# Applying Item Response Theory In Validating An Indonesian Language Profeciency Test

M Sudaryanto[1], D Mardapi[2], S Hadi[3]
[1,2,3]Yogyakarta State University, Yogyakarta, Indonesia
[1]Sebelas Maret University, Solo, Indonesia

{[1]memet.sudaryanto2016@student.uny.ac.id, [1]memetsudaryanto@staff.uns.ac.id, [2]djemari.@uny.ac.id, [3]samsul.hd@gmail.com}

**Abstract.** The Indonesian language test is a measurement of competence in understanding the discourse of written and oral media. The understanding of examining the dimensions of the test needs to be analyzed to look for factors that can explain the relationship or correlation between the various independent indicators observed. This study investigated the application of Item Response Theory (IRT) in analyzing packages of Indonesian test for foreign speakers. The test consists of 3 competencies that measure the ability in listening, reading, and grammar. The analysis was carried out through two stages of testing, namely (1) testing the validation and estimation of test reliability, and (2) measuring the ability to examine with the Rasch Model to determine the level of difficulty of the item. The stages of testing are testing content validation through the Aiken index, testing the instrument dimensions, confirmatory factor analysis, KR-20 reliability estimation, estimation of alpha reliability from Cronbach, and estimation of reliability through item information function. The difficulty level of the test item is measured by IRT Model package of the freeware R. The results of the study indicate that the Indonesian language test consists of 3 dimensions that correlate to explain latent factors. Based on test validation testing, the Indonesian language test instrument measures the competency according to the test construction. Test reliability is estimated by three methods showing quite high results. Based on the reliability estimation the test shows that the instrument has the consistency of the test results indicated by a coefficient above 0.7.

***Keywords****: MIRT, Validation, Reliability, Indonesian Language Test*

## 1 INTRODUCTION

In addition to the function of communication, language has a central function as a cultural identity of society. Indonesia recognizes Indonesian as the national language, the language of unity, and the language of communication. These three functions underlie the importance of standardized and applicable Indonesian language both nationally and internationally. Moreover, currently Indonesian language is being pioneered as an international language in the Asian Region [1]. The main language identification is divided into four skills, namely, listening, speaking, reading and writing skills. Listening skills are closely related to speaking skills, as well as reading skills and integrated writing skills [2]. Basic language knowledge includes grammar, lexeme, and four language skills such as listening, speaking, reading, and

ICEL 2019, March 23-24, Malang, Indonesia

DOI 10.4108/eai.23-3-2019.2284880

writing, all of which must go hand in hand to support the internationalization process of Indonesian language [3].

One of the tools for internationalizing Indonesian language is an instrument used to measure the ability of examining. Foreign speakers who wish to settle in certain countries (with the aim of studying or working) are required to be proficient in using the language of purpose. Each country develops language tests according to their respective characteristics, such as TOEFL (English), DELE (Spanish), ALPT (Arabic), and TOCFL/TOP (Chinese) which are used to test the language skills of the country. The test is used to measure the competency of examinee before taking classes or working in an institution in the destination country [4].

Language assessment standards are not only the preparation of good items, but need to be specified in determining the competencies to be measured. Determining the boundaries of competency categories is one of the most important things in the development of tests, administration, and assessment reports [5]. In addition, a pattern that shows the competency of test participants and the level of difficulty of items according to [6] can be measured on the same scale. So far, language skills possessed by foreign citizens have not been standardized. Every foreign student with various skill levels who will enter Indonesia is finally accepted without minimum completeness criteria.

Listening competency tests measure ability in (a) paying attention to the right speech, words, diction, and sentence elements, (b) determining the reasons why, (c) understanding the various meanings of context instructions, (d) distinguishing facts and fantasies and those that are relevant and not relevant, (e) deciding, (f) drawing conclusions, (g) determining answers to specific problems, (h) determining new information or additional information on a topic, (i) translate, interpreting expressions, idioms, and languages that are not commonly used, and (j) act objectively and evaluatively to determine the authenticity of the truth or the existence of prejudice.

Reading is defined as an activity/process in which the reader controls the source of information, elaborates on meaning and strategy, monitors his understanding, and uses social context to reflect his response [2] The focus of testing language skills is to show effective sentences that show that the delivery process by the speaker or writer. The reception process is receptive by the partner to take place in a complete manner so that the purpose of the speech delivered can be captured as complete information.

### 1.1 Research Method

$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_j)}$$

*Where,*
*$\theta$ = ability*
*$b_i$ = difficulty parameter*

The principle of the rasch analysis model is the opportunity for the student to answer a single item correctly with the student's ability compared to the difficulty level of the item. [7]. [8] Explain when the ability value of a group is transformed, so the mean (average) becomes 0 and the standard deviation is 1, the value of bi ranges from -2 to +2. The bi value close to the -2 number means that the item is too easy, and the bi value is close to +2 so the item is too difficult. [9] Hambleton states that one logistics parameter is often referred to as Rasch. This study focuses on developing receptive skills test instruments, namely listening, reading, and applied skills are skills to respond to rules. The indicators and grids measured ultimately greatly influence the content validity by knowing the expert agreement on the item index [8].

Validity is proven by confirmation in developing objective test constructs. Each aspect is broken down into a number of indicators, which are compiled again into items that are assembled to get a complete instrument. The instrument was tested on foreign students to find out its usage and estimation of the reliability coefficient of the measurement results. Expert Judgment is the provision of input from experts on the validation of the contents of the instruments prepared. Content validity (aiken index) construct validity with CFA (Confirmatory Factor Analysis) instrument analysis using empirical data and analyzed with the help of Lisrel or MPlus application. In addition, the estimated reliability of the instrument is determined by calculating the Alpha correlation coefficient from Cronbach for the test results. The reliability through the IRT is estimated by the suitability of the characteristics of the items with the type and purpose of the test, which greatly determines the quality of the test items [10].

### 1.2 Result and Discussion

The results of testing the validity and reliability of Indonesian language test instruments show significance estimations. Indonesian language proficiency test for foreign speakers with 3 instruments namely tests of listening, reading, and responding to rules. Listening tests measure students' ability to know information in daily speech, descriptive speech, exposition, argumentation, and procedural. This test consists of 31 test items in the form of dialogue and monologue. Reading tests measure simple sentences, chronological sentences, information descriptions, persuasion, explanatory processes, persuasive sentences, narratives, and written discourse generalizations. This test consists of 40 test items. The rule response test measures grammar competence in word formation, word classification, and semantic relations. This test consists of 29 test items.

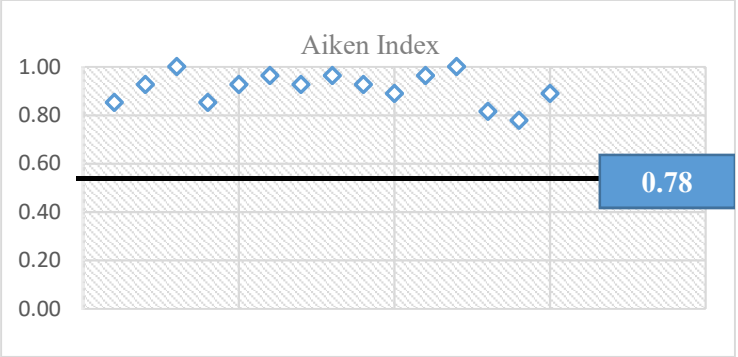
### 1.3 Validity Content Test with Aiken

Content validity indicates the test instrument reflects the complete range of attributes studied and is usually carried out by seven or more experts.

**Table 1.** Distribution of Rater Scoring toward Items.

| | R1 | R2 | R3 | R8 | R9 | S1 | S2 | S3 | S8 | S9 | ∑S | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MY_01 | 3 | 4 | 4 | 4 | 3 | 2 | 3 | 3 | 3 | 2 | 23 | 0.85 |
| MY_02 | 3 | 4 | 3 | 4 | 4 | 2 | 3 | 2 | 3 | 3 | 25 | 0.93 |
| MY_03 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 27 | 1.00 |
| MY_04 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 23 | 0.85 |
| MY_05 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 25 | 0.93 |
| MB_01 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 26 | 0.96 |
| MB_02 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 25 | 0.93 |
| MB_03 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 3 | 2 | 3 | 26 | 0.96 |
| MB_04 | 4 | 4 | 3 | 3 | 4 | 3 | 3 | 2 | 2 | 3 | 25 | 0.93 |
| MB_05 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 24 | 0.89 |
| KD_01 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 3 | 2 | 3 | 26 | 0.96 |
| KD_02 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 27 | 1.00 |
| KD_03 | 3 | 3 | 4 | 4 | 3 | 2 | 2 | 3 | 3 | 2 | 22 | 0.81 |
| KD_04 | 4 | 2 | 4 | 3 | 3 | 3 | 1 | 3 | 2 | 2 | 21 | 0.78 |
| KD_05 | 3 | 4 | 4 | 4 | 4 | 2 | 3 | 3 | 3 | 3 | 24 | 0.89 |

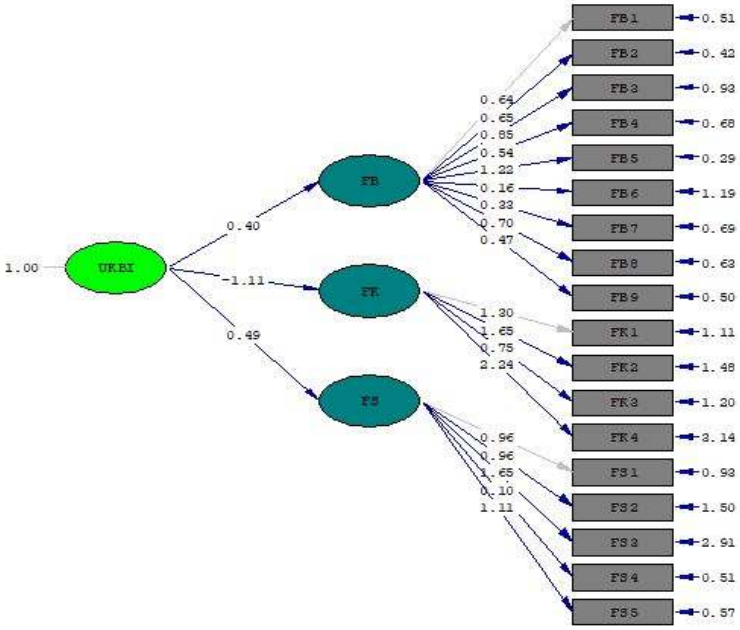
Aiken [11] determines the item validation index with the number of experts as many as 9 rater, and 4 rating options are 0.78. Based on the content validation test from Aiken, it was shown that each item of the Indonesian language test was valid. The emphasis of the content validation approach is one of expert or professional judgment. Proof of content validity in general does not define the measuring domain accurately because the sample of human behavior that is intended to be used as an item is quite a lot. The accuracy of content validity can be achieved if on the development of an instrument, the measuring domain is well defined and the instrument item is correctly written [12].




### 1.4 Confirmatory Factors Analysis

Confirmatory Factor Analysis (CFA) is known as a component test tool that is useful in finding the construct form of a collection of manifest variables, or testing a variable on the assumption of the manifest that built it. Confirmatory analysis is very suitable for testing a variable theory on the manifest or indicators that build it, where the variable is assumed to be only measurable with these indicators. Confirmatory factor analysis aims to evaluate patterns of relationships between several constructs. Some indicators build each construct. The confirmatory analysis model is usually not assumed to be the direction of the relationship between constructs, but only the correlative relationship between constructs.

One function of the CFA is to test construct validity, by confirming whether certain items are in the same indicator as other items so that they measure the same dimensions. Based on the validity test by confirming the analysis factor with the help of Lisrel 8.50 program on 3 components of the factors FB (Reading Factor), FK (Grammar Factor), and FS (Editing Factor). Estimates are the numbers in arrows; these are the results of rough data calculations. For further calculation the regression is calculated, or the parameters are calculated from the original data [13].




The procedure determines whether a theoretical model fit with the data is called the test of goodness of fit. There are quite a number of criteria used to determine whether a model is fit or not.

### 1.5 Construction Reliability Estimation

The construct reliability test is estimated by the output data on Lisrel 8.50 (the results of the Confirmatory factor analysis test), which shows the relationship between the observed variable and the latent variable. Assuming a congeneric scale [14] with a standardized latent construct, ω can be estimated as

$$CR = \frac{\left(\sum_{j=1}^{i} \lambda_i\right)^2}{\left(\sum_{j=1}^{i} \lambda_i\right)^2 + \left(\sum_{j=1}^{i} \delta\right)}$$

Where $\lambda_l$ represent the factor loading of item *i* onto a single common factor and θ represent the unique variance of item *i*. The value of Λ indicates the relationship of the variable, while the number in the δ measurement error shown by each error value. The estimation of construct reliability uses the loading factor for each indicator to compile and error index of each indicator. CR formula is calculated by formula.

**Table2.** Construction Reliability Estimation.

| | Λ | δ | | Λ | δ |
|---|---|---|---|---|---|
| FB1 | 0.67 | 0.55 | FK2 | 0.8 | 0.35 |
| FB2 | 0.71 | 0.5 | FK3 | 0.56 | 0.68 |
| FB3 | 0.66 | 0.56 | FK4 | 0.78 | 0.39 |
| FB4 | 0.55 | 0.7 | FS1 | 0.71 | 0.5 |
| FB5 | 0.92 | 0.16 | FS2 | 0.62 | 0.62 |
| FB6 | 0.14 | 0.98 | FS3 | 0.69 | 0.52 |
| FB7 | 0.37 | 0.86 | FS4 | 0.13 | 0.98 |
| FB8 | 0.66 | 0.56 | FS5 | 0.83 | 0.32 |
| FB9 | 0.55 | 0.69 | Totals | 11.13 | 10.32 |
| FK1 | 0.78 | 0.4 | | | |
| | | | $\sum \lambda^2$ | 123.88 | |
| | | | RK | 0.923 | |

Based on the tested construct, the Indonesian language test instrument consists of 3 components (listening, reading, and rule responses). The results of the estimated reliability of the test indicate that each indicator has a consistency of 0.92 if tested on a wide-scale test participant.

### 1.6 Alpha Reliability Test from Cronbach

To estimate the different items that measure the same construct used composite reliability. Composite in question is the combined score of the instrument compiler. One of the estimation tools used is the Alpha Formula from Cronbach.

$$\alpha = \left(\frac{K}{K-1}\right)\left(1 - \frac{\sum {\sigma_i}^2}{{\sigma_i}^2}\right)$$

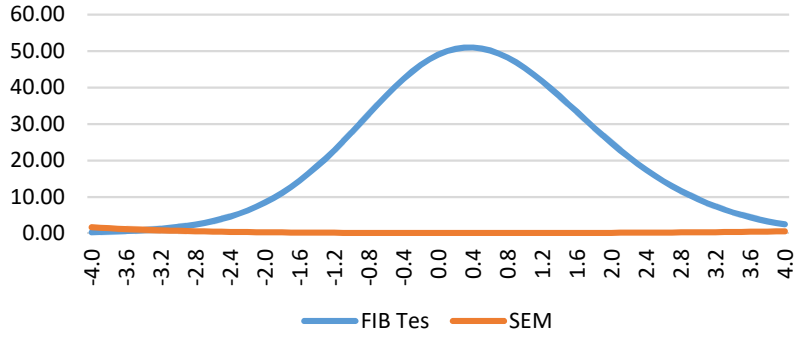
With α is the instrument reliability coefficient, k is the number of questions in the instrument, $\sum {\sigma_i}^2$ is the number of variances of the instrument, and ${\sigma_i}^2$is the total score variance.

**Table3**. Estimated Alpha Reliability from Cronbach.

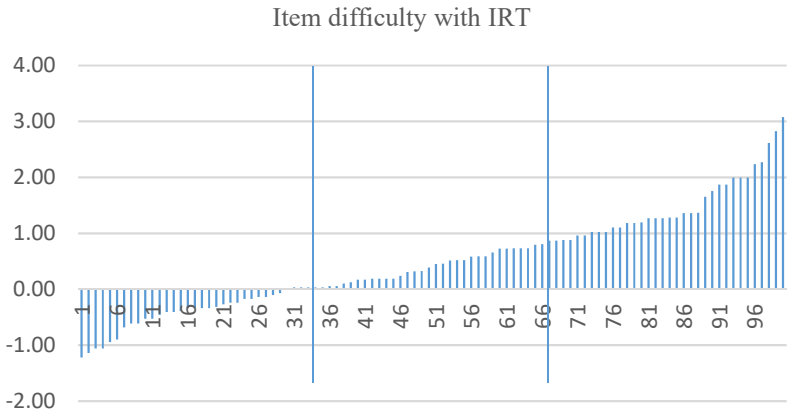
| Cronbach's Alpha | N of Items |
|---|---|
| .700 | 100 |

### 1.7 Test Information Function Reliability Test Item difficulty with MIRT

The test information function shows the ability of the item to explain the latent variables measured. Listening skill is important to be tested, namely to know and identify other skills, especially the development of ideas for speaking skills and writing skills. In addition, the essence of listening aside from self-expression also adds insight into which each language learner can get more than mere ingredients but also additional information in the form of science.




Reading is a means to receive information in written communication activities with the aim of obtaining information, capturing messages, and understanding meaning. The essence of using language must be authentic because language is a practical science and needs to be practiced. The more frequent and consistent use of good grammar, the more the language users do not have to bother to explore the use of standard language spelling.




## 2 CONCLUSIONS

The difficulty level of the test item is measured by IRT Model package of the freeware R. The results of the study indicate that the Indonesian language test consists of 3 dimensions that correlate to explain latent factors. Based on test validation testing, the Indonesian language test instrument measures the competency according to the test construction. Component test tool to find construct form of a collection of manifest variables, or testing a variable on the assumption of the manifest that built it shows P=0.339. That means, all the latent variable can be well measured through observed variables. The reliability test estimated by the KR-20 and the test information function produce identical results. In the other hands, the analysis show that one of the tools for internationalizing Indonesian language is an instrument used to measure the ability of examining is ready to use. Based on the qualitative analysis find an errors language test information in four fields, namely phonological errors, morphology, syntax,

and semantic fields. Some mistakes that must be understood by foreign speakers include: sentence ineffectiveness, word selection errors, affix use errors, incomplete sentence functions, prepositional misuse, word order reversal, use of passive construction, conjunctive usage errors, 'yang' usage errors, and errors in plural formation.

## REFERENCES


[1] J. Jackson, "Globalization, internationalization, and short-term stays abroad," *Int. J. Intercult. Relations*, 2008.

[2] E. Tschirner, "Listening and Reading Proficiency Levels of College Students," *Foreign Lang. Ann.*, 2016.

[3] F. Djajasudarma, "PERGESERAN PERAN BAHASA INDONESIA," *Ranah J. Kaji. Bhs.*, 2018.

[4] K. Saddhono, "Kajian Sosiolinguistik Pemakaian Bahasa Asing dalam Pembelajaran Bahasa Indonesia untuk Penutur Asing (BIPA)," *Kaji. Linguist. dan Sastra*, 2012.

[5] G. J. Cizek and M. B. Bunch, *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. 2007.

[6] D. Mardapi and B. Kartowagiran, "Pengembangan Instrumen Pengukur Hasil Belajar Nirbias dan Terskala Baku," *J. Penelit. dan Eval. Pendidik.*, 2019.

[7] R. P. Chalmers, "mirt : A Multidimensional Item Response Theory Package for the R Environment ," *J. Stat. Softw.*, 2015.

[8] R. K. Hambleton and R. W. Jones, "Comparison of classical test theory and item response theory and their applications to test development," *Educ. Meas. Issues Pract.*, 1993.

[9] R. K. Hambleton and H. Swaminathan, *Item response theory : principles and applications*. 1985.

[10] M. Djemari, *Pengukuran Penilaian & Evaluasi Pendidikan*. Yogyakarta: Nuha Media, 2012.

[11] L. R. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," *Educ. Psychol. Meas.*, 1985.

[12] H. Retnawati, "Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index)," *Res. Eval. Educ.*, 2017.

[13] M. Picard and L. Velautham, "Developing Independent Listening Skills for English as an Additional Language Students," *Int. J. Teach. Learn. High. Educ.*, 2016.

[14] G. J. Geldhof, K. J. Preacher, and M. J. Zyphur, "Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework," vol. 19, no. 1, pp. 72–91, 2014.