ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research, volume 100

5th South East Asia Development Research (SEA-DR) International Conference

Characteristics of School Examination Test of Biology Subject

Dyah Febria Wardhani
SMPN 1 Paramasan
Banjar, Indonesia
dyah2302@gmail.com

Suratno, Aminuddin P. Putra
Universitas Lambung Mangkurat
Banjarmasin, Indonesia.

***Abstract*— The results of the analysis of a test is indispensable for making the policies that should be done by a teacher in improving the quality of teaching and learning. This study aimed to describe characteristics of the school examination for Biology subject in Banjar Regency in the year of 2014/2015. The quantitative analysis was based on modern test theory using software version MICROCAT Big steps 2.30 on the response of the students' answers to the tests made in the Department of Education and Ministry of Religious Affairs. The results showed that the tests made in the Ministry of Religion have suitability items with the model of Rasch and better quality than the tests made in the Department of Education. Both tests had a level of difficulty that was good, discriminating power which was unacceptable, and the distribution of response answers that was not good. Statistics showed that the reliability of items was preferential while the reliability of persons was weak. Moreover, it was still found the students who responded inconsistently.**

Keywords—Biology, Characteristic, Examination

I. INTRODUCTION

Curriculum, learning process, and assessment are very important three-dimensions in education and interrelated between each other [1]. Assessment can be used as a measuring tool for teachers and students in relation to the analysis of the success rate of the learning process [2]. The analysis is indispensable for the creation of policies that should be done by a teacher to improve the quality of teaching and learning process [3].

Data of assessment utilization results are very useful, not only for teachers but also for students, principals and education supervisors in order to improve the development of education in schools both in terms of planning, executing or implementing, assessment, monitoring, or determining the outcomes of education [4].

Learning achievement test is a test tool in education that at least provides four functions, namely to evaluate students, motivate and help students, provide feedback for teachers, and develop instruction [5]. The test which is not yet standardized will provide information about the students' ability that is bias and inaccurate so that the data obtained are still doubtful [6].

Multiple choice question is a kind of objective test that is most widely used by teachers [7]. This test consists of a question that refers to the matter subject (stem) and a set of two or more options that consist of possible answers from the statement. The best answer is an answer key (key) and the other is called by distractors [8]. This matter is very versatile type for use because it can measure a wide range of learning outcomes ranging from simple to complex, can be adapted to most types of subjects, is very easy to apply, and the majority of standard tests use this form [9].

Levels of cognitive complexity and depth of knowledge become important aspects in learning outcomes assessment standard because they are used at the level of matter in the blueprint for a summative test [10]. One of international studies regarding students' cognitive ability is TIMMS (Trends in Mathematics and Science Study) which states that the average percentage of correct answers in a question of understanding is always higher than the percentage of correct answers on the question of the application and reasoning [11]. This is not in accordance with science learning process that has characteristic of scientific process or scholarly work based on the ability to think and problem resolution [12].

Broadly speaking, the analysis of items is divided into two, namely the approach of classical test theory and item response theory. In classical test theory, level of difficulty and discriminating power largely determines the quality of items. However, the characteristics of the item which is resulted by classical test theory changes depending on the group of examinees [13] [14]. The analysis model of achievement test that is based on item response theory is more credible [15].

Item response theory is a theory of general statistics regarding the exam and test characteristics and how those characteristics attributed to the ability as measured by the questions in the test. The response of students to the item can be measured dichotomous and politomus [16]. If the data are scored in dichotomous, the use of Rasch models for tests analysis is very appropriate [17]. Rasch model application in the development of the test can be an excellent tool in evaluating performance because it can make a selection matter so well that the results are good and valid [18]. In addition, this model will produce an independent measurement [19].

Rasch models show the degree of difficulty on item response theory in an objective assessment. This model is able to prove the validity and can be used to replace the factor

analysis on classical test theory and test for undimensional on the written test [20]. Analysis of this model can be done with the help of a computer program, one of which is a program Bigsteps and Winsteps [21][18][22]. Mathematically, the characteristic of function item in Rasch models expressed by Hambleton, Swaminathan, Rogers (1991), Fahmi (2011), and Ratnaningsih & Isfarudi (2013) in the following equation[16][23][24]:

$$Pi\ (\theta) = \text{with } i = 1, 2, ..., n \qquad (1)$$

Information:
Pi (θ): Opportunities enabled participants answered correctly θ can answer correctly the first item.
θ: the level of ability of the test taker
bi: the difficulty level butirke-i
e: natural numbers whose value approaches 2,718
n: number of items in the test
D: constant-value of 1.7 as a standard deviation of distribution logistics

Opportunities for someone to answer an item correctly is a function of the ability of the participants and the level of difficulty of the grain. The value of b indicates a point on a scale of standard capabilities (0.1) where it is likely to give the correct response was 0.50 (meaning 50% chance to be able to answer the item correctly) [25]. Level of difficulty of items moves from - ∞ to + ∞. Nevertheless significant value typically moves on a scale of -3 to +3 in logit units (log odds units). Otherwise good grain is grain that has a difficulty level ranges from -2 up to +2 [16].

II. METHODS

This research is non experimental quantitative research using descriptive methods implemented in Banjar district. The data source is the answer sheets of students on tests made in the Department of Education and Ministry of Religious Affairs of Banjar district. The population in this study were 557 students who gave the response on the school exam tests student on science subjects (Biology material) at the school year 2014/2015. The population consisted of 381 students response from junior high school and 176 students response of MTs. The students' answers were analyzed quantitatively using MICROCAT program Big steps 2:30 version.

The primary variable of this study was characteristic of Biology items used on school exams in Banjar district. Sub research variables are wright (map of persons and items), item (item fit, item measure, Ptbis, item option frequencies, quality), abilities' students (person fit, person measure), and a summary of the analysis.

III. RESULT AND DISCUSSION

A. Map of Wright (map of persons and items)

Based on the analysis of Big steps program, wright map obtained from tests made in the Department of Education and Ministry of Religious Affairs of Banjar regency is shown in Fig. 1.

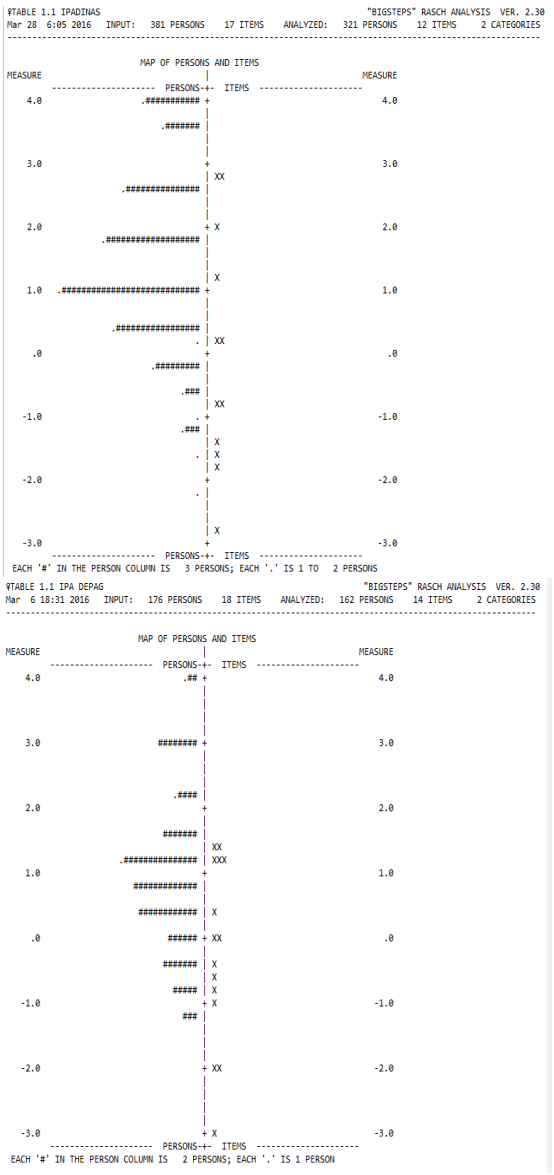

Fig. 1. Map of *Wright* Test made in Department of Education and Ministry of Religious Affairs in Banjar district

Fig. 1 shows a blank intervals (> 3 logit and> 2 logit) on the map of test items made in Department of Education and Ministry of Religious Affairs. In addition, there were the vacancy of person maps in the interval (-3 to -2) on tests made in the Ministry of Religion. Accordance with the opinion [32], it indicates that the tests made in the Department of Education and Ministry of Religious Affairs of Banjar district does not function optimally if it is imposed on students who have a high level of ability and tests made in the Ministry of Religion with a low difficulty level works less optimal. Therefore, tests are made by the Department of Education and Ministry of Religious Affairs of Banjar district cannot measure all students' abilities. Follow-up to do is that test developers need to add some points on the test that have a high difficulty level in order to fill the void interval and eliminate items that are less functional.

B. Items

Characteristics of test items made in Department of Education and Ministry of Religious Affairs of Banjar district can be seen in Table 1.

TABLE I. CHARACTERISTICS OF ITEMS

The Test Maker		D	K
Item fit	Amount	12	15
	Percentage	70.59	83.33
Item measure	High	2	-
	Moderate	9	13
	Low	1	2
Item Difficulty		-2.79-(2.88)	-5.41-(1.39)
Indeks of Discrimination		0.06-0.26	0.09-0.31
Distribution of the option	Good	4	8
	Not Good	8	7
Quality	Good	9	13
	Enough	3	2
	Not Good	5	3

[a.] *Source: Data Processing Result

Information:
D: Education Office of Banjar District
K: Ministry of Religious Affairs of Banjar District

Based on the results, the test items made in Education Office were not good while the test items made in Ministry of Religious Affairs were good, but equally not at 100% so that the two tests still have items that do not fit [26]. Based on the opinion [19], the items that do not fit show misconceptions about the items. Items that fit with the model mean the items function consistently with what is expected by the model [27]. Further research is needed on the causes of problems that do not fit according to the rules-writing and make repairsment before re-testing.

Based on the results, level of difficulty in the two tests was good even though they still found the questions that were very difficult and very easy [26]. The follow-ups to do by test developers are as follows: First, the items included in the medium difficulty level is included in the question bank and can be removed again in a test of learning outcomes in the future. Secondly, items which have a high degree of difficulty can be discarded, reexamined, and any time can be used in the tests, which were very tight as a selection test. Third, items with low difficulty level also have three possibilities, namely being discarded, reexamined, or can be used in the tests that are loose in the sense that most of the testee will pass in these tests[28].

The discriminating power of both tests had low values. Although [26] states that the value of the discriminating power is good [26] as it does not have a negative value, but according to [29], the discriminating power of the two tests was unacceptable because it had a value ≤0,40[29]. Follow-up to do is revise the item, put it forward again in the upcoming achievement test, and then analyze it again whether the discriminating power is increasing or not. The other follow-up is to dispose of the item.

In accordance with the results of [26], the distribution of answers from the both tests was not good. Follow up on the results of the analysis is that selection of answers that have been able to function properly can be used again on the upcoming tests, but those not functioning properly can be revised or replaced with other options [26]. This is accordance with what expressed by [26], that the questions whose the response distribution is unfavorable needs to be revised [29].

The percentage of question which is not good on the test made in Education Office and Ministry of Religious Affairs of Banjar district sequentially is 29.41% and 20%. The classification [30] and both the quality of the test were good [31].

C. Test Participant

Characteristics of testee on the exam 2014/2015 school year are presented in Table 2.

TABLE II. CHARACTERISTICS OF TESTEE

The Test Maker		D	K
Person fit	Amount	355	155
	Percentage	93.16	88.07
Person measure	High	89	30
	Moderate	200	107
	Low	92	30
Student Level		-2.24-(4.45)	-1.15-(3.78)

[b.] Source: Data Processing Result

Percentage of less than 100% indicates that there is still a testee who was inconsistent in answering the questions. Rasch modeling can detect their response patterns that do not fit, the mismatches answers given based on the ability compared to the ideal model [19]. The information of response pattern that does not fit can be known further by looking at schallogram. Through the matrix Guttman, the direct cause of the pattern of response is not appropriate. Further research is needed on the causes of students who have no appropriate response and the follow up is necessary on those students according to the results of causes analysis.

Testee capability is divided into low, medium, high, according to classification of [19]. Students who have the medium ability was more than those with the low and high ability. The average ability of the test taker in analysis results is different from the results of the estimation. This is because the results of the analysis only calculate a score which is not extreme, both minimum and maximum, whereas the data contained some test takers who have the maximum score.

D. Statistics Summary

Statistical summary of items and the test taker is presented in Table 3.

TABLE III. STATISTICS SUMMARY

The Test Maker		D	K
Item	Measure	1.24	0.86
	SEM	0.84	0.69
	INFIT MNSQ	1.00	1.00
	OUTFIT MNSQ	0.87	0.99
	Reliability	0.47	0.57
Testee	Measure	0.00	0.00
	SEM	0.18	0.20
	INFIT MNSQ	1.00	1.00
	OUTFIT MNSQ	0.87	0.99
	Reliability	0.99	0.97

[c.] Source: Data Processing Result

Statistics summary shows that the average ability of the test taker is higher than the level of difficulty of the questions. Reliability learners are weak while reliability is a special item. Moreover, it has good conditions for measurement [19],[32]. Thus, the test made in the Department of Education and Ministry of Religious Affairs of Banjar Regency has the testees' ability that is higher than the level of difficulty of the test were tested. Testees' answer consistency is not good although it has good conditions for measurement.

IV. CONCLUSION

Characteristics of the test at the school exam for Biology subject at Banjar district in 2014/2015 have match items with the Rasch model, have good difficulty levels, unacceptable discrminating power, and answer distribution which was not good. The quality test was pretty good, special test reliability was found, and reliability of the test participants was weak. Moreover, there were still students who responded inconsistently.

REFERENCES

[1] S. Surapranata, (2004). Panduan Penilaian Tes Tertulis: Implementasi Kurikulum 2004, Bandung: PT. Remaja Rosdakarya Offset, 2004.

[2] D. Rahayu, & U. Azizah, Pengembangan Instrumen Penilaian Kognitif Berbasis Komputer dengan Kombinasi Permainan "*Who Wants To Be A Chemist*" pada Materi Pokok Struktur Atom untuk Kelas X SMA RSBI. *Conference Proceedings of Chemistry UNESA 2012*, Surabaya: UNESA. ISBN: 978-979-028-550-7, 2012, pp.41-50.

[3] A. Jihad, & A. Haris, Evaluasi Pembelajaran, Yogyakarta: Multi Pressindo, 2012.

[4] Azhary, Analisis assessment soal ujian sekolah mata pelajaran bahasa indonesia di SMP Negeri 17 Palu. *E-Jurnal Bahasantodea* 4 (1), 2016, pp.39-47. Retrieved from http://jurnal.untad.ac.id/jurnal/index.php/Bahasantodea.

[5] T. Foltynek, "A new approach to the achievement test item evaluation: the correctness coefficient," *Journal on Effeciency and Responsibility in Education and Science* 2 (1), 2009, pp.28-40.

[6] Suwarto, "Pengembangan tes dan analisis hasil tes yang terintegrasi dalam program komputer," *Jurnal Penelitian dan Evaluasi Pendidikan* 12(1), 2009, pp.40-56. DOI: http://dx.doi.org/10.21831/pep.v13i1.1401.

[7] Sukardi. Evaluasi Pendidikan: Prinsip dan Operasionalnya, Jakarta: Bumi Aksara, 2012.

[8] D. Dibattista, & L. Kurzawa, "Examination of the quality of multiple choice items on classroom test," *The Canadian Journal for the Scholarship of Teaching and Learning* 2(2), 2011, pp.1-23. doi:10.5206/cjsotl.rcacea.2011.2.4.

[9] T. Miller, S. Chahine, N. E. Gronlund, Measurement and Assessment in Teaching 10th ed. Pearson Education Ltd. United States of America, 2009.

[10] S. E. Embretson, & R. C. Daniel, "Understanding and quantifying complexity level in mathematical problem solving items," *Psychology Science Quarterly* 50(3), 2008, pp.328-344.

[11] E. Rofiah, N. S. Aminah, E. Y. Ekawati, Penyusunan instrumen tes kemampuan berpikir tingkat tinggi fisika pada siswa SMP, *Jurnal Pendidikan Fisika* 1(2), 2013, pp.17-22.

[12] Tasiwan; S.E. Nugroho; Hartono, "Pengaruh *advance organizer* berbasis proyek terhadap kemampuan analisis-sintesis siswa," *Jurnal Pendidikan Fisika Indonesia* 10(1), 2014, pp.1-8.

[13] H. H. Scheiblechner, "Rasch and pseudo rasch models: suitableness for practical test and applications," *Psychology Science Quarterly,* 51(2), 2009, pp.181-194.

[14] N. Guler, G. K. Uyanik, G. T. Teker, "Comparison of classical test theory and item response theory in terms of item parameters," *European Journal of Research on Education.* International Association of Social Science Researsch-IASSR, 2 (1), 2014, pp.1-6. ISSN: 2147-6284.

[15] S. Senarat, S. Tayraukham, C. Piyapimonsit, S. Tongkhambanjong, "Developmnet of an item bank of order and graph by applying multidimensional item response theory*," Canadian Social Science*, 8(4), 2012, pp.21-27. Doi: 10.3968/j.css.1923669720120804.1263.

[16] R. K. Hambleton, H. Swaminathan, H. J. Rogers, Fundamental of Item Response Theory, Newbury Park: Sage Publication Ins., 1991.

[17] S. Golia, "The Assessment of DIF on rasch measures with an application to job satisfication," Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation, 1(1), 2010, pp.16-25. Doi: 10.1285/i2037-3627v1n1p16.

[18] N. S. Sukor, K. Osman, T.M.T. Soh,. "Chemistry test item development: assesing conceptual understanding among Malaysian students," *Asian Social Science,* 9(16), 2013, pp.126-132. Doi: 10.5539/ass.v9n16p126.

[19] B. Suminthono, & W. Widhiarso, Aplikasi Pemodelan Rasch pada Assesmen Pendidikan, Cimahi: Trim Komunikata Publishing House, 2015.

[20] W. W. Chiang, "Ninth grade student' self assessment in science: a rasch analysis approach," *Procidia Social and Behavioral Science,* 176, 2015, pp.200-210. Doi:10.1016/j.sbspro.2015.01.462.

[21] S. H. Ariffin, R. Idris, N. M. Ishak, "Differential item functioning in malaysian generic skills instruments (MyGSI)," *Jurnal Pendidikan Malaysia*, 35(1), 2010, pp.1-10. Retrieved from http://e.journal.ukm.my/jpend/article/view/13315

[22] I. F. Hidayati, & D. Rosana, "Penerapan rasch model berbasis irt dalam analisis soal UAS fisika SMA kelas XI menggunakan program *bigsteps* sebagai acuan pembuatan perangkat soal yang berkualitas," *Jurnal Universitas Negeri Yogyakarta* 2(5), 2013, pp. 1-7.

[23] Fahmi, "Perbandingan nilai ujian nasional dan ujian sekolah mata pelajaran matematika SMA program IPA tahun pelajaran 2010/2011," *Jurnal Pendidikan dan Kebudayaan* 17(6), 2011, pp.608-614.

[24] D. J. Ratnaningsih, & Isfarudi (2013), "Analisis butir tes objektif ujian akhir semester mahasiswa Universitas Terbuka berdasarkan teori tes modern," *Jurnal Pendidikan Terbuka dan Jarak Jauh* 14 (20): 2013, pp.98-109.

[25] M. D. Beer, "Use differential item functioning (DIF) analysis for bias analysis in test construction," SA Journal of Industrial Psychology,

30(4), 2004, pp.52-58. Retrieved from http://sajip.co.za/index.php/sajip.article/viewFile/175/172.

[26] K. Mulyana, "Karakteristik soal tes masuk SMP Negeri di Kabupaten Bantul," *Jurnal Penelitian dan Evaluasi Pendidikan* 10 (2), 2007, pp.235-248.

[27] Kustriyono, "Penyusunan perangkat soal ujian akhir mata pelajaran sains- biologi SMP dalam rangka pengembangan bank soal," *Jurnal Penelitian dan Evaluasi Pendidikan* 2 (6), 2004, pp.175-198.

[28] A. Sudijono, *Pengantar Evaluasi Pendidikan*, Jakarta: Rajawali Pers, 2015.

[29] D. Anggraini, & P. Suyata, "Karakteristik soal UASBN mata pelajaran bahasa indonesia di Daerah Istimewa Yogyakarta pada tahun pelajaran 2008/2009," *Jurnal Prima Edukasia* 2 (1), 2014, pp.57-65.

[30] M. Nurung, "Kualitas Tes Ujian Akhir Sekolah Berstandar Nasional (UASBN) IPA SD Tahun Pelajaran 2008/2009 di Kota Kendari". *Tesis*, Universitas Negeri Yogyakarta, 2008.

[31] S. Rogayah, & Ekaria, "Evaluasi taraf sukar butir tes matematika USM PMB STIS Tahun 2007/2008 dan tahun 2008/2009 dengan model Rasch," *Jurnal Aplikasi Statistika & Komputasi Statistik, UPPM-STIS* 2 (1), 2010, 76-91.

[32] B. Suminthono, & W. Widhiarso, Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial, Cimahi: Trim Komunikata Publishing House, 2015.