Development of the Academic Achievement Test for Undergraduate Students

Alfrits Roul Sinadia*
Faculty of Education
Universitas Klabat Airmadidi
Manado, Indonesia
*alfritssinadia@unklab.ac.id

Surya Jatmika
Faculty of Education
Universitas Muhammadiyah Surakarta
Surakarta, Indonesia
2sj795@ums.ac.id

***Abstract*—This study was conducted due to the in availability of a lecture-made test developed through a scientific process used to measure undergraduate students 'achievement. It aimed to develop a good quality test which could be used from semester to semester. As a developmental quantitative descriptive study, it analyzed the items of the test that were tried out on 54 undergraduate students. Their answers were analyzed in terms of validity, item difficulty, discrimination power, distracter effectiveness, and reliability. First, the validity analysis showed that all items were considered valid. Then the difficulty index analysis showed that more than a half of the items were categorized moderate but the rest were easy or difficult. Similarly, more than a half of the items were categorizing good in terms of item discrimination power. In terms of distracter effectiveness, it was found that some items had all well-functioning distracters; some others had three or two nonfunctioning distracters. Finally, the reliability estimation analysis showed that the consistency index of the test already met the minimum reliability index required. Based on the findings, the test developer is recommended to improve the items which need revisions or change them with new items for another tryout.**

Keywords: achievement test, item difficulty, discrimination power, distracter effectiveness

I. Introduction

At a university located in Manado city, North Sulawesi, Indonesia, an academic achievement test for undergraduate students was developed by one of its lecturers. The main reason of conducting this development was the subject taught by the lecturer (Educational Measurement & Evaluation) had no any sumative test for use that had been developed through quantitative item analyses to assess the quality of the test. For this reason, after being designed and tried out, the test items were analyzed in terms of validity, reliability, item difficulty, item discrimination power, and distracter effectiveness. Theoretically, the test items developed in this study went through a process of analyses called item analysis which is made up of the last three types of analysis [1]. This type of analysis is commonly conducted to make sure the items are qualified because the quality of a test is determined by the quality of its items [2].

Some previous studies in education and another field conducted similar analyses to analyze the quality of the test items used in educational measurement. In two previous researchs, Sharif et al., analyzed the multiple choice items of a university's examination items in terms of item difficulty and distracters function and, in a like manner [3], Odukoya et al., analyzed the multiple choice test items of compulsory university courses [4]. Reference Rehman et al., conducted quite similar analyses on the quality of a medical and dental undergraduate multiple-choice test through analyses on the following variables [5]: item difficulty index, item discrimination index, and distracter effectiveness (an additional variable which was not anayzed in the first two studies). Similarly, these studies generally conducted the analysis of validity, reliability, and item analysis that consisted of item difficulty, item discrimination power, and distracter effectiveness analyses which were also found in several local studies conducted by Iskandar [6], Suryani [7], and Akbar et al., [8]. In these cases, these studies similarly came up with two types of results: some items were already categorized qualified, but some others were not. In these studies, some items were rejected due to being too difficult or too easy [9], having poor discrimination power [10], or both of them: having too difficult or too easy items and poor discrimination power [11]. Another reason, for items to be disqualified is having ineffective distracters [12]. Differently, another study found that all distracters of the multiple choice items functioned well [13].

This study aimed to develop a qualified test built through a scientific quantitative process. Practically, the developed test would be continually used by the lecturer to measure students' academic achievement from semester-to semester and/or improve its items for future use like the test development conducted by Hofer et al., [14]. Specifically, qualified test items would be stored in a viable item bank and be used later in the future when needed [15]. In other words, the availability of the test will fill the gap of fact where there was no standardized test the lecturer could use to measure students'achievement at the end of semester over the last few years.

II. Methods

This study was a developmental study using quantitative approach and descriptive analyses. In the process of

Copyright © 2020 The Authors. Published by Atlantis Press SARL.
This is an open access article distributed under the CC BY-NC 4.0 license -http://creativecommons.org/licenses/by-nc/4.0/.

development, the test tryout involved 54 respondents, students who were taking the subject: Educational Measurement & Education during the 1st and 2nd semester for the school year 2018-2019. It was started from January to December 2018. The procedure of development was made up of several steps consisting of preparing the test blueprint, item construction, proving validity, item try out, and item analysis. The items were distributed on four Bloom's cognitive levels. These consisted of 7 items to measure knowledge, 26 items for comprehension, two items for application, and 13 items for analysis (total 48 items). The test blueprint later contained 11 types of competences measured using three types of test items, namely multiple choice, matching, and true/false items (see Table 1). Next, the test items were constructed based on the blueprint. On this stage, 30 multiple choice items (MC1-MC30), six matching items (MA1-MA6), and 12 true/false items were successfully constructed (TFA1-TFA6 and TFB1-TFB6). Then the items were analyzed in terms of face and content validation. This was conducted by three educational experts through Aiken's V analysis.

The next step was to try it out to the target respondents, undergraduate students who were taking the Educational Measurement and Evaluation subject. These respondents consisted of 54 students who were divided into two groups: 19 students took the test on the 1st semester and 35 students took the test on the 2nd semester of school year 2018-2019. Their answers were then tabulated and analyzed.

The analysis conducted to study the item difficulty levels referred to the following formula:

$$p = \frac{\text{Number of Examinees Correctly Answering the Item}}{\text{Number of Examinees}} \qquad (1)$$

The results of calculation were later interpreted using the following scales: < 0.30 = difficult; 0.30 – 0.70 = moderate; > 0.70 = easy [1]. Items whose difficulty levels indexes were lower than 0.30 or higher than 0.70 were removed because they were too difficult or too easy items.

The levels of item discrimination power were calculted using the following formula:

$$D = P_t - P_b \qquad (2)$$

D = Index of Discrimination Power
Pt = Proportion of Examinees in the Top Group Who Correctly Answer the Item
Pb = Proportion of Examinees in the Bottom Group Who Correctly Answer the Item

TABLE I. ITEM DISTRIBUTION ON COMPETENCIES

Competency	Item Type (Code)	Number of Items
Understanding of Measurement, Assessment, and Evaluation	Multiple Choice (MC)	4
Ability to Identify Different Measurement Scales	Multiple Choice (MC)	3
Application of Basic Mathematics of Measurement	Matching (MA)	6
Ability to Identify Different Kinds of Scores	Multiple Choice (MC)	3
Understanding of Reliability	Multiple Choice (MC)	2
Understanding of Validity	Multiple Choice (MC)	2
Understanding of Item Analysis	Multiple Choice (MC)	3
Understanding of How to Develop a Test	Multiple Choice (MC)	3
Ability to Identify Selected-Response Items	Multiple Choice (MC)	7
Ability to Identify Constructed-Response Items	True/False (TFA & TFB)	12
Understanding of Performance & Portfolio Assessment	Multiple Choice (MC)	3
Total Items		48

The number of examinees from the top and bottom group was determined by taking 27% of the examinees who have the highest scores and 27% of the examinees who have the lowest scores of the data which are normally distributed [16]. To interpret the results, a guide proposed by Hopkins [17] was used. It is categorized into five scales as folows: > 0.40 (very good), 0.30 – 0.39 (good); 0.11 – 0.29 (moderate); 0.00 – 0.10 (poor). Next, a distracter is categorized effective if it is chosen by at least five percent of the examinees. When there is less than 2.5 percent of examinees choose the distracter, it is categorized ineffective [18].

The test items which passed the screening analysis of difficulty level, discrimination power, and distracter effectiveness then went through the reliability analysis. It was conducted by estimating the test's coefficient Alpha which is suitable for both dichotomous or polytomous items [1]. The test is considered reliable if it has the reliability index of at least 0.70 [19]. Overall, each item quality was evaluated in terms of the previously stated characteristics such as its validity, item difficulty, discrimination power, distracter effectiveness, and reliability (conducted for the whole items).

III. RESULTS AND DISCUSSION

In the content validity analysis, results of expert judgement showed that all items (48 items) had Aiken's V values between 0.92 and 1.00 or they were all valid to measure all the constructs they had to measure. Next, through the item analysis, some items were found having been qualified, but some others were not. First, results of item difficulty analyses showed that there were 30 items whose p indexes fell between 0.30 and 0.70 (21 multiple choice items, two matching items, and seven true/false items) or under the category of items with moderate difficulty level. Out of the 48 items, 12 items were categorized easy because their p values were higher than 0.70.

On the other, six items had the p values less than 0.30 or fell under the category of difficult items. Items with moderate difficulty levels are considered qualified to use, but those which were considered easy and difficult were changed with new items or totally eliminated because they are not able to indicate the existing difference among examinees [16].

TABLE II. ITEM DIFFICULTY LEVEL AND DISCRIMINATION POWER

No	Item Code	Difficulty Level (p)	Description of p	Discrimination Power (D)	Description of D
1	MC1	0.80*	Easy*	0.30	Good
2	MC2	0.39	Moderate	0.57	Very Good
3	MC3	0.48	Moderate	0.71	Very Good
4	MC4	0.54	Moderate	0.57	Very Good
5	MC5	0.54	Moderate	0.30	Good
6	MC6	0.41	Moderate	0.43	Very Good
7	MC7	0.30	Moderate	0.57	Very Good
8	MC8	0.35	Moderate	0.57	Very Good
9	MC9	0.43	Moderate	0.57	Very Good
10	MC10	0.63	Moderate	0.57	Very Good
11	MC11	0.41	Moderate	0.71	Very Good
12	MC12	0.69	Moderate	0.71	Very Good
13	MC13	0.69	Moderate	0.71	Very Good
14	MC16	0.52	Moderate	1.00	Very Good
15	MC17	0.41	Moderate	0.57	Very Good
16	MC18	0.37	Moderate	0.30	Good
17	MC21	0.63	Moderate	0.71	Very Good
18	MC22	0.31	Moderate	0.43	Very Good
19	MC28	0.30	Moderate	0.43	Very Good
20	MA3	0.43	Moderate	0.83	Very Good
21	MA4	0.35	Moderate	0.67	Very Good
22	TFA1	0.31	Moderate	0.86	Very Good
23	TFA5	0.44	Moderate	0.71	Very Good
24	TFA6	0.63	Moderate	0.30	Good
25	TFB2	0.44	Moderate	0.57	Very Good
26	TFB3	0.74	Moderate	0.43	Very Good
27	TFB4	0.30	Moderate	0.43	Very Good
28	TFB5	0.70	Moderate	0.30	Good

Similarly, the results of discrimination power analyses showed that most items had very good discrimination power. Out of 48 items, 32 items were categorized having very good discriminating power (*D* values were higher than or equal to 0.40). Seven items were categorized good (having *D* values between 0.30 and 0.39) and the other nine items were disqualified because of having *D* values which were lower than 0.30. Some of them even had negative *D* values indicating that there were flaws during the process of determining the answer key or other kinds of flaws.

Before moving to distracter analysis, the items were screened in terms of item difficulty and discrimination power indexes. After having screened the items, 28 combined items (19 multiple choice items, two matching items, and seven true/false items (see Table 2). As a special note, item MC1 was an easy item (*p = 0.80*) which was intentionally put in the beginning of the test in order to motivate examinees work on the rest items [16]. After the screening process, the distracters of 19 multiple choice items were analyzed in terms of their function effectiveness to distract examinees from choosing the correct answers. When a distracter was chosen by at least 2.50% of examinees, that distracter is considered effective [20]. Reffering to this reference, the results of analysis showed that six items had all functioning distracters (MC5, MC6, MC9, MC16, MC22, and MC28) (see Table 3). Secondly, eight items had three functioning distracters (MC2, MC3, MC4, MC7, MC10, MC12, MC13, and MC18). Finally, the other five items had only two functioning distracters. These consisted of MC1, MC8, MC11, MC17, and MC21. In conclusion, all nonfunctioning distracters of 13 items need to be changed (for distracters chosen by nobody) or improved (for distracters chosen by less than 2.50% examinees).

Reliability analysis was the last type of analysis conducted on the 28 items. Using the analysis of coefficient Alpha, it was found that the reliability index of the test was 0.64 which means that there is a measurement error possibility of 0.36. This figure was lower than the minimal reliability index required for a test, that is 0.70 with the error measurement possibility of 0.30. Consequently, five items which had high measurement errors were removed out of the 28 items. The items were made up of TFA6, TFB2, TFB3, TFB4, and TFB5. On the remaining 23 items, the coefficient Alpha analysis was conducted again and the result showed that the reliability index of the test had changed to 0.71 or had met the minimum requirement of reliability.

TABLE III. EFFECTIVENESS OF MULTIPLE CHOICE ITEM DISTRACTERS

No	Item	Percentage of Choosing Distracters					Functioning Distracters				
		A	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
1	MC1	7	0	*	2	9	Yes	-	*	-	Yes
2	MC2	*	37	20	4	2	*	Yes	Yes	Yes	-
3	MC3	20	22	*	15	2	Yes	Yes	*	Yes	-
4	MC4	2	37	4	7	*	-	Yes	Yes	Yes	*
5	MC5	*	20	9	15	4	*	Yes	Yes	Yes	Yes
6	MC6	28	*	22	9	4	Yes	*	Yes	Yes	Yes
7	MC7	22	20	*	37	0	Yes	Yes	*	Yes	-
8	MC8	50	17	2	2	*	Yes	Yes	-	-	*
9	MC9	28	*	4	4	33	Yes	*	Yes	Yes	Yes
10	MC10	2	4	22	4	*	-	Yes	Yes	Yes	*
11	MC11	54	*	9	2	0	Yes	*	Yes	-	-
12	MC12	24	7	*	0	4	Yes	Yes	*	-	Yes
13	MC13	17	4	9	2	*	Yes	Yes	Yes	-	*
14	MC16	7	24	*	11	9	Yes	Yes	*	Yes	Yes
15	MC17	*	48	2	0	7	*	Yes	-	-	Yes
16	MC18	41	*	13	7	2	Yes	*	Yes	Yes	-
17	MC21	24	2	13	2	*	Yes	-	Yes	-	*
18	MC22	41	15	*	13	7	Yes	Yes	*	Yes	Yes
19	MC28	20	7	*	11	35	Yes	Yes	*	Yes	Yes

In summary, out of the 48 test items constructed in the beginning, there were 28 items (19 multiple choice items, two matching items, and seven true/false items) which were qualified in terms of item difficulty and discriminating power. Next, out of the 19 multiple choice items, there were six items with all functioning distracters, however, there were 13 items which had two or three non-functioning distracters which need to be changed or improved. In order to acquire enough reliability coefficient on the test, five items which potentially lower the reliability index down to 0.70 were removed from the test. Finally, there were 23 items which had met the qualifications as follows: the item difficulty indexes ranged between 0.30 to 0.70 or having moderate item difficulty levels; the discriminating power indexes ranged between 0.30 to 1.00 or being categorized as items with good or very good discriminating power; and when combined as a test, it had enough reliability index. Having met the qualifications, the 23 items were ready for use in measurement activities (see Table 4).

Based on the results and discussion stated previously, a few conclusions were drawn. First, out of the 28 items which passed the early screening process, six multiple choice items with all distracters function effectively had met the minimum qualifications that they are ready for use in a measurement process. Similarly, two matching items and two true/false items which met the minimum qualifications are also ready for use. Secondly, based on the results and discussion stated previously, a few conclusions were drawn. First, out of the 28 items which passed the early screening process, six multiple choice items with all distracters function effectively had met the minimum qualifications that they are ready for use in a measurement process. Similarly, two matching items and two true/false items which met the minimum qualifications are also ready for use. Secondly, the other 13 multiple choice items can only be used after having improved their nonfunctioning distracters. The same treatment need to be done toward the five true/false items removed because of the potential to lower the reliability coefficient due to some measurement errors; to be qualified, it is recommended that these items are improved. Alternatively, some new paralleled-content items can be added with a purpose to lift up the reliability coefficient because, theoretically, an increase of reliability coefficient tend to be followed by an increase of validity index of a test.

TABLE IV. ITEM ANALYSIS RESULTS

No	Item Code	*p* Acceptable	*D* Good	Description of Distracter Function
1	MC1	Yes	Yes	Two Need Improvement
2	MC2	Yes	Yes	One Needs Improvement
3	MC3	Yes	Yes	One Needs Improvement
4	MC4	Yes	Yes	One Needs Improvement
5	MC5	Yes	Yes	All Distracters are Effective
6	MC6	Yes	Yes	All Distracters are Effective
7	MC7	Yes	Yes	One Needs Improvement
8	MC8	Yes	Yes	Two Need Improvement
9	MC9	Yes	Yes	All Distracters are Effective
10	MC10	Yes	Yes	One Needs Improvement
11	MC11	Yes	Yes	Two Need Improvement
12	MC12	Yes	Yes	One Needs Improvement
13	MC13	Yes	Yes	One Needs Improvement
14	MC16	Yes	Yes	All Distracters are Effective
15	MC17	Yes	Yes	Two Need Improvement
16	MC18	Yes	Yes	One Needs Improvement
17	MC21	Yes	Yes	Two Need Improvement
18	MC22	Yes	Yes	All Distracters are Effective
19	MC28	Yes	Yes	All Distracters are Effective
20	MA3	Yes	Yes	-
21	MA4	Yes	Yes	-
22	TFA1	Yes	Yes	-
23	TFA5	Yes	Yes	-

IV. CONCLUSION

In terms of distracter effectiveness, it was found that some items had all well-functioning distracters; some others had three or two nonfunctioning distracters. Finally, the reliability estimation analysis showed that the consistency index of the test already met the minimum reliability index required. Based on the findings, the test developer is recommended to improve the items which need revisions or change them with new items for another tryout

REFERENCES

[1] C.R. Reynolds, R.B. Livingstone, and V. Willson, Measurement and Assessment in Education, 2nd ed., New Jersey: Pearson Education, 2009.

[2] S. Azwar, Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar, 2nd ed., Yogyakarta: Pustaka Pelajar.

[3] M.R. Sharif, M.H. Asadi, A.R. Sharif, and M. Sayyah, “Examining the multiple choice educational examinations of college students,” Biomed. Pharmacol. J., vol. 6 (1), pp. 23-27, 2013.

[4] J.A. Odukoya, O. Adekeye, and A.O. Igbinoba, “Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university,” Qual. Quant., vol.52, pp. 983-997, March 2017.

[5] A. Rehman, A. Aslam, and S.H. Hasan, “Item analysis of multiple choice questions,” Pakistan Oral Dent. J., vol. 38 (2), pp. 291-293, April-Juni 2018.

[6] A. Iskandar and M. Rizal, “Analisis kualitas soal di perguruan tinggi berbasis TAP,” J. Pene. Eva. Pend., vol 21 (2), pp. 12-23, December 2017.

[7] Y. E. Suryani, “Pemetaan kualitas soal ujian akhir semester pada mata pelajaran bahasa Indonesia SMA di Kabupen Klaten,” J. Pene. Eva. Pend., vol 21 (2), pp. 142-152, December 2017.

[8] M. N. Akbar, H. Firman, and L. Rusyati, “Developing science virtual test to measure student’s critical thinking on living things and environmental sustainability theme,” J. Phys.: Conf. Ser. 812012106, 2017.

[9] S. M. J. A. Marie and E. Sreekala, “Relevance of item analysis in standardizing an achievement test in teaching of physical science in B.Ed syllabus,” i-manager’s J. Edu. Techno., vol 12 (3), pp. 30-36, October-December 2015.

[10] C. Boopathiraj and K. Chellamani, “Analysis of test items on difficulty level and discrimination index in the test for research in education,” Inter. J. of Soc. Scie. & Int. Res., vol. 2 (2), pp. 189-193, February 2013.

[11] A.A. Danuwijaya, “Item analysis of reading comprehension test for post-graduate sutdents,” English Rev.: J. of English Edu., vol. 7 (1), pp. 29-40, December 2018.

[12] S. Toksöz and A. Ertunç, “Item analysis of a multiple-choice exam. Adv. Lang. Lit. Stud., vol. 8 (6), pp. 141-146, December 2017.

[13] M. Kusumawati and S. Hadi, “An analysis of multiple choice questions (MCQs): item and test statistics from mathematics assessments in senior high school,” Res. Eva. Edu., vol. 4 (1), pp. 70-78, November 2018.

[14] S.I. Hofer, R. Schumacher, and H. Rubin, “The test of basic mechanics conceptual understanding (bMCU): using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton’s mechanics,” Inter. J. STEM Edu., vol. 4 (18), pp. 1-20, 2017.

[15] G. Mehta and V. Mokhasi, “Item analysis of multiple choice questions—an assessment of the assessment tool,” Int. J. Health Scie. Res., vol. 4 (7), pp. 197-202, July 2014.

[16] M.J. Allen and W.M. Yen, Introduction to Measurement Theory, California: Wadsworth, 1979.

[17] K. D. Hopkins, Educational and Psychological Measurement and Evaluation, 8th ed., Boston: Allyn & Bacon, 1998.

[18] R. Zulaiha, Bagaimana Menganalisis Soal dengan Program Iteman, Jakarta: Puspendik.

[19] D. Mardapi, Teknik Penyusunan Tes dan Nontes, Yogyakarta: Mitra Cendikia, 2008.

[20] D. Mardapi, Pengukuran, Penilaian, & Evaluasi Pendidikan, Yogyakarta: Nuha Medika, 2012.