by the students doing Package 1, Package 2, Package 3, Package 4, and Package 5. This means that the randomization of item numbers has no effect on the item difficulty index, which means that constructing parallel tests by randomizing the item numbers and alternative answers is good to do, and this research has proved that applying this method will result in parallel tests.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Los Angeles, CA: Wadsworth.

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of Classical Test Theory and Item Response Theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, ESJ*, *12*(28), 263–284. https://doi.org/10.19044/esj.2016.v12n28p263

Azwar, S. (2013). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: Pustaka Pelajar.

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

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

Bichi, A. A. (2016). Classical Test Theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, *2*(9), 27–33. https://doi.org/10.26643/ijss.v2i9. 6690

Center for Educational Assessment. (2014). *Laporan pengolahan Ujian Nasional tahun ajaran 2014/2015* (Unpublished). Jakarta: Center for Educational Assessment of Republic of Indonesia.

Fernandes, H. J. X. (1984). *Testing and measurement*. Jakarta: National Education Planning, Evaluation, and Curriculum Development.

Field, A. (2009). *Discovering statistics using SPSS* (3rd 3d.). London: Sage Publications.

Güler, N., Uyanik, G. K., & Teker, G. T. (2014). Comparison of Classical Test Theory and Item Response Theory in terms of item parameters. *European Journal of Research on Education*, *2*(1), 1–6.

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

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Kronmüller, K.-T., Saha, R., Kratz, B., Karr, M., Hunt, A., Mundt, C., & Backenstrass, M. (2008). Reliability and validity of the knowledge about depression and mania inventory. *Psychopathology*, *41*(2), 69–76. https://doi.org/10.1159/000111550

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

Mardapi, D. (2014). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Litera.

Mehrens, W. A., & Lehmann, J. L. (1973). *Measurement and evaluation in education and psychology*. New York, NY: Holt, Rinehart, and Winston.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

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

Purnama, D. N. (2017). Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum. *REiD (Research and Evaluation in Education)*, *3*(2), 152–162. https://doi.org/10.21831/reid.v3i2.18121

Putro, N. H. P. S. (2013). Karakteristik butir soal ulangan kenaikan kelas sebagai persiapan bank soal Bahasa Inggris. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(1), 92–114. https://doi.org/10.21831/pep.v15i1.1089

Rasyid, H., & Mansur, M. (2008). *Penilaian hasil belajar*. Bandung: CV Wacana Prima.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207–230. https://doi.org/10.3102/10769986004003207

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

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Rohmawati, R. (Ed.). (2013). Kurikulum 2013, 87 persen guru kesulitan cara penilaian. Retrieved January 6, 2018, from https://unnes.ac.id/berita/87-persen-guru-kesulitan-soal-penilaian-kurikulum-2013.html

Sanjaya, W. (2010). *Kurikulum dan pembelajaran*. Jakarta: Kencana.

Santoso, A. (2013). Pemilihan butir alternatif pada tes adaptif untuk peningkatan keamanan tes. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran*, *43*(1), 1–8. https://doi.org/10.21831/jk.v43i1.1953

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

Surya, A., & Aman, A. (2016). Developing formative authentic assessment instruments based on learning trajectory for elementary school. *REiD (Research and Evaluation in Education)*, *2*(1), 13–24. https://doi.org/10.21831/reid.v2i1.6540

Werheid, K., Hoppe, C., Thone, A., Muller, U., Mungersdorf, M., & von Cramon, D. Y. (2002). The adaptive digit ordering test clinical application, reliability, and validity of a verbal working memory test. *Archives of Clinical Neuropsychology*, *17*(6), 547–565. https://doi.org/10.1093/arclin/17.6.547

Zaman, A., Kashmiri, A.-U.-R., Mubarak, M., & Ali, A. (2008). Students ranking, based on their abilities on objective type test: Comparison of CTT and IRT. *Edu-Com International Conference*, 591–599. Retrieved from https://ro.ecu.edu.au/ceducom/52/

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* or *reid.ppsuny@gmail.com*