Jurnal Penelitian dan Evaluasi Pendidikan
Volume 25, No 1, June 2021 (97-107)
Online: http://journal.uny.ac.id/index.php/jpep

Applying Item Response Theory model for evaluating item and test properties of academic potential test for students with disability

Sukaesi Marianti*; Dian Putri Permatasari; Unita Werdi Rahajeng
Universitas Brawijaya
Jl. Veteran, Ketawanggede, Lowokwaru, Kota Malang, Jawa Timur 65145, Indonesia
*Corresponding Author. E-mail: s.marianti@ub.ac.id

ARTICLE INFO

Article History
Submitted:
16 February 2021
Revised:
28 June 2021
Accepted:
14 July 2021

Keywords
admission selection; disability; computer-based academic potential test; item response theory

Scan Me:

ABSTRACT

Universitas Brawijaya (UB) is one of the pioneers of inclusive education in higher education in Indonesia. One of the innovations in the policies related to inclusive education is affirmative action admissions special for students with disabilities, namely *Seleksi Mandiri Penyandang Disabilitas* (Independent Selection for Person with Disabilities), which focuses on accommodating admissions selection for students with disabilities who want to enroll in bachelors or vocational programs. A part of this admission selection is the test called the Computer-Based Academic Potential Test. This study aims to evaluate, from a psychometric perspective, the psychometric properties of the potential academic test. The approach used in this study is the item response theory (IRT) framework, which is mostly used for evaluating psychometric quality at both item-level and test levels. This study's IRT model is a two-parameter logistic model that includes difficulty parameter and discrimination parameter. The result of this study exhibited that the three subtests of the Computer-Based Academic Potential Test, in general, have satisfying results from the 2PL model estimation. The result also showed that most of the item difficulties ranged from medium to very difficult.

How to cite:
Marianti, S., Permatasari, D., & Rahajeng, U. (2021). Applying Item Response Theory model for evaluating item and test properties of academic potential test for students with disability. *Jurnal Penelitian dan Evaluasi Pendidikan, 25*(1), 97-107. doi:https://doi.org/10.21831/pep.v25i1.38808

INTRODUCTION

Since 2012, Universitas Brawijaya (UB) has developed a special admission system for students with disabilities: *Sistem Penerimaan Khusus Penyandang Disabilitas* or SPKPD (Special Admission System for Person with Disabilities). The system is part of affirmative action to address a problem where the education level attainment of people with disabilities is still much lower than those without disabilities. Initially, the system provided for people with disabilities to enter the university was still limited. Even if they want to participate in national selection into public universities, they must face less adaptive and less accommodating selection models for people with special needs. Palombi (2000) states that standardized test models often do not consider the special needs of people with disabilities, so there will be unfairness if regular test scores are used for the decision making in the admission of students with disabilities.

SPKPD is designed as a system that considers the interests of people with disabilities, which will then be adjusted to the departments and programs in UB. The SPKPD implementation is fully submitted to *Pusat Studi dan Layanan Disabilitas* (PSLD), a service center for students with disabilities in UB. Initially, PSLD developed SPKPD using only interview and observation methods. Administratively-qualified prospective students with disabilities are invited to participate in interviews and observations in several activity settings (Pratiwi et al., 2018).

https://doi.org/10.21831/pep.v25i1.38808

Since 2018, SPKPD has been modified by adding a Computer-Based Academic Potential Test. Some of the underlying considerations are (1) advice from program managers related to general standards that prospective students with disabilities must own, (2) the increasing number of prospective students with disabilities participating in the SPKPD makes the implementation of interviews or observations inefficient and impractical. Besides, the Computer-Based Academic Potential Test results are used as the basis for selecting prospective students to be interviewed by the manager of the selected program.

In 2019, UB Rector Regulation No. 33 of 2019 changed the term of SPKPD to SMPD, which stands for *Seleksi Mandiri Penyandang Disabilitas* (Independent Admission for Person with Disabilities). In general, the admission selection process does not change from the previous selection system, which aims to select and obtain information about the academic potential of prospective students. The underlying reason in the selection system is to provide opportunities for prospective students with disabilities by going through the selection process. Although prospective students with disabilities are not given the same selection test as regular prospective students, ideally, the quality of tests for both groups is psychometrically acceptable.

The computer-Based Academic Potential Test, as a part of the admission selection, is expected to represent the academic potential of prospective students with disabilities, characterized by mastery of basic academic abilities, including language and numeric. Therefore, the Computer-Based Academic Potential Test sub-test consists of Bahasa Indonesia, English, and Mathematics. A selection system involving a Computer-Based Academic Potential Test that is psychometrically feasible is one way to ensure the quality and readiness of prospective students with disabilities to study in college. Basically, to study in college, one must have a minimum requirement. Without the selection process, the university does not have enough information to know which students are ready and not academically ready to attend college education. In addition, if there are academically not ready students, it will be difficult to attend courses in college because they do not have the adequate basic academic ability.

Wolanin and Steele (2004) explained that in terms of admission of students with disabilities, each course must still consider the minimum academic requirements of prospective students. In general, without a well-designed selection model, prospective students with disabilities in universities are vulnerable to getting caught up in the charity model paradigm. In the charity model paradigm, prospective students with disabilities are the parties entitled to mercy (Rukmantara & Lesmana, 2018). Of course, the spirit is not in line with UB policy that opens the opportunity to study in universities for prospective students with disabilities as a form of social reconstruction and fulfillment of human rights equality.

A good quality test is a test that has good psychometric characteristics through a series of psychometric analyses to obtain evidence that it is feasible to use. One indication of psychometric feasibility is that the items function accurately and fairly to all test takers. A test that psychometrically functions optimally is a test that produces a score that truly represents the test taker's ability so the scores obtained from the test results can be used for decision making.

In order to create a well-design test, psychometric evaluation is inevitable. Evaluation of the psychometric characteristics of a test involves psychometric analysis to prove that a test is not very easy and not too difficult and can distinguish participants with high and low abilities. A very popular approach used to evaluate psychometric characteristics is IRT, also known as latent traits theory or modern theory. The advantage of IRT is this theory's ability to describe the relationship between the ability, difficulty of the item, and the probability of answering correctly on a particular item (Zoghi & Valipour, 2014).

Item Characteristic Curve (ICC)

An ICC is a curve used to describe the relationship between the ability or characteristics of a test taker, the characteristics of an item, and the probability of answering correctly on the

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

item. There are item parameters used in ICC, namely, item difficulty (b), item discrimination (a), and guessing (c) (Schmidt & Embretson, 2012). The number of parameters used as fixed parameters depends on the selected model. This research uses the 2PL model, which involves two parameters, a and b.

Information Functions

The information function is intended to demonstrate the ability of an item and/or a test in providing precise information at a certain level of ability (theta). A high information value represents higher precision in providing information about test takers at a certain level of ability. IRT has information functions at the item level (item information) and the test level (test information) (Baker, 2001; Hambleton et al., 1991).

Two Parameter Logistic Model (2PL Model)

Birnbaum developed the 2PL model in 1968, where the logistics of the model were easier to work with than normal. The probability of the test taker answering correctly on an item based on the 2PL model is written on Formula (1), where $P_i(\theta)$ is the probability is the probability of test taker at a certain level of ability to answer the question correctly, θ is the ability of the test taker, b is the difficulty level of the item, a is the discrimination power of the item, and e is the constant value, 2.718.

$$P_i(\theta) = \left[\frac{1}{1+e^{-a(\theta-b)}}\right] \dots\dots\dots\dots\dots (1)$$

The use of the 2PL model in this study is based on the comprehensiveness of the 2PL model compared to the 1PL model since the 2PL model includes item discrimination parameters. Compared to the 3PL model, the 2PL model has fewer parameters. However, in the calibration process, the 2PL model is easier to achieve convergence. In the 3PL model, the difficulty of achieving convergence often occurs because the scale of the guessing parameter is different from the other two parameters.

Therefore, this study has several important points, including (1) evaluating the psychometric characteristics of the Computer-Based Academic Potential Test used for the admission selection for prospective students with disabilities during the period 2018 to 2019, (2) evaluating the characteristics of items and the amount of information based on the IRT framework, and (3) evaluating the characteristics of the test and the amount of information that the test can provide. Furthermore, important findings in this study can be useful to obtain a scientific basis in deciding whether it is necessary to reconstruct new assessments in the future, as a basis for deciding whether the test can be used to determine the score of prospective students.

RESEARCH METHOD

This research is quantitative psychometric research that aims to evaluate the psychometric characteristics of the Computer-Based Academic Potential Test. This research was conducted in four stages: test review, data collection, data analysis, and interpretation and decision making, as described in Figure 1.

The first stage was a test review. This stage involved studying the test equipment used to select prospective students with disabilities in-depth, such as examining the basis of the theory used and the construction study used, considering the number of dimensions or structural factors. It also involved the study of test quality evaluation techniques that have been done and studying the techniques of estimating the test taker's scores by considering the type of construction and psychometric quality.

https://doi.org/10.21831/pep.v25i1.38808

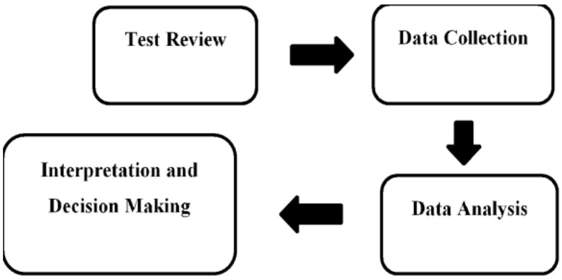


Figure 1. Four-stage Flowchart of Research in Psychometric Analysis

The second stage was data collection. It was done by collecting all the test taker's answers for the last five years (the test taker's identity was confidential). The data was only an answer response for all items in the test. The response had been coded into quantitative data, i.e., scores of 1 and 0, based on actual or false answers.

After collecting quantitative data, the next stage was analyzing the data based on the type of the previously-described construct. The analysis technique used is Item Response Theory (IRT) for the type of dichotomy response. The IRT model used is a logistic model with two parameters (2PL). The 2 PL Model is useful for estimating test items' characteristics based on item difficulty parameters (b) and item discrimination (a). The researchers also estimated the function of item information and tests to find out how informative an item and test in providing information about the test taker's ability.

The fourth stage, as the last stage, was interpretation and decision-making based on the results of data analysis, which will then lead to a decision on whether an item is feasible for university entrance selection for UB prospective students with disabilities. After the items were examined psychometrically, the interpretation was carried out at the test level. If most of the items are of good quality, then a group of such items can be concluded as a set of psychometrically qualified tests for use.

This research was conducted in *Pusat Studi Layanan Disabilitas* (PSLD) of Universitas Brawijaya, using test result data for the last two years. Samples in this study are 60 prospective students with disabilities who participated in UB entrance selection in the range of 2018 to 2019. All personal identities of the study sample were not used as research data, so that the test taker's identity was not listed in the research report.

The Computer-Based Academic Potential Test is a selection test for UB students with disabilities. The test is administered in groups and with a limited time of 90 minutes to work on the whole question. This test aims to measure academic capability consisting of three subtests: Bahasa Indonesia, English, and Mathematics. Each subtest consists of 15, ten, and five items, respectively. Each item is given a score of 1 or 0, based on a true or false answer.

FINDINGS AND DISCUSSION

A crucial IRT assumption test was conducted before analyzing the data using the item response theory (IRT) technique, known as unidimensionality. Unidimensionality testing aims to determine whether a subtest measures only one latent trait (Zanon et al., 2016).

In this study, a dimensionality test was conducted using the confirmatory factor analysis (CFA) technique. The results of the CFA can be seen in Table 1. Each subtest is modeled as unidimensional. Table 1 shows that all three subtests have a unidimensional model that fits the data. This conclusion is based on the cut-off point. The CFI fit index is said to be a reasonable fit with a minimum value of 0.90 (Wang & Wang, 2019). A value of SRMR less than 0.08 is considered a good fit (Hu & Bentler, 1999), and it is acceptable when the value is less than 0.10 (Kline, 2016). A value of RMSEA is said to be a fair fit if it falls between 0.05-0.08 and is said to be a close fit if it is less than 0.05 (Byrne, 1998).

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v25i1.38808

Table 1. Fit Indices of the Three Subtests

Subtest	Index		
	CFI	**SRMR**	**RMSEA**
Bahasa Indonesia	0.921	0.088	0.037
English	0.987	0.065	0.031
Mathematics	0.922	0.070	0.063

After the assumption test is completed, further analysis is performed using the item response theory (IRT) technique. Based on data analysis, information about item characteristics and test characteristics are obtained. Item characteristics are represented by discrimination power, item difficulty level, item characteristic curve (ICC), item information function (IIF). Test characteristics are indicated by the test characteristic curve (TCC) and test information function (TIF). Both the item characteristics and the test characteristics are very important information as the basis for deciding whether the test used for selection is a psychometrically feasible test.

Item Characteristics

The item characteristics were estimated for each sub-test, and the results of the estimation of all three sub-tests are shown in Table 2. The analysis results show that most items from all sub-tests tend to have moderate difficulty levels, and some items have difficulty levels in the range of difficult to very difficult, such as item 13 in the Bahasa Indonesia sub-test and item 5 in the Mathematics sub-test. The item discriminations fall in the range of 0.301-2.267, which is low-very high (Baker, 2001). This shows quite good results, especially since there is no negative discriminatory power.

Table 2. Item Parameter Estimation of the Three Subtests

Item		Item Parameter		Item		Item Parameter	
		α	β			α	β
Bahasa Indonesia	1	1.170	0.753	English	1	1.193	0.832
	2	0.502	3.374		2	1.349	0.689
	3	0.876	-0.095		3	0.665	0.907
	4	0.301	2.355		4	2.267	0.884
	5	0.768	0.699		5	1.080	1.179
	6	1.570	1.005		6	2.165	0.899
	7	0.489	2.561		7	0.774	-0.628
	8	1.901	1.779		8	0.634	2.067
	9	1.819	0.222		9	1.663	0.265
	10	1.053	0.636		10	0.764	0.919
	11	1.495	-0.202	Mathematics	1	0.751	1.767
	12	0.437	2.219		2	0.921	-0.085
	13	0.523	4.071		3	1.089	0.463
	14	0.783	0.791		4	0.820	1.072
	15	0.453	2.536		5	0.471	4.471

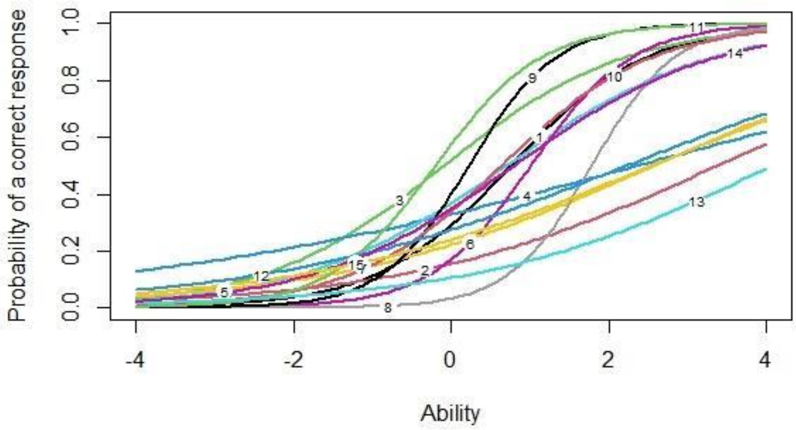
The parameters in Table 2 are more clearly illustrated in the ICC, as shown in Figure 2. The ICC illustrates how item parameters and person parameters interact with each other in a single frame. Based on the ICC on each sub-test, most items show a fairly sharp shape resembling the letter S and no ICC that has reversed direction. This is in line with Table 2, which indicates the item parameters tend to be good, and there is no negative item discrimination (Wu, 2017). However, some items look flatter, that is, item 13 in the Bahasa Indonesia sub-test. The ICC form of item 13 is flat. The item parameter (α=0.523, β=4.071) indicates that the item cannot distinguish variations in test taker's ability at a low-moderate level of ability, but rather can optimally distinguish variations in test taker's ability at a very high level of ability.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

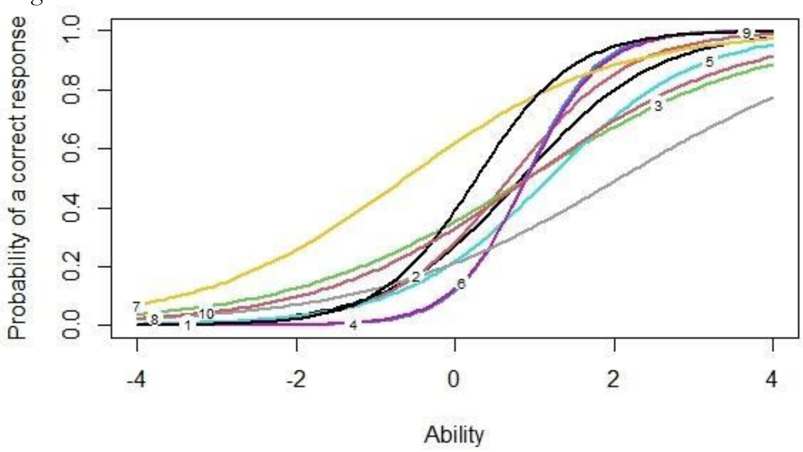
https://doi.org/10.21831/pep.v25i1.38808

After ICC is obtained, the Item information function (IIF) is obtained. Shown in Figure 3, IIF generally affirms the quality of items illustrated by ICC (Figure 2). Items that have a good ICC shape (with a sharp S shape) tend to show a high IIF curve. Items that indicate the peak of a high curve spread at a moderate-high range. Some items have a flat ICC followed by a flat IIF curve, that is, item 13 in the Bahasa Indonesia sub-test has a very flat curve because basically, the item is too difficult so almost no test taker can answer correctly. Besides, Figure 3 also shows that no item has an IIF with a peak that is at a low level of ability. This indicates that all three sub-tests did not have items that functioned optimally at the low ability level.

a. Bahasa Indonesia



b. English



c. Mathematics

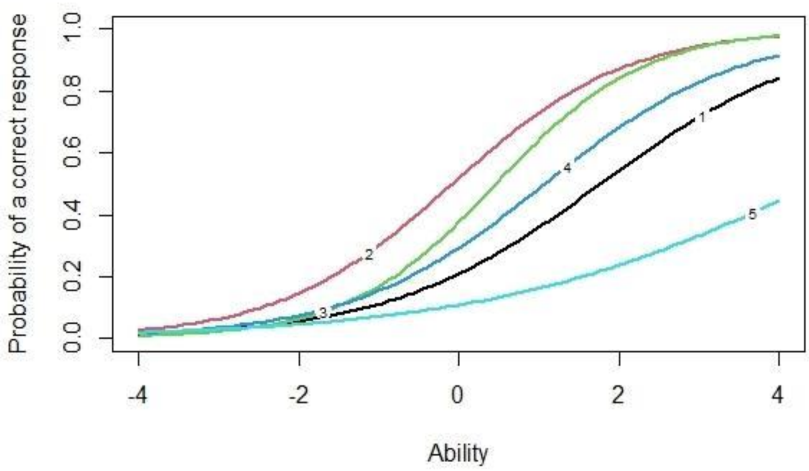


Figure 2. Item Characteristic Curve (ICC) of the Three Subtests

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v25i1.38808

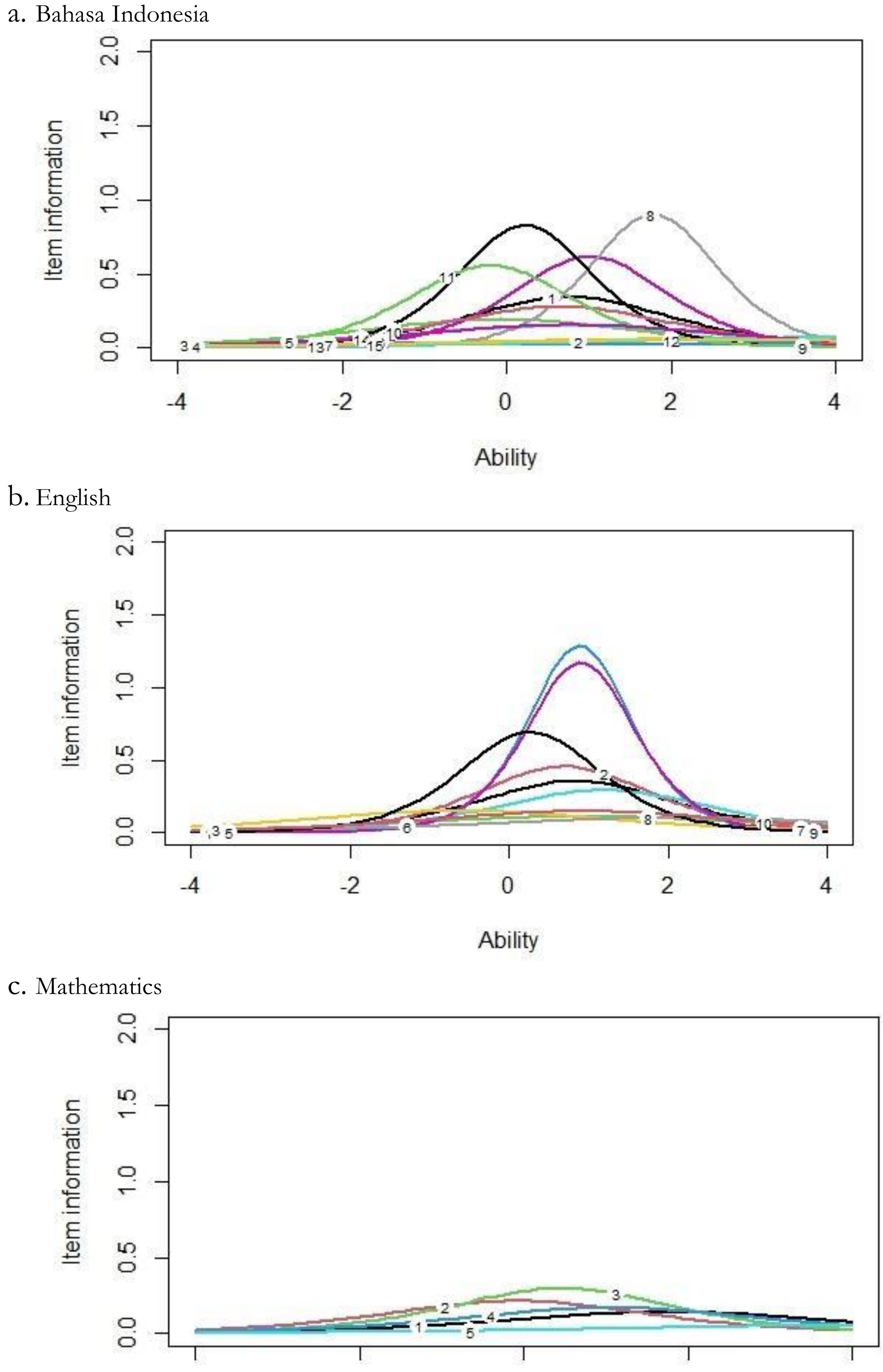


Figure 3. Item Information Function (IIF) of the Three Subtests

Test Characteristics

Beside the item's characteristics, it is also necessary to provide test characteristics to present the psychometric qualities of a set of items. Some important test characteristics are TCC and TIF. Both characteristics are obtained from the accumulation of ICCs and IIFs. TCC and TIF are obtained for each subtest (Bahasa Indonesia, English, and Mathematics) in this study. TCC for all three sub-tests is shown in Figure 4, and TIF for the three sub-tests is in Figure 5.

https://doi.org/10.21831/pep.v25i1.38808

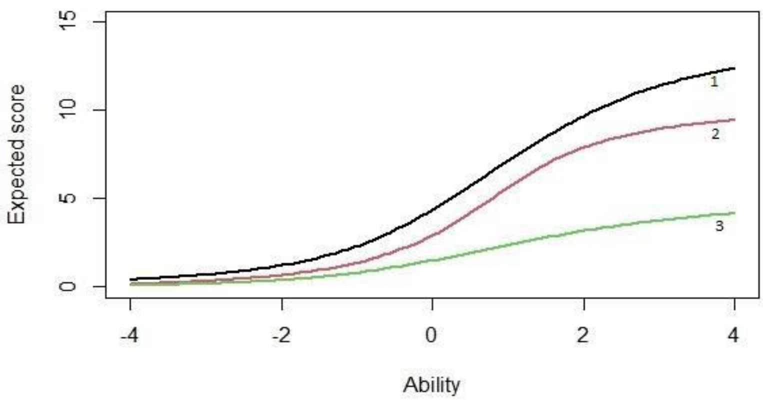


Figure 4. Test Characteristic Curve (TCC) of the Three Subtests. 1 = Bahasa Indonesia, 2 = English, 3 = Mathematics

Figure 4, where the Y-axis is the expected score/true score, and the X-axis is the ability, illustrates the relationship between true score and ability. In practical situations, TCC has an important role in transforming ability into a true score. This makes it very easy for test participants who are not familiar with scaling used in IRT (Baker & Kim, 2017). Figure 4 shows that Bahasa Indonesia is a subtest that has the best psychometric quality among the three sub-tests, while the Math sub-test tends to have a flat curve compared to the other two sub-tests.

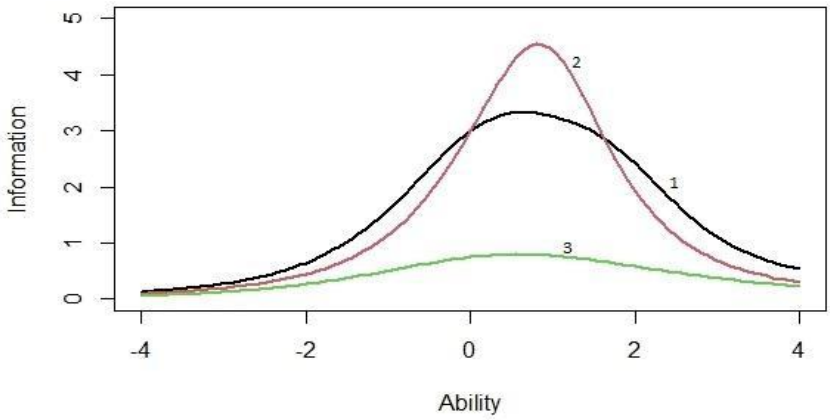


Figure 5. Test Information Function (TIF) from Three Subtests. 1 = Bahasa Indonesia, 2 = English, 3 = Mathematics

Along with IIF, TIF demonstrates the information function at the test level. Figure 5 shows all three sub-tests have the maximum information function at the moderate level, which tends to the high level of ability. This is in line with the IIF showing that the three sub-tests do not have items that can provide maximum information at low ability levels. Besides, the Mathematics subtest has the lowest maximum score (curve peak) compared to other sub-tests.

This research was done to analyze the feasibility of the Computer-Based Academic Potential Test used in admission selection for students with disabilities from a psychometric perspective, both at the item level and test level. IRT is used in this research because it has been proven that IRT has a strong performance in showing the quality of items and tests. Based on the IRT framework, item characteristics and personal characteristics are estimated separately (invariant property of IRT). Moreover, the information function at the item level (IIF) and the test level (TIF) is used to illustrate the test results' precision (An & Yung, 2014; Fan, 1998).

The results of the data analysis showed that all three sub-tests had a fairly good test quality shown from TCC and TIF. According to Baker and Kim (2017), TCC and TIF are the

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v25i1.38808

accumulation of ICC and IIF. It is said to be satisfactory when TCC exhibits a curve that is not flat or a sharp S shape. TIF is informative when it has high information, followed by a low standard error. In this study, both TCC and TIF demonstrated that the Bahasa Indonesia sub-test performed the best, followed by English and Mathematics.

Basically, TIF is intended to show how precise a test estimates ability. The more items in a test, the more informative a test is in describing the characteristics of the test taker, and it is characterized by the high maximum value of information (curve peak) (Baker & Kim, 2017). That is why in Figure 5, it appears that the Mathematics sub-test has the flattest curve compared to the other sub-tests. This is because the number of items in the Math sub-test is the least. In practical situations, TIF is generally more attractive than IIF. For example, in selecting prospective students with disabilities, ideally, the cut-off score of ability is determined in advance to determine the candidates who are accepted or who are not accepted as students. In this case, the curve peak of the TIF of the selection test should be at the cut-off score.

For more details, in each sub-test, items in the Bahasa Indonesia sub-test tend to exhibit the difficulty level in the medium to the difficult range. However, two items (2 and 13) are too difficult with β values that are extremely high (3.374 and 4.071), as shown in Table 2, so are the items in the Math sub-test, which are at moderate to difficult levels, and there is one very difficult item, that is, item 5. Unlike the others, the items in the English sub-test exhibit moderate difficulty levels and no extreme values.

The good news from the results of this study is that no ICC reverses direction (reversed S), since there are no negative discrimination parameters with the range between low to very high. Along with the item difficulty, the item discrimination of all three sub-tests also exhibited pretty good results. However, test developers need to be more aware of some items regarding the difficulty level of the item and the item discrimination. The weak discrimination power of items will affect the ICC's slope to be flatter (Zanon et al., 2016). This indicates that the item cannot detect differences among levels of ability. It could be because the item is ambiguous, the item is too difficult, or the item is too easy. When the item is too difficult, then almost all test takers will answer the test incorrectly. If the incorrect answer is coded 0, almost all item responses are 0, and the variation becomes very small. Likewise, most test-takers will respond to the item correctly on items that are too easy. If most of the test takers gave the same response, then the item will not be able to distinguish different levels of ability.

The aforementioned explanation also shows that there is a relationship between item discrimination and item difficulty. A study conducted by Sim and Rasiah (2006) found a non-linear relationship between the power of discrimination and the item's difficulty level. The curve shape depicting the relationship is curved downwards (vault), which means that the item discrimination is low when the item difficulty level is low, and increases as the item difficulty level increases, but it begins to go lower when the item difficulty is too difficult. This can be seen in this study (Table 2), which has an extremely high item difficulty value, followed by lower discrimination.

The problem to note is that there are no easy items in the three sub-tests that can function optimally at low levels of ability. Ideally, in a test, it is necessary to have an IIF peak that spreads from low to high ability levels. Thus, it is very necessary to have items that function well and are informative in providing information about the ability of test-takers at all levels of abilities. The Computer-Based Academic Potential Tests used in SMPD cannot provide reliable information for test-takers with a low level of ability, so that the test still needs to be improved, taking into account the informative items which spread evenly at all levels of ability. In short, the test needs to have items of all difficulty levels from easy to difficult.

In developing measuring instruments, things that can be an obstacle for people with disabilities need to be considered. For example, as for the blind, problems in the form of images or maps will be difficult to interpret by screen reader applications. Therefore, if the blind students get problems with many pictures, then he/she will not accurately capture the informa-

https://doi.org/10.21831/pep.v25i1.38808

tion about it. In general, people with hearing impairment/deaf are vulnerable to language comprehension deprivation, considering that deaf education in Indonesia has not sided with the natural language learning model of the deaf. This leads to a relatively slower and more profound knowledge gain, similarly for people with intellectual disabilities who can process information slowly. For both types of disabilities, special norms are needed to be developed.

Administrative challenges are another obstacle that occurs when administering a test to persons with disabilities, since people with disabilities have different characteristics and special needs. To be fair, it is necessary to note the administration of tests that accommodate people with disabilities' specific needs, for examples, they are elaborated as follows. (1) For the blind test takers, this computer-based test should be readable by the application layer reader precisely following the writing pronunciation. In some cases, screen readers often do not have an accurate ability to read Indonesian text. It is better to be concerned about the time provided for the test takers with disabilities, mostly blind test takers. (2) For the physically impaired, using the computer used for the test's work should not prevent them from working using their limbs, especially for students who have difficulty moving their hands and upper limbs. There need to be special adjustments for people with disabilities who have a less adaptive physical posture with the size of most computer and keyboard products (e.g., dwarfs). (3) For the deaf, there is still a need for a Sign Language Interpreter to explain the procedure of conducting the test at once to facilitate if the test taker has difficulty doing the test.

These specific needs also implement tests that should be done in different rooms for people with disabilities. For example, deaf participants cannot be placed in a room with mental disabilities such as autism or ADHD, because the deaf will do a lot of movement as a form of communication for those who will become a significant distractor for autism or ADHD.

In the context of this study, the limited number of participants in each testing period was also an obstacle in the development of measuring instruments, given that large and representative samples are indispensable in the analysis of psychometric characteristics. Therefore, it is necessary to do continuous data banks so that later psychometric analysis for accommodating measuring instruments for people with disabilities can be better. Furthermore, the items in the subtests have been administered only to the prospective students with disabilities, so there is no evidence that the quality of these items is different between the two populations, which are prospective students with disabilities and prospective students without disabilities. In the future, it would be more interesting if the subtest were administered to the two populations so that the item characteristics of the two populations can be compared.

CONCLUSION

The study found that computer-based tests used in SMPD UB have three sub-tests containing items with satisfying performance. However, some items are still too difficult for the test takers. Besides, the three sub-tests also do not have easy items, so it is very difficult to get information about a test taker with low ability levels. In general, the test has good items with moderate to difficult difficulty levels. It is very effective for measuring the ability of test-takers ranging from moderate to high level.

REFERENCES

An, X., & Yung, Y.-F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS364*, 1–14. https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

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

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

https://doi.org/10.21831/pep.v25i1.38808

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using* R. Springer.

Byrne, B. M. (1998). *Structural equation modeling with Lisrel, Prelis, and Simplis*. Psychology Press. https://doi.org/10.4324/9780203774762

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (Volume 2)*. SAGE Publications.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.

Palombi, B. J. (2000). Recruitment and admission of students with disabilities. *New Directions for Student Services*, *2000*(91), 31–39. https://doi.org/10.1002/ss.9103

Pratiwi, A., Lintangsari, A. P., Rizky, U. F., & Rahajeng, U. W. (2018). *Disabilitas dan pendidikan inklusif di perguruan tinggi*. Universitas Brawijaya Press.

Rukmantara, A., & Lesmana, B. (2018). Inclusive education and SDGs: Snapshots from the field. In M. Anwar (Ed.), *International Conference on Sustainability Development Goals for Disabilities (ICSDGD)* (pp. 1–17). Asosiasi Profesi Pendidikan Khusus Indonesia (APPKhI). http://appkhi.or.id/Proceedings ICSDGD.pdf

Schmidt, K. M., & Embretson, S. E. (2012). Item response theory and measuring abilities. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology - Volume 2: Research methods in psychology* (2nd ed., pp. 451–473). John Wiley & Sons.

Sim, S. M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals of the Academy of Medicine, Singapore*, *35*(2), 67–71.

Wang, J., & Wang, X. (2019). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.

Wolanin, T. R., & Steele, P. (2004). *Higher education opportunities for students with disabilities: A primer for policymakers*. The Institute for Higher Education Policy.

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, *59*(4), 453–470. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/04_Wu.pdf

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 18. https://doi.org/10.1186/s41155-016-0040-x

Zoghi, M., & Valipour, V. (2014). A comparative study of Classical Test Theory and Item Response Theory in estimating test item parameters in a linguistics test. *Indian Journal of Fundamental and Applied Life Sciences*, *4*(s4), 424–435. https://www.cibtech.org/sp.ed/jls/2014/04/JLS-051-S4-052-VALIPOUR-COMPARATIVE.pdf

ISSN (print) 2685-7111 | ISSN (online) 2338-6061