The Characteristics of *Digital Assessment Bloom* for Indonesian Junior High School

Yuriska Melia Sari*, Badrun Kartowagiran
Educational Research and Evaluation Department, Graduate School of Yogyakarta Stated University, Yogyakarta, Indonesia
yurizka.melia2016@student.uny.ac.id

Heri Retnawati
Mathematics Education Department, Yogyakarta Stated University, Yogyakarta, Indonesia

Shofan Fiangga
Mathematics Education Department, Universitas Negeri Surabaya, Surabaya, Indonesia

***Abstract*—In this recent years, the integration of technology in learning and assessment has become a major need in education. In Indonesia, CBT has been used as a computer-based national examination system (UNBK), a computerized exam as a media to display the problem and process of answering it. In addition, Digital Assessment Bloom (DAB) research is a form of tool development a computer-based assessment for the ninth grade of Junior High School based on Bloom's revised taxonomy. This study aims at: (1) understanding the characteristics of Digital Assessment Bloom test items using the Item Response Theory and (2) determining the student's ability on the trial exam instruments. This study used quantitative approach assisted with SPSS and BILOGMG 3.0 software. The developed DAB has been piloted on a limited scale in two junior high schools selected by random sampling technique. Results showed that DAB instrument fulfilled the assumptions on item response theory and in accordance with the use of 3-Parameter Logistic models. Furthermore, the results indicated that there was an item classification of 16 good items stored in item bank for the sake of better use of the assessment.**

Keywords—Assessment; Digital assessment bloom; Item response theory.

I. INTRODUCTION

The fact that the use of technology in the classroom practice has become more salient and resistible. The integration of technology in educational assessment answers effectively the recent challenge in the education field by developing and carrying out most of evaluation methods. In fact, the computer-based assessment performs better than the use of conventional method in a way of recording both of students' knowledge and cognitive abilities [1]. One of implementation form of technology in assessment is CBT (Computer-Based Testing), of which an innovative approach in educational evaluation might be conducted in faster and easier ways [2].

The recent study concerning evaluation field have been focusing on the development of the use of computer-based evaluation with various approaches and techniques based on their purpose on evaluation and testing practices. In Indonesia, CBT has been used as the evaluation system in high school national final examination, or familiarly known as UNBK (Ujian Nasional Berbasis Komputer). Since 2016, the examination has implemented computer as a medium to present test items and answering process [3]. Another form of CBT is the Digital Assessment Bloom in which the revised Bloom's taxonomy has been considered as the main framework development [4].

Innovation in evaluation and testing practices is important. However, those innovations should use standardized principle in analysing measurement data as the main reference in a way that the analysis would involve systematic procedure before a valid conclusion can be made from the analysis [5]. A high quality instrument is not only analysed theoretically (content, construct, and language) but also empirically [6]. In fact, there are two approaches in analysing the data resulted from measurement test, which are Classical Test Theory (CTT) and Item Response Theory (IRT) [7]. The CTT might result several weaknesses such as; the item difficulty level and the discriminant potency rely heavily on the used sample, the test result cannot be generalized outside the context used, and the reliability concept that remains unclear. Nevertheless, those weaknesses could be overcome by using the concept of IRT developed by [8]. They stated that IRT is basically developed using two postulates which are traits (person's ability dimension) and the relationship between subject in a certain item and the latent ability tool in a form of item characteristic curve. The traits are used for predicting subject's ability in a tested item.

Several assumptions applied in IRT cover unidimensional, local independent, and item characteristics function that reveal the actual relationship between unobserved variables (e.g. ability) with observed variable (e.g. responses) [8]. The unidimensional assumption stated that the developed items only measure one factor of the subject's ability. Unidimensional assumption can be fulfilled by calculating the ration between the first and the second chi-square whereas the higher ration conclude that the model is unidimensional [9]. As the second assumption, local independent suggests that the examinee responds toward a certain item that does not correlate with other tested item. The last assumption claims that the parametric invariance from which the item characteristics are not dependent with the examinees' ability, parametric distribution, and the examinee characteristic is not dependent with the test item's characteristic [6].

The assumption used in IRT results three logistic models in describing the item characteristic function using particular parameter. One-parameter logistic model, which is also called Rasch' Model, describes the model of item characteristic function with parameter of difficulty or we can symbolize with b [8]. Meanwhile, the two-parameter logistic model uses parameter of difficulty level (b) and discriminant potency (a). There is an additional parameter named pseudo-guessing (c) as the three-parameter logistic model besides parameter of difficulty level (b) and discriminant potency (a).

In this study, a vast scale measurement result would be analysed. The developed Digital Assessment Bloom is used for eliciting the mathematics problems item test characteristic using IRT. Beside, this study also verify the assumptions in IRT which are unidimensional, local independent, and parametric invariance on Digital Assessment Bloom. Furthermore, a suitable logistic model is verified and good test items are analysed.

II. METHODS

This descriptive-evaluative research used quantitative approach to analyse the ninth grade of junior high school students in Surabaya who had already participated in the Digital Assessment Bloom test. This study describes IRT's assumptions namely unidimensional, local independence, and parameter invariance. Besides, this study also examined suitable logistic model for determining difficulty level, discriminate power, and pseudo-guessing. In addition, this study also revealed and elaborated good and bad items in Digital Assessment Bloom test.

This study was carried out in five public and private junior high schools in Surabaya with 947 participants. The participants were enrolled in the ninth grade for 2015/2016 academic year who further faced Computer based National Examination (UNBK) in 2016. Data collection in this study used documentation technique by collecting participant responses from the developed Digital Assessment Bloom. SPSS and Bilog-MG 3.0 were used to analyse both the IRT's assumptions, suitable logistic models for Digital Assessment Bloom, and suitable item parameters based on the logistic model. The parameter of difficulty level (b) could be seen from Threshold column. The discriminant item power (a) could be seen from Slope column. While pseudo-guessing (c) could be revealed from Asymptote column in Bilog-MG analysis results. Moreover, in determining whether the test item was good or not, Table 1 helps show the criteria used [10].

TABLE I. GOOD TEST ITEM CRITERIA IN IRT

Parameters	Value	Note
Discriminate power (*Slope*)	0.4 - 2	Good
Difficulty level (*Threshold*)	-2 - 2	Good
Pseudo-Guessing (*Asymptote*)	0 - 1/k (k = the number of alternatif answers)	Good
Model matching-test	>0.05	Fit Model

There were several techniques in examining IRT's assumptions and model matching test using either SPSS or BILOGMG 3.0. In examining IRT's assumptions, the unidimensional assumption could be tested using three different techniques covering Eigen-value analysis from inter-items matrices, Stout test on unidimensional assumption, and residual-based index resulted from unidimensional solution (De Mars, 2010). Whereas, in this study, eigenvalue analysis was used to examine unidimensional assumption using SPSS. Moreover, local independency could be fulfilled by proofing unidimensional assumption [11]. There was a developed procedure to examine local independency of two items by using chi-square [8]. However, Hambleton's method would be inefficient if implemented in an instrument that had a number of test items. Furthermore, parametric invariance assumption could be seen on the group of different test items, by which gender difference was used in this study.

Model matching test was conducted to determine whether the resulted IRT analysis was already suitable with either one of the following logistic parametric models namely 1-PL, 2-PL or 3-PL. This test was carried out to determine feasibility of the test item used. There were two ways used to examine matching model based on the analysis of students' responses. They were statistics methods and Item Characteristic Curve method from which, in this study, statistics method was used. In details, this statistics method examined significant probability value from BILOGMG 3.0. If the value of sig $< \alpha$, then item did not match with the model and a calculation on what extend the model did not match with the test items. During matching test, the model with the least incompatibility would be used for the next analysis.

III. RESULTS AND DISCUSSION

In this study, Digital Assessment Bloom that had been implemented in wide scale was analysed for the IRT's assumptions, which are unidimensional, local independency, and parametric invariance. Furthermore, this paper discusses the suitable model in Digital Assessment Bloom instrument either 1-PL, 2-PL or 3-PL. From the analysis results, the characteristics of good test item would be elicited.

A. *IRT's assumptions test*

[8] suggested that the IRT was based on three assumptions namely unidimensional, local independency, and parametric invariance. Unidimensional was described as each test item measuring one certain ability of which eigenvalues of inter-items variance-covariance matrices used [6]. This test was carried out using SPSS to cope with the preliminary analysis. Sample's sufficiency analysis was shown by Chi-Square value on Bartlett test of 3807.538 with the degree of freedom 276 and

p-value less than 0.01. This result implied that the size of participants in this study was already enough to apply unidimensional assumption. Based on factor analysis using SPSS, a scree plot was obtained (see Figure 1). The diagram showed that the Eigenvalue started to decline horizontally from the third value. This suggested that there was only one dominant factor namely the first Eigenvalue in the developed Digital Assessment Bloom. The second Eigenvalue had significant effect on variance component that could be explained. In summary, this instrument, at least, measured two factors on which the first factor was the dominant one.

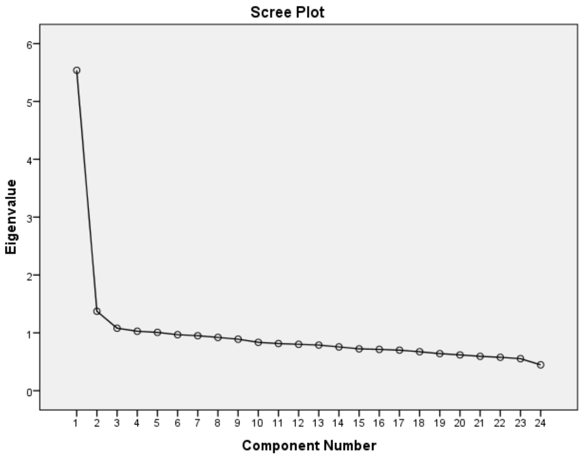

Figure 1: Scree Plot from Factor Analysis Result

The second assumption concerning local independency was revealed using unidimensional characteristic of the participant's responses on Digital Assessment Bloom instrument. Based on unidimensional test, this instrument was able to measure one dimension in mathematics ability. The unidimensional analysis showed that there was no correlation between participants' responses on an item toward different item [12, 13]. This suggested that local independence assumption had been fulfilled.

The third assumption was parametric invariance on which the test item characteristic was not influenced by the distribution of the participants' ability and the characteristic parameter of the participant was independent toward the test item characteristics [6]. Item parametric estimation on different group of participants was used to prove parametric invariance assumption. In this study, an analysis on discriminate power (a), difficulty level (b) and pseudo-guessing (c), as parametric invariance assumptions, on different gender group of participants was described in scatter plot (see Figure 2).

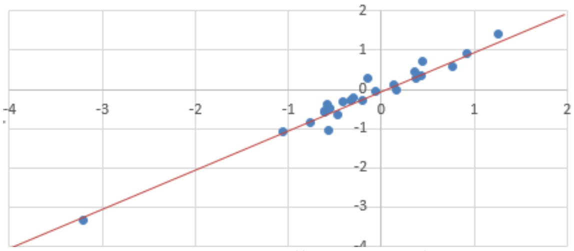

Figure 2: Parametric Invariance on Difficulty Level from Male and Female Group

Based on the result of parametric invariance test on male and female groups shown in Figure 2, there was a significant correlation proven from the scatter plot by which the scatter was approaching a line through initial point and slope 1. Therefore, it suggested that the developed instrument had already satisfied the assumption in which the parameters on different gender group was invariance [14]. In short, three assumption analyses had shown that the developed Digital Assessment Bloom tested in wide scale had fulfilled unidimensional, local independency, and parametric invariance as IRT's assumption.

B. *Matching test on logistic parametric model*

[6] stated that comparing chi square value from three developed models could be used for logistic model testing. The test items in the instrument were compatible with the model of the calculated value of chi-square, of which it showed less than the chi-square value of the table. Moreover, the compatibility model could also be seen from significant value (sig<α) in which the model showed incompatible with the developed model. Based on the result of matching test on Digital Assessment Bloom items regarding IRT Model 1, 2, and 3 Logistic, there were 10 out of 24 items compatible with 1-PL model; 15 out of 24 items compatible with 2-PL model; and 20 out of 24 items compatible with 3-PL model. Hence, the most compatible model for the Digital Assessment Bloom (DAB) was Logistic Parametric 3 (PL-3). This suggested that the model used in this IRT study was 3-PL. These findings seemed to be consistent with previous researchers [15,16]. [15] indicated that the three-parameter logistic model showed higher significant relationship to CTT in term of item discrimination and item difficulty, whereas, the two-parameter logistic model revealed lower significant relationship to CTT in term of item discrimination. Futhermore, [16] stated that the three-parameter logistic model was the most comparable induced with IRT.

C. *Good and non-good items classification*

After matching test conducted, 3-PL model was used to estimate items and ability. The parameter analysed using Bilog MG 3.0 in this study was discriminate power (a), difficulty level (b) and pseudo-guessing (c). Table 2 shows the result of the analysis.

TABLE II. CHARACTERISTICS OF DAB INSTRUMENT BASED ON IRT PL-3 MODEL

Item	Subject	a	b	C	Note
1	Similarity (C2)	1.436	-1.087	0.190	GOOD
2	Similarity (C5)	1.606	-0.522	0.343	NOT GOOD (c > 0.25)
3	Similarity (C1)	1.436	-0.399	0.135	GOOD
4	Similarity (C5)	1.582	-0.209	0.376	NOT GOOD (c > 0.25)
5	Similarity on Triangle (C4)	0.566	-0.664	0.195	GOOD
6	Congruence on Triangle (C2)	1.946	0.934	0.140	GOOD
7	Similarity and Ratio (C4)	2.822	0.071	0.124	NOT GOOD (c > 0.25)

Item	Subject	a	b	C	Note
8	Congruence on Triangle (C3)	0.874	-0.582	0.294	NOT GOOD (c > 0.25)
9	3 Dimensional Curved Figure (C5)	1.200	-0.508	0.238	GOOD
10	Cylinder (C1)	1.513	0.374	0.097	GOOD
11	Lateral and Based Face Identification on Cone (C6)	0.332	-0.950	0.25	GOOD
12	Identification on Cylinder and Cone volume relationship (C6)	1.745	0.370	0.381	NOT GOOD (c > 0.25)
13	Cylinder in Problem Solving (C3)	1.494	-0.190	0.242	GOOD
14	Cone Area Surface (C4)	1.309	-0.811	0.25	GOOD
15	Single Data Frequency (C1)	1.051	0.103	0.227	GOOD
16	Single Data Mean (C2)	1.736	0.677	0.147	GOOD
17	Mean Concept (C3)	1.314	0.601	0.208	GOOD
18	Setting New Rule from the Given Mean (C6)	1.506	-0.355	0.391	NOT GOOD (c > 0.25)
19	Central Measurement (C5)	1.190	0.018	0.351	NOT GOOD (c > 0.25)
20	Mean Data (C2)	0.924	0.116	0.243	GOOD
21	Median Data (C3)	1.198	0.405	0.208	GOOD
22	Central Measurement (C5)	0.992	0.025	0.438	NOT GOOD (c > 0.25)
23	Bar Diagram (C4)	0.518	1.926	0.25	GOOD
24	Line Diagram (C1)	1.471	1.323	0.25	GOOD

In accordance with Table III and PL-3 Model regarding parameter analysis, it can be concluded that there were 8 not-good test items and 16 good test items. Then, the not-good items were replaced while the good items were saved in DAB test item bank to be used later for evaluation.

IV. Conclusion

The IRT's Assumptions have fulfilled the DAB instrument. The unidimensional assumption suggests that there is one dominant factor which is mathematics ability. The local independence assumption is generated from the unidimensional case. While, the parameter invariance assumptions are concluded from the analysis result of participant responses on different gender group.

Based on the matching test using statistical analysis to determine the compatible model for the DAB instrument, 3-PL model is the most suitable model with discriminate power parameter (a) 0.322 to 2.822; parameter of difficulty level (b) -1.087 s/d 1.926; and pseudo-guessing parameter (c) 0.097 s/d 0.438. Further analysis shows that there are 8 not-good test items and 16 good test items.

The weakness of this study is that there is no analysis on participants' ability estimation resulted from the DAB instrument. Therefore, for further study, a more comprehensive study should be administered by elaborating participant ability estimation using either Bayesian method or maximum likelihood method. In addition, implementing DAB in different region might foster the study impact in a vast way.

Acknowledgments

Authors wish to acknowledge assistance or encouragement from colleagues. They also thank for special works by technical staffs or financial supports from organizations.

References

[1] M. Jamil, Tariq, and Shami, "Computer-Based vs Paper-Based Examinations: Perceptions Of University Teachers", TOJET: The Turkish Online Journal of Educational Technology, vol. 11(4), pp. 10-19, 2012.

[2] M. Thurlow, S. S. Lazarus, D. Albus, and J. Hodgson, Computer-based testing: Practices and considerations (Synthesis Report 78). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2010.

[3] BSNP, Prosedur Operasional Standar Penyelenggaraan Ujian Nasional Tahun Pelajaran 2015/2016. Jakarta: Kemendikbud, 2015.

[4] Y. M. Sari, "The Development of Digital Assessment Bloom As Assessment Tools In Junior High School", Proceeding of International Conference on Research, Implementation and Education of Mathematics and Sciences 2014, 2014.

[5] M. Elvira, and S. Hadi, "Karakteristik butir soal ujian semester dan kemampuan siswa SMA di kabupaten Muaro Jambi" Jurnal Evaluasi Pendidikan, vol. 4 (1), pp. 58-68, 2016.

[6] H. Retnawati, Validitas, Reliabilitas & Karakteristik Butir (Panduan untuk Peneliti, Mahasiswa, dan Psikometrian). Yogyakarta: Parama Publishing, 2016.

[7] D. Mardapi, Pengukuran, Penilaian, dan Evaluasi Pendidikan. Yogyakarta: Nuha Medika, 2012.

[8] R.K. Hambleton, H. Swaminathan, and H.J. Rogers, Fundamentals of Item Response Theory. London: Sage Publication, 1991.

[9] J. Hattie, "Methodology Review: Assessing Unidimensionality of Test and Items", Applied Psychological Measurement. Vol. 9, pp. 139-164, 1985.

[10] N. Huda, and D. Mardapi, "Komparasi model penskoran berdasarkan teori respons butir pada soal ujian nasional mata pelajaran matematika", Jurnal Evaluasi Pendidikan, vol. 3(1), pp. 56-66, 2015.

[11] C.E. De Mars, IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood: SSi, 2010.

[12] C. E. DeMars, and J. N. Jacovidis, "Multilevel IRT: When is local independence violated? Electronic board", presented at the annual meeting of the National Council on Measurement and Education, Washington, DC, 2016.

[13] K. Hidayati, and H. Retnawati, "Pendeteksian Keberfungsian Butir Diferensial (Differential Item Functioning, DIF) Menggunakan Indeks Perbedaan Probabilitas pada Data Politomus dengan Model Generalized Partial Credit Model (GPCM)", Matematika dan Pendidikan Karakter dalam Pembelajaran, pp. 978–979, 2011.

[14] E. S. Kim, and M. Yoon, "Structural Equation Modeling : A Testing Measurement Invariance : A Comparison of Multiple-Group Categorical CFA and IRT Testing Measurement Invariance : A Comparison of Multiple-Group Categorical CFA and IRT, pp. 37–41, 2011.

[15] N. Abedalaziz, and C. H. Leng, “The Relationship between CTT and IRT Approaches in Analyzing Item Characteristics”, The Malaysian Online Journal of Educational Science, vol. 1(1), pp. 64–70, 2010.

[16] C. Nukhet, “A Study of Raven Standard Progressive Matrices Test’s Item Measures Under Classic and Item Response Models: An Empirical Comparison. Ankara University”, Journal of Faculty of Educational Science, vol. 35(1-2), pp. 71-79, 2002.