International Journal of Instruction *October 2019 ● Vol.12, No.4*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp.OnlineFirst

Received: 21/01/2019
Revision: 21/05/2019
Accepted: 27/05/2019
OnlineFirst:23/07/2019

Construction of Test Instrument to Assess Foreign Student's Competence of Indonesian Language through Objective Test

Laili Etika Rahmawati
Universitas Sebelas Maret, Universitas Muhammadiyah Surakarta, Indonesia, *laili_etika@student.uns.ac.id, laili.rahmawati@ums.ac.id*

Sarwiji Suwandi
Universitas Sebelas Maret Surakarta, Indonesia, *sarwijiswan@staff.uns.ac.id*

Kundharu Saddhono
Universitas Sebelas Maret Surakarta, Indonesia, *kundharu_s@staff.uns.ac.id*

Budhi Setiawan
Universitas Sebelas Maret Surakarta, Indonesia, *kaprodipbi@staff.uns.ac.id*

Evaluation on the foreign student's competence of Indonesian language is required to assess their achievement. This research aimed to identify the difficulty level of test instrument, analyze the difference of foreign students' competence for Indonesian language and to analyze the correlation of students' competence over various objective test formats. This research was carried out between August 2017 to August 2018 to the 16 students of Muhammadiyah Universities involved in the Darmasiswa scholarship and the developing countries cooperation program. The formats of objective test were: true-false, directed gap filling, multiple choice, correct-incorrect, gap filling and matching. Kruskal-Wallis and bivariate correlation were carried out as the statistical analysis. The difficulty level of test instrument was between the very easy to moderate level. As the impact, students' score achievement was ranging from medium to high. Significant difference was identified both in the difficulty level and score achievement.

Keywords: achievement, difficulty, correlation, objective, score, test formats

INTRODUCTION

Language is an important tool to develop communication between individuals (Armstrong & Ferguson, 2010). There are approximately one thousand languages been identified all over the world (Ronen et al., 2014). Language is used as a tool to deliver information, while in the social life it is used to improve the interactions. However, in a multilingual community, transfer of knowledge is difficult to carry out (Fletcher-Chen,

Citation: Rahmawati, L. E., Suwandi., S., Saddhono, K., & Setiawan, B. (2019). Construction of Test Instrument to Assess Foreign Student's Competence of Indonesian Language through Objective Test. *International Journal of Instruction*, 12(4), .

2015). Indonesian language has two important values, including as communication tool and as the identity of the country (Nugraheni, 2015). Indonesian is the mandatory language in the national educational activities in Indonesia. Thus, any foreign students should have the competence in Indonesian.

The number of foreign students in Indonesia has increased since the first time Dharmasiswa program was established (Poedjiastutie, 2009). Dharmasiswa is a program organized by the Minister of Education and Culture to provide a short non degree scholarship to the foreign students which countries had established diplomatic relationship (Wilujeng, 2015). Students from various countries and cultural backgrounds come to Indonesia to learn Indonesian language and cultures. As the consequences, language teaching should include cultural aspect as the part of teaching materials (Saddhono, 2018).

Nowadays, Indonesian language is one of the popular languages learned in many countries as a second language (Junpaitoon, 2017). BIPA is a learning program which is specialised for the foreign speaker (Kusmiatun, 2016). The importance of BIPA has increased along with the increasing interest of foreign learners in Indonesian language (Suyoto, 2016). Currently, there are 251 institutions in 22 countries which provide BIPA training.

There are several motivations of foreign students to learn Indonesian, including: to understand Indonesian language, to conduct scientific research, to work and live in Indonesia (Suyitno, Susanto, Kamal, & Fawzi, 2017). BIPA plays important roles in connecting the learners and the local community. Thus, BIPA learning is carried out by referring various life apects, including: 1) historical-phylosophical; 2) geographical; 3) demographic; 4) economics; 5) political; 6) socio-cultural; and 7) global phenomena. Evaluation on the content of BIPA textbook has been carried out to improve its quality (Saddhono, 2018). However, evaluation on the competence test instrument is rarely carried out.

CONTEXT AND REVIEW OF LITERATURE

Language competence is an important aspect in communication (Armstrong & Ferguson, 2010). Previous research showed that language competence could improve the performance of a company (Ward, 2010). In the social interaction, language competence is a mean to develop politeness which could be evaluated through the identification of various maxims in a conversation (Sari, 2018). Theoretically, there are three main competencies of languages, including grammatical, sociolinguistic, and strategic competencies. However, sociolinguistic competence is the most crucial in the communicative competence (Sandlund, Sundqvist, & Nyroos, 2016). Language competence test involves various components, including vocabulary, grammar and pronunciation (Chiedu & Omenogor, 2014).

Test is an important process to measure the learning achievement. However, instead of the students' achievement, the performance of teacher is also important. Thus, test instrument should be well constructed so it would be valid, reliable and have appropriate difficulty (Lebagi, Nadrun, & Darmawan, 2014). A reliable, valid, practical and fair test

instrumet needs to be developed to provide a qualified assessment tool of Indonesian language competence.

Every foreign student may achieve different level of competence of Indonesian regardless their country of origin and the duration of their study (Wilujeng, 2015). In order to provide appropriate information regarding the result of BIPA learning, development of appropriate competence test instrument is required. This research aimed to identify the difficulty level of test instrument, analyze the difference of foreign students' competence for Indonesian language and to analyze the correlation of students' competence over various objective test formats.

METHOD

This research was carried out in three Muhammadiyah Universities in Indonesia which provide Darmasiswa scholarship and developing countries cooperation program in August 2017 to August 2018. The subjects of this research were the foreign students who were studying in the selected universities, while the object of the research was the test instrument of competence. The instrument was developed based on the *Criterion-Referenced Language Test Development (CRLTD) Workshop*. The test instrument was examined for its validity and reliability by some experts. The experts included were four lecturers specialised in BIPA from four different universities. The experts were selected after having discussion with the supervisors by considering their experience. The instruments was examined with delphi method. During the research, 16 students from the selected universities were involved in this research.

The research was carried out through an experiment. Data collection on the competence of foreign students was measured by a test. The formats of test acted as the treatment, while the test scores acted as the result. Construction of test instrument involves two test types, the supply and selection tests. There were six formats of test structure carried out in the research, including: 1) alternate response test by true – false options; 2) directed gap filling; 3) multiple choice; 4) alternate response test by correct – incorrect options; 5) gap filling; and 6) matching. Each test consisted of one reading topic followed by five related questions. Thus, total number of questions for the whole experiment was 30 items.

Data analysis was carried out to identify the difficulty level of test items, difference of difficulty level among test formats, difference of students' competence for Indonesian language over different test formats, and the correlation between students' competence over various test formats. However, since the distribution of obtained data were not normal, non parametric data analysis was carried out. Statistical analysis for the difference was conducted by Kruskal-Wallis, while post hoc analysis was conducted through pairwise comparison with Mann-Whitney U. Bivariate correlation analysis was conducted with Kendall's tau. Difficulty level of test formats was categorised into 5 levels, including: very easy ($d \leq 20\%$); easy ($20 < d \leq 40\%$); moderate ($40\% < d \leq 60\%$); hard ($60\% < d \leq 80\%$); and very hard ($d > 80\%$). The student's achievement was also categorised into 5 levels, including: very low (score ≤ 1); low ($1 <$ score ≤ 2); medium ($2 <$ score ≤ 3); high ($3 <$ score ≤ 4); and very high (score > 4).

FINDINGS

The difficulty level of test instrument was the first factor to be analysed. Level of difficulty was calculated based on the proportion of wrong answer to the total respondents of each test item. The questions were categorized to three levels of difficulty, including easy, moderate and hard. Analysis result on the distribution of competence test's difficulty level is presented in table 1.

Table 1
Proportion of Test Difficulty Level

No.	Test Format	Proportion of Test Difficulty Level (%)		
		Easy	Moderate	Hard
1.	True / False	80	20	0
2.	Gap Filling (Directed)	100	0	0
3.	Multiple Choice	40	40	20
4.	Correct / Incorrect	40	60	0
5.	Gap Filling	80	20	0
6.	Matching	100	0	0
	Overall proportion	73.33	23.33	3.33

Analysis on the difficulty level of test instruments showed that overall the items, 73.33% was categorized as easy, 23.33% was moderate, and only 3.33% was hard. However, the difficulty was varied among test formats. The test with multiple choice format was the only test with hard item category, while all items for the test formats of directed gap filling and matching were easy. Statistical analysis to further understand the difference of difficulty level among test formats is presented in table 2.

Table 2
Difference of Difficulty Level among Test Formats

No.	Test Format	Range of Difficulty Level (%)	Category Range (By Item)	Average Difficulty (%)	Category (Average)
1.	Gap Filling (Directed)	0.00 – 6.25	Very Easy	3.75a	Very Easy
2.	Matching	0.00 – 25.00	Very Easy – Easy	13.75ab	Very Easy
3.	True / False	6.25 – 37.50	Very Easy – Easy	15.00ab	Very Easy
4.	Gap Filling	12.50 – 43.75	Very Easy – Moderate	21.25b	Easy
5.	Correct / Incorrect	0.00 – 50.00	Very Easy – Moderate	27.50ab	Easy
6.	Multiple Choice	0.00 – 93.75	Very Easy – Very Hard	41.25ab	Moderate

Notation: column with similar letters indicates the insignificant difference

Table 2 shows that the difficulty level of the test formats applied in this research was ranging from very easy to moderate. However, item based analysis was ranging from very easy to very hard. Multiple choice test was the only test format which had the most variable difficulty level. Statistical analysis for all test fromats with Kruskal-Wallis showed chi-square value of 9.736 and probability 0.083. Further analaysis was carried out with Mann-Whitney test to analyze partial comparison between test formats. As the result, significant difference on the difficulty level was only obtained between the directed gap filling and ordinary gap filling formats. The analysis resulted the M-W value of 0.000 and probability 0.008 which indicates significant different of both test formats. Average difficulty level of directed gap filling test format was only 3.75% (very

easy), while the ordinary gap filling provided difficulty level of 21.25%. However, the average difficulty level of gap filling test format was not the highest. Insignificant difference of four other test formats toward directed gap filling and ordinary gap filling test formats was related to the distribution of difficulty levels.

Data analysis on the student's test achievement showed that there are variations on the frequency distribution. Analysis was based on the number of correct answer. If all questions were answered correctly, the score was 5 (very high), while if none of them was correct, the scored was 0 (null). Detailed analysis result of score frequency distribution is presented in table 3.

Table 3
Frequency Distribution of Students' Objective Test Achievement

Test Achievement	Frequency (Number of Students)						
	True/ False	Directed Gap Filling	Multiple Choice	Correct/ Incorrect	Gap Filling	Matching	Overall
Very High	7	14	1	4	5	10	7
High	6	1	4	5	6	3	9
Moderate	3	1	6	5	4	2	0
Low	0	0	3	1	1	0	0
Very Low	0	0	2	1	0	1	0
Null	0	0	0	0	0	0	0
Total	16	16	16	16	16	16	16

Table 3 shows that every student succeeded to provide correct answer for each test format. Unfortunately, the data showed that there were several students which obtained low (and very low) achievements, such as for the multiple choice, correct/incorrect, gap filling, and matching tests. However, overall result showed that there were only two levels of achievements, including the very high and high. This indicates that students' competency toward the objective test was variable. Some students may be superior at certain tests but inferior at the remaining tests.

Valuation on the student's competence over various test formats was relied on the individual score achievement. The score represents the correct answer obtained by each student for each test formats. Thus, the range was between 0 to 5. Table 3 shows the score distribution obtained from the experiment along with the average score for each test formats as well as the statistical analysis result.

Table 4
Difference in Student's Score Achievement among Test Formats

No.	Test Format	Score Range	Average Score	Category
1.	Multiple Choice	1 – 5	2.94a	Medium
2.	Correct / Incorrect	1 – 5	3.63ab	High
3.	Gap Filling	2 – 5	3.94bc	High
4.	True / False	3 – 5	4.25bc	Very High
5.	Matching	1 – 5	4.31cd	Very High
6.	Gap Filling (Directed)	3 – 5	4.81d	Very High

Notation: column with similar letters indicates the insignificant difference

Table 4 shows that the distribution of score was varied among test formats. The multiple choice, correct/incorrect and matching test formats had the widest range, including 1 as the lowest to 5 as the highest. The true/false and directed gap filling test formats had the narrowest score range with 3 as the lowest to 5 as the highest. Among the test formats, multiple choice test format resulted the avarage score below 3 which indicated moderate students' competence. The correct/incorrect and gap filling had the average scores within the range of 3 to 4 which indicated high students' competence. Overall the test formats, very high students' achievement was obtained from the true/false, matching and dirrected gap filling test formats with average score of 4.25, 4.31 and 4.81 (within the range 4 to 5).

Statistical analysis with Kruskall Wallis showed that there was significant difference on students' competence among test formats. Chi-square value resulted from the analysis was 29.117 and with the probability of 0.000. Post hoc analysis showed there were four groups of students' competence. Table 3 shows detailed difference on students' competence among tes formats.

Correlation analysis was carried out to evaluate the relationship of students' achievement between test formats. The capability of the students in working on certain test format could be related to another test formats. By understanding the relationship pattern, improvement of students' competence could be carried out concurrently. Table 5 showed the correlation of foreign students' competence of Indonesian language over various test formats.

Table 5
Correlations in Students' Score Achievement between Test Formats

Test Format	Coefficient of Correlation				
	Gap Filling (Directed)	Multiple Choice	Correct / Incorrect	Gap Filling	Matching
True / False	0.165	0.046	0.619*	0.071	-0.475*
Gap Filling (Directed)	##	0.286	0.268	-0.197	0.220
Multiple Choice	##	##	0.159	-0.022	0.037
Correct / Incorrect	##	##	##	0.284	-0.269
Gap Filling	##	##	##	##	-0.126

Notation: * indicates significant correlation

Analysis of correlation showed that there were only two significant relationships among the students' competence over test formats. The true/false test format was significantly related to the correct/incorrect test format with the correlation coefficient as much as 61.9%. The true/false test format was also related to the matching test format. However, the correlation was negative with the coefficient as much as -47.5%. The results indicated that the correlation between true/false and correct/incorrect test format was strong, while the correlation between the true/false and matching test formats was fair.

DISCUSSION

Test Difficulties

An assessment of students' achievement is also act as a feedback for the teacher's performance (Jandaghi, 2011). The low achievement of the students is considered as a failure of teaching performance. Thus, based on the analysis result as shown in Table 3, students are generally weak at multiple choice test format and strong at directed gap filling and true/false test formats. This implicates that the teacher should evaluate the teaching methods in order to improve the students' capability in identifying the distractor in the multiple choice test format.

The result of this research implicate that the instrument for language competence testing can be grouped into four level of difficulties. Directed gap filling and matching test formats can be grouped to the level of "very easy", since all the question items are categorised as easy. True / false and gap filling test formats are grouped to the level of "easy", since 80% of the test items are categorised as easy, while the other 20% are moderate. Correct / incorrect test format is grouped into the level "moderate", since 40% of the questions are categorised as easy and 60% are moderate. While multiple choice test format is grouped to the level "hard", since 20% of the items are categorised as hard.

The recommended range of difficulty level of a test instrument is variable. However, suggested difficulty level of objective test is between 30%-70% (Fourie, Summers, & Zweygarth, 2010). Unfortunately, the average difficulty level of objective test in this research showed the index between 3.75%-41.25%. Only multiple choice test format had the average difficulty index over 30%. However, item based analysis showed the difficulty index of some test items with >30% in matching, true/false, gap filling, and correct/incorrect test formats.

Language testing or assesment is used to understand the competence of the user (speaker) for communication (Chiedu & Omenogor, 2014). Thus, evaluation on the achievement of foreign students competence for Indonesian language is required to assess the result of their study. The assessment could be conducted through subjective or objective test. However, this research focused on the objective test in order to obtain a well constructed test instrument.

Objective test is a method to evaluate the achievement of a student (learner) toward certain knowledge. In the objective test, the testee needs to identify the correct answer among a number of alternatives (Igbojinwaekwu, 2015). The advantages of objective test are: rapid, reliable and repeatable. However, to conduct an objective test, certain standard which requires subjective judgement and valuation is required (Singham, Birwal, & Yadav, 2015).

Objective test is considered as reliable method for the assessment of language competence. In the objective test, test items are independent whereas the capability to answer one question correctly is not related to another question (Fox, 2012). In the objective test, each format of test is considered to have certain difficulties (Adebule,

2009). However, the difference may not be significant. Objective test has several advantages, such as: easy to develop, easy to administer, easy to score, efficient, credible, and effective for factual knowledge (Stecher et al., 1997). Instead of its advantages, objective test also has some disadvantages, such as incapable to explore the student's ideas (Lebagi et al., 2014).

Based on the result, there was a difference in the difficulty level of objective test from the instrument developped in this research. In this research, the index of difficulty shorted from the lowest to highest was directed gap filling, matching, true / false, gap filing, correct / incorrect and multiple choice. Similar result was obtained from the previous research which showed that multiple choice test format is more difficult than the true / false test format (Adebule, 2009). Another research also showed that the difficulty level of completion (gap filling) test format was higher than the matching test format although the difference was not significant (Osundare & Omirin, 2016). Similar result was also obtained from this research.

The difficulty level of each test formats resulted in this research showed that the test instrument has an appropriate qualification for the evaluation on the foreign student's competence of Indonesian language. Difficulty leveling is required in the subjective or objective tests, because it shows the quality of the test (Lebagi et al., 2014). Among the formats of objective tests, multiple choice is considered as the most flexible and useful (Alade & Omoruyi, 2014). However, the result of this research showed that it is the most difficult test format. The difficulty could be caused by the existence of distractions among the answers (Mahjabeen et al., 2017). Thus, it requires exact knowledge to answer the questions.

Difficulty index of test instrument is purposed to identify various teaching/learning aspects, such as: evaluation on the teaching materials and evaluation on student's strength/weakness over the taught materials (Johari et al., 2011). Based on the outcome of an assessment, the teacher needs to evaluate or modify the method or materials taught to the students. In the other side, the students can evaluate their weaknesses and modify their learning method to improve their competences.

Function of Test Formats

The completion test format (gap filling) has the function as a tool to measure the vocabulary of the students which is characterised as context-dependent test (Bagheridoust & Karagahi, 2013). Thus, instead of knowing the correct answer, the testee should also choose the correct word regarding the topic of test. In order to obtain a good achievement on the completion test format, the testee should understand the meaning and its appropriateness of a word to be used. The context is important aspect in the instrument of completion test format.

Alternate response test formats (true/false and correct/incorrect) are purposed to evaluate students competence in understanding the concept of the questions (Hubbard, Potts, & Couch, 2017). The alternate response test formats are considered as the bridge between the multiple choice and free response test formats, because the testees are asked to choose an answer but also identify the conception of the statement provided. Since

alternate response test formats are nearly similar to multiple choice test format, the result should not be significantly different. Previous research even showed that score achievement of the testee on the multiple choice and correct/incorrect test format was similar (Temel, Özgür, & Yilmaz, 2012). Even though the result is not exactly the same, this research also indicate that score obtained from both test formats were consecutive. The alternate response test formatis related to the memorization (Orluwene & Otuata, 2017).

Matching test format is context-independent test which is suggested as one of the best format for the assesment of vocabulary (Bagheridoust & Karagahi, 2013). Thus, although both the matching test format and completion test format have the function to measure the vocabulary, the performance ot the testee in matching test format is generally better than the completion test format. Similar result was obtained in this research whereas the difficulty of gap filling test format was higher than the matching test format. However, the directed gap filling had lower difficulty because the questions were provided with options of the answer.

Objective test is consisted of various test formats, such as multiple choice, gap filling, matching and alternate response. However, there might be possibility that the testee's competence in one test format is related to another test formats. A test format could be developed from another test format, such as the alternate response test format which apply the same priciple with the multiple choice test format (Hubbard et al., 2017). It could also be developed by differing the dependency to the context such as matching and completion test formats (Bagheridoust & Karagahi, 2013). Thus, the test formats within the objective test could be considered as gradual test construct.

Even though the basic principle of some test formats are related to another formats, the outcome could be significantly different. However, some formats of test may be correlated each other. As the result of this research, a significant correlation was obtained between the true/false test format and correct/incorrect test format as well as the true/false test format and matching test format. However, the correlation between the true/false and matching test formats was negative. It proves that objective test could be correlated each other between test formats. Previous research also proved that there are significant correlation between some test formats, such as multiple choice, alternate response and completion (Orluwene & Otuata, 2017). However, the relationship could be different among field of knowledge.

Even though true/false and correct/incorrect test formats basically have similar format, but the correlation was not strong enough. Generally, the mistakes in the alternate response test format such as true/false and correct/incorrect are caused by a misconception of the testee toward the question (Hubbard et al., 2017). Negative weak correlation between the true/false and matching test formats indicates that the competence of a language is not linear toward its aspects.

Students' Competence

Basically, Indonesian language for foreign students (BIPA) is devided into three levels of competence, including elementary (basic), intermediate, and advanced (Yahya,

Andayani, & Saddhono, 2018). Placement test based on their Indonesian language competence is carried out before the course is started. Thus, the students may achieve different level of initial advancement. Further, the students are grouped based on their competence level. However, this research was carried out to the students with basic level of Indonesian language competence.

Regardless the respective result of objective test formats carried out to the students, only two group of achievement were obtained from the cummulative score, including very high and high groups. This result shows that generally the students have appropriate competence of Indonesian language. However, each student has weaknesses at some particular aspect of the language, such as vocabulary for completion and matching test formats (Bagheridoust & Karagahi, 2013) or conception for multiple choice and alternate response test formats (Hubbard et al., 2017).

In the language competence test, the testee needs to understand the context of the questions. Within the BIPA course, various teaching materials are involved in order to develop the students' general knowledge instead of the language. The differences of the test result could be affected by the topic of discussion. In BIPA teaching, various topic could be selected, such as introduction, daily activity, transportation, vacation, etc (Ningsih, Rasyid, & Muliastuti, 2018). Previous research showed that the students preferred tourism topic for the BIPA learning. Culinary, culture, art and entertainment are alternative topic which are preferred. However, the level of preference was not as much as the tourism topic (Kusmiatun, 2016).

Regardless the result of competence test, the achievement of the students in learning Indonesian is also influenced by the teaching and learning methods. The student's learning strategy is one of the influencing factor to the achievement of their competence. Previous research showed that foreign students in the Dharmasiswa program have various learning method to improve their experiences and achievements (Wilujeng, 2015). Another research showed that integrative learning could increase students' understanding on both language and culture (Andayani, 2016).

A good language test should have some characteritics, such as: reliable, valid, practical and fair (Chiedu & Omenogor, 2014). Thus, evaluation and improvement of test instrument needs to be carried out periodically. In order to improve the quality of objective test, evaluation of test instrument to the item level is suggested (Siri & Freddano, 2011). Test-level and item-level assessment may show difference in the measurement comparability. Previous research showed that at item-level the degree of comparability was between moderate to low, while at test-level the degree of comparability was high (Oliveri, Olson, Ercikan, & Zumbo, 2012). Item analysis is important to determine the quality of test instrument (Suruchi & Rana, 2014). Moreover, it also has the function to identify the defective test items and the mastery of individual testee on the examined topic.

The finding of this research implicates that the students have various types of capability regarding Indonesian language. Thus, in order to improve their capability, improvement of teaching method by the teacher/lecturer is required (Johari et al., 2011) as well as the

learning method by the students (Wilujeng, 2015). The instrument examined in this research is appropriate to assess the students competence, both in general term and specific term regarding Indonesian language. The distribution of students' performance of respective test formats proved that the instrument could act as the identifier of their strengths and weaknesses in Indonesian language learning.

CONCLUSION

Based on the test format, the test instrument constructed for this research showed the average difficulty level ranging from easy to moderate with significant difference obtained from the directed gap filling and gap filling test formats. The students' competence on Indonesian language was between medium to very high, whereas some significant differences was obtained among test formats. Significant positive correlation of students' competence over test formats was obtained between the true/false and correct/incorrect test formats, while true/false and matching test formats showed negative significant correlation. This research implies that the test instrument only represented four difficulty levels ranging from very easy to hard and the students' competence in Indonesian language is varied ranging from moderate to very high. Further improvement on the test instrument is required, especially in order to provide appropriate range of the test difficulty level.

REFERENCES

Adebule, S. O. (2009). Reliability and levels of difficulty of objective test items in a Mathematics achievement test: A study of ten senior secondary schools in five local government areas of Akure, Ondo State. *Educational Research and Review*, *4*(11), 585-587.

Alade, O. M., & Omoruyi, I. V. (2014). Table of specification and its relevance in educational develoopment assessment. *European Journal of Educational and Development Psychology*, *2*(1), 1-17.

Andayani. (2016). Improving the language skills and local cultural understanding with integrative learning in teaching Indonesian to speakers of other languages (TISOL). *International Journal of Language and Linguistics*, *3*(2), 44-53.

Armstrong, E., & Ferguson, A. (2010). Language, meaning, context, and functional communication. *Aphasiology*, *24*(4), 480-496.

Bagheridoust, E., & Karagahi, M. (2013). The effect of context-dependent and context-independent test design on Iranian EFL learners' performance on vocabulary tests. *International Research Journal of Applied and Basic Sciences*, *4*(8), 2129-2136.

Chiedu, R. E., & Omenogor, H. D. (2014). The concept of reliability in language testing: Issues and solutions. *Journal of Resourcefulnes and Distinction*, *8*(1), 1-9.

Fletcher-Chen, C. (2015). Impact of language diversity and social interaction on knowledge transfer. *US-China Education Review A*, *5*(3), 159-180. http://doi.org/10.17265/2161-623X/2015.03.001.

Fourie, S., Summers, B., & Zweygarth, M. (2010). Difficulty and discrimination indices as quality assurance tools for assessments in a South African problem-based pharmacy programme. *Pharmacy Education*, *10*(2), 119-128.

Fox, J. (2012). Language assessment methods. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-6). Oxford, UK: Blackwell Publishing Ltd. http://doi.org/10.1002/9781405198431.wbeal0606.

Hubbard, J. K., Potts, M. A., & Couch, B. A. (2017). How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. *CBE-Life Sciences Education*, *16*(ar26), 1-13. http://doi.org/10.1187/cbe.16-12-0339.

Igbojinwaekwu, P. C. (2015). Effectiveness of guided multiple choice objective questions test on students' academic achievement in senior school mathematics by school location. *Journal of Education and Practice*, *6*(11), 37-49.

Jandaghi, G. (2011). Assessment of validity, reliability and difficulty indices for teacher-built physics exam questions in first year high school. *Arts and Social Sciences Journal*, *16*(1), 1-4.

Johari, J., Sahari, J., Abd Wahab, D., Abdullah, S., Abdullah, S., Omar, M. Z., & Muhamad, N. (2011). Difficulty index of examinations and their relation to the achievement of programme outcomes. *Procedia-Social and Behavioral Sciences*, *18*, 71-80. http://doi.org/10.1016/j.sbspro.2011.05.011.

Junpaitoon, P. (2017). Enrichment of vocabulary in BIPA learning for beginner Thai students. *Journal of Innovative Studies on Character and Education*, *1*(1), 88-103.

Kusmiatun, A. (2016). Topik pilihan mahasiswa Tiongkok dalam pembelajaran BIPA Program Transfer Kredit di UNY. *Litera*, *15*(1), 138-146.

Lebagi, D., Nadrun, & Darmawan. (2014). Analizing difficulty level of subjective test used by an English teacher. *English Language Teaching Society*, *2*(2), 1-14.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of Pakistan Institute of Medical Sciences*, *13*(4), 310-315. Retrieved from http://www.indianjournals.com/ijor.aspx?target=ijor:ijone&volume=9&issue=3&article=024.

Ningsih, S. A., Rasyid, Y., & Muliastuti, L. (2018). Analisis kebutuhan materi ajar membaca BIPA A1 dengan pendekatan deduktif di SD D'Royal Moroco. *Pembelajar*, *2*(2), 85-91. http://doi.org/10.26858/pembelajar.v2i2.5974.

Nugraheni, A. S. (2015). Pengembangan program profesionalisme dosen pengajar Bahasa Indonesia untuk Penutur Asing (BIPA) di ASEAN. *AL-BIDAYAH*, *7*(1), 89-101. Retrieved from http://digilib.uin-suka.ac.id/25203/.

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for

investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, *12*(3), 203-223. http://doi.org/10.1080/15305058.2011.617475.

Orluwene, G. W., & Otuata, E. A. (2017). Gender influence on the efficacy of multiple-choice, alternate response and completion objective test formats non students achievement in economics. *European Journal of Research and Reflection in Educational Sciences*, *5*(1), 71-80.

Osundare, A. G., & Omirin, M. S. (2016). Determining the differences in the difficulty and discriminating indices of chemistry completion and matching test formats. *European Journal of Research and Reflection in Educational Sciences*, *4*(2), 65-70.

Poedjiastutie, D. (2009). Culture shock experienced by foreign students studying at Indonesian universities. *TEFLIN Journal*, *20*(1), 25-36. http://doi.org/10.15639/TEFLINJOURNAL.V20I1/25-36.

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *PNAS*, ES616-ES622. http://doi.org/10.1073/pnas.1410931111.

Saddhono, K. (2018). Cultural elements in the Indonesian textbooks as a foreign language (BIPA) in Indonesia. *KnE Social Sciences*, *3*(9), 126-134. http://doi.org/10.18502/kss.v3i9.2619.

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass*, *10*(1), 14-29. http://doi.org/10.1111/lnc3.12174.

Sari, Y. (2018). Wujud kesantunan berbahasa mahasiswa asing program darmasiswa di Universitas Gadjah Mada. *Jurnal Gramatika*, *4*(1), 118-128.

Singham, P., Birwal, P., & Yadav, B. K. (2015). Importance of objective and subjective measurement of food quality and their inter-relationship. *Journal of Food Processing & Technology*, *6*(9), 1000488. http://doi.org/10.4172/2157-7110.1000488.

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia-Social and Behavioral Sciences*, *29*, 188-197. http://doi.org/10.1016/j.sbspro.2011.11.224.

Stecher, B. M., Rahn, M. L., Ruby, A., Alt, M. N., Robyn, A., & Ward, B. (1997). *Using alternative assessments in vocational education*. Washington D.C.: RAND. Retrieved from https://www.rand.org/content/dam/rand/pubs/monograph_reports/2007/MR836.pdf.

Suruchi, & Rana, S. S. (2014). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in biology. *Paripex Indian Journal of Research*, *3*(6), 56-58. http://doi.org/10.15373/22501991.

Suyitno, I., Susanto, G., Kamal, M., & Fawzi, A. (2017). Cognitive learning strategy of

BIPA students in learning the Indonesian language. *IAFOR Journal of Language Learning*, *3*(2), 175-190.

Suyoto. (2016). The condition of Indonesian society in the perspective of BIPA program development. *The Journal of Kanda University of International Studies*, *28*, 327-342.

Temel, S., Özgür, S. D., & Yilmaz, A. (2012). The effect of different types of test on preservice chemistry teachers' achievement related to "chemical bonding." *Problems of Education in the 21st Century*, *41*, 123-129.

Ward, S. A. (2010). The road to foreign language competency in the United States: A leadership perspective. *Journal of Leadership Studies*, *4*(3), 6-22. http://doi.org/10.1002/jls.20173.

Wilujeng, N. C. S. (2015). How do the Darmasiswa students learn Indonesian language in Yogyakarta State University, Indonesia? *Journal of Education*, *8*(1), 10-15.

Yahya, M., Andayani, & Saddhono, K. (2018). Hubungan penguasaan kosakata dengan kesalahan diksi dalam kalimat bahasa indonesia mahasiswa BIPA level akademik. *Jurnal Kredo*, *1*(2), 53-70.