Eurasian Journal of Educational Research 76 (2018) 187-202

Eurasian Journal of Educational Research
www.ejer.com.tr

Analytical Weighting Scoring for Physics Multiple Correct Items to Improve the Accuracy of Students' Ability Assessment

Wasis[1] , Kumaidi[2], Bastari[3], Mundilarto[4], Atik Wintarti[5]

ARTICLE INFO

Article History:

Received: 31 May 2018

Received in revised form: 9 Jun. 2018

Accepted: 5 Jul. 2018

DOI: 10.14689/ejer.2018.76.10

Keywords

Analytical weighting scoring, accuracy of estimation, physics aptitude

ABSTRACT

Purpose: This is a developmental research study that aims to develop a model of polytomous scoring based-on weighting for multiple correct items in the subject of physics. Weighting was analytically applied based on question complexity and imposed penalties on wrong answers. **Research Methods:** Within the development model, Fenrich's development cycle, consisting of analysis, planning, design, development, implementation, evaluation, and revision, was selected throughout the cycle. The multiple correct items used have 3–4 options. The items were implemented to 140 upper secondary school students and 410 first-year undergraduate students. The students' physics ability was analyzed using the Quest program, and the results of dichotomous and polytomous scoring were compared. **Findings:** The results of this study showed that the analytical weighting scoring based on a complexity and penalty system on the developed assessment items generated a higher number of scoring level categories (three to seven categories) than that of dichotomous scoring (only two categories), estimated students' physics abilities more accurately and in greater detail, had an approximate distribution closer to the normal distribution, and produced a standard deviation smaller than that of dichotomous scoring. Thus, the analytical weighting scoring for multiple correct items in this study was able to produce a more accurate estimation of physics ability than those using dichotomous scoring. **Implications for Research and Practice**: It is recommended that the assessment of physics ability using multiple-correct items on a large scale can apply the analytical weighting scoring based on the complexity of the content and a penalty system.

[1] Corresponding Author: Department of Physics, Universitas Negeri Surabaya, INDONESIA. e-mail: wasis@unesa.ac.id, ORCID: 0000-0002-4437-5141

[2] Departement of Psychology, Universitas Muhammadiyah Surakarta, INDONESIA. e-mail: kum231@ums.ac.id, ORCID: 0000-0002-2736-6139.

[3] Balitbang Kemendikbud, INDONESIA. e-mail: bastari@kemdikbud.go.id, ORCID: 0000-0003-3442-3063.

[4] Department of Physics, Universitas Negeri Yogyakarta, INDONESIA. e-mail: mundilarto@uny.ac.id, ORCID: 000-0003-2891-4317.

[5] Department of Mathematics, Universitas Negeri Surabaya, INDONESIA. e-mail: atikwintarti@unesa.ac.id, ORCID: 0000-0001-6514-1881.

Introduction

The study of scoring on selected-response items is still continuing in educational assessment in recent decades. Previous studies have examined this issue for the following purposes: to evaluate teachers' competency (Martin & Itter, 2014), to develop construction and scoring on selected-response items (Emaikwu, 2015), to compare the assessment using the question of multiple choice and constructed response (Hickson, Reed, & Sander, 2012; Stankous, 2016), to obtain immediate feedback assessment techniques (Merrel et al., 2015), and to propose a more comprehensive framework of writing a more sophisticated format of selected response items in hopes of reduced guesswork (Bush, 2015). However, all of those purposes are designed to improve the validity and reliability of the selected response items developed within those studies (Ali, Carr & Ruit, 2016)

Despite the various weaknesses of selected-response assessment items, including the issue of validity and reliability, facts show that this method of testing is still dominantly used, especially in large-scale tests with speedily delivered results (Oosterhof, 2003; Rodriguez, 2005; Merrel et al., 2015). This is because the execution of the test with selected-response items takes less time, requires a quick process, is easy to score, and has a high degree of objectivity (Wooten et al., 2014; Baghaei & Dourakhshan, 2016).

However, the use of selected-response items to measure, especially, the mastery of physics ability, has a number of weaknesses, particularly when the scoring is done using a dichotomous model. The topics of physics are known as a continuum, ranging from very simple to very complicated cognition, so the assessment instrument of physics ability, should include a representative scope of content and competence (Klein, et. al., 2017). Thus, when a person seeks to understand or master the subject of physics, his/her understanding or mastery can be in a position among all the possible positions in the continuum. The person's ability, position can be anywhere and is not limited to the lowest or highest position. Therefore, if a person's ability is assessed, then the result of his judgment cannot be forced to be dichotomous, which is to say high or low. Furthermore, the problem-solving in physics also has a number of stages (Adeyemo, 2010). Therefore, if a person's ability to solve a physics problem is assessed, the results of his or her judgment should not only be indicated in the final result but should demonstrate a knowledge of all stages leading up to find the outcome. The results of the assessment should be able to describe the student's achievements at each stage. Moreover, the phases of solving physics problems generally reflects various levels of complexity. When a person is able to complete a stage, then the assessment of his or her performance should also vary depending on the weight of its complexity. Based on the reasons listed above, the scoring of physics ability should be done analytically with regard to the results of each stage. According to Wiseman (2012) analytical scoring provides better assessment results than holistic scoring.

The use of selected-response items in assessing physics ability often causes students to answer the questions by guessing. To reduce the inaccuracies that may result from this practice, the incorrect answer needs to be punished under the

understanding that one should be cautious because penalties for incorrect answers could intimidate students (Holt, 2006). In analytical scoring, penalties can be integrated by weighting of each stage of solving physics problems. Therefore, another scoring model of selected response questions should be developed to test physics ability that considers completion stages, the weighting of each of those stages, and the responsibility to answer questions. It is hoped that with such scoring patterns that even though the assessment uses selected responses item, the judgment of the results will have many categories, or in other words, that they may be polytomous. Bond & Fox (2007) stated that adding a number of categories to the scoring will increase the reliability of the measurement. If the measurement of the results is reliable, then the assessment based on the results will be more accurate.

The current study aims to develop a polytomous scoring model based on weighting, focusing on complexity and assigning penalties for the incorrect answer of each stage of solving a given physics problem. In order for each stage to be analytically scored, an alternative multiple-choice question, or so-called as multiple-correct item, was selected. Questions in multiple-correct items are more efficient because they can generate better memory retention and provide more correct information (Bishara & Lanzo, 2015). Also, one question can be designed to measure multiple capabilities of the same dimensions (Haladyna et al., 2002), and it was also empirically proven to be more reliable than multiple choice items (Frisbie, 1992). Thus, the results of this study are expected to produce a model of polytomous scoring for multiple correct items so as to be able to estimate students' ability in physics more accurately.

Method

Research Design

This is a development research study following Fenrich's model (2004) (see Figure 1). The result of this development is a polytomous scoring model for multiple correct items. The model encompasses the phases of analysis, planning, design, development, and implementation; each of which provides an opportunity for possible evaluation and revision.

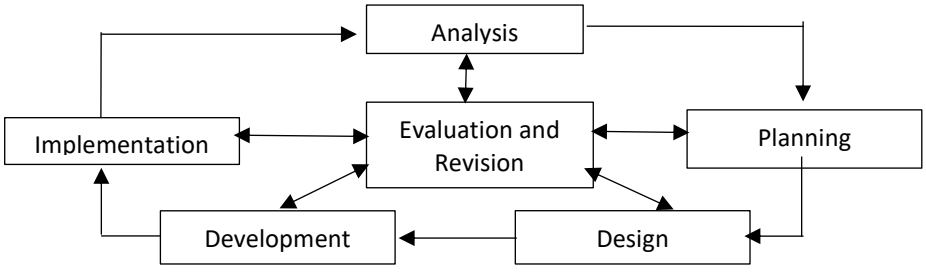


Figure 1. Fenrich model development cycle (Fenrich, 2004)

The analysis was carried out by focusing on the scoring method, which uses a partial-credit model instead of the dichotomous scoring model that only considers two options: Yes or No. The use of a partial-credit model on multiple choice questions has been investigated by several scholars (see e.g. Wright & Masters, 1982; Grunert et al., 2013 and King et al., 2004). However, the scores of each option are based on partial assumptions, for example 0.25, 0.50, 0.75, and 1. This study's analytical scoring is designed according to the weight of the complexity of each option and imposes penalties for incorrect answers. This relates to Ayedemo's (2010) statement suggesting that problems in physics typically require multi-stage problem-solving strategies. Therefore, the scoring needs to consider the weight of each stage. Also, this follows Holt's (2006) suggestion that the prudence of the punishment can reduce the carelessness of students in guessing as they answer multiple choice questions.

The scoring design described above can be applied to multiple-correct items, i.e. multiple-choice items with more than one correct option. This item format was selected because it can measure the number of stages of problem solving within physics. With multiple-correct item formats, it is possible that the stem of the problem in the items is in the form of a physics problem while the options are in the form of stages of problem solving. All options can be designated as being correct or incorrect. Based on previous research, it has been proven that multiple-correct items are more reliable and provide more advantages than traditional multiple-choice items (with only one correct option) (Frisbie, 1992; Haladyna et al., 2002; and Bishara & Lanzo, 2015).

The process of evaluation and revision were also carried out reciprocally throughout the development and test phases as shown by the two-way arrows in Figure 1. The results of the analysis emphasize the planning, design, and development of the problem along with the scoring model. The results of the four steps were evaluated by the experts, which in this case are physics lecturers through the validation mechanism. Validation results were used to revise the design of problems in the multiple correct items developed as well as their scoring. When the implementation phase was conducted, the activities of evaluation and revision were also conducted so that the representative respondents, either senior high school students (high, middle, and low) or undergradute students (physics, chemistry, and biology majors) were equally obtained.

Research Sample

The sample for this study were 140 high school students, coming from three different high schools in Surabaya, Indonesia and 410 first-year undergraduate students from four different majors within the Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia.

Research Instruments and Procedures

There were two types of multiple-correct item format developed in this study: items with three options and items with four options. The number of items for each

type is fifteen. The following is an example of a multiple-correct item with three options complete with its weighted analytical scoring model developed in this study.

A 200-gram bullet is fired vertically upward from the ground with a speed of 60 m/s. If the gravitational acceleration g = 10 m/s², then:
(a) The maximum height reached by the bullet is 180 meters
(b) At the highest point, the energy of the bullet is 360 joules
(c) At its highest point, the bullet has neither speed nor acceleration

Because this is in the form of a multiple-correct item, then each option is likely to be true and the participants are permitted to choose more than one option. The analytic scoring guide for each of the above responses is shown in Table 1.

Table 1

Analytical Scoring Guide

Completion stage	Analytic score for each stage	Total score for each option
The bullet reaches its highest point when $v_t = 0$	1	3
The maximum level reached is given by $\frac{V_0^2}{2g}$ $h_{\max} = \frac{60^2}{2.10} = 180$ meter	1	
➔ Proposition given in option (a) is correct	1	
At the highest level, the bullet will not be moving; thus, for a moment, $v_t = 0$, so $E_{K\ highest} = 0$.	1	3
The bullet energy's at its highest level = potential energy = initial kinetic energy. $E_{highest} = E_p = mgh = (0.2)(10)(180) = 360$ joule	1	
or $E_{highest} = E_{Ki} = ½\ mv^2 = (½)(0,2)(60)^2 = 360$ joule	1	
➔ Proposition given in option (b) is correct		
The object moves vertically up and then moves down to the earth. At its highest point, the object has no movement for a moment, so it has no speed.	1	2
The object has a constant acceleration, including when it is at its highest point, that is, gravitation acceleration.	1	
➔ The proposition given in option (c) is incorrect		

Based on the above scoring guidelines, if the test taker chooses option (a), he/she gets a score of 3; does not select option (a), gets a score (-3); selects option (b), gets a score of 3; does not select option (b) gets a score (-2); selects option (c), gets a score of

(-2); and does not select option (C), gets a score of 2 because the proposition in option (c) is incorrect. The following scoring categories are obtained for all possible response patterns to this question as follows:

Table 2

Polytomous Scoring Based on Analytical Weighting and Penalty System

Pattern of responses			Score for each option			Total score	Category
(a)	(b)	(c)	(a)	(b)	(c)		
✓			3	-3	2	2	4
	✓		-3	3	2	2	4
		✓	-3	-3	-2	-8	1
✓	✓		3	3	2	8	6
✓		✓	3	-3	-2	-2	3
	✓	✓	-3	3	-2	-2	3
✓	✓	✓	3	3	-2	4	5
			-3	-3	2	-4	2

Table 2 shows that in the polytomous scoring model developed in this study, the ability of test participants can be assigned to one of six categories, namely categories 1 through 6. Whereas if using dichotomous scoring, there are only two possible categories, namely score 1 for participants who choose options (a) and (b) but not (c) and score 0 for participants who choose any other combination.

The developed model of polytomous scoring, as illustrated in Table 1, has the following stages: (a) developing analytic scoring guidelines for each problem-solving step by considering the weight of its complexity, (b) identifying all possible response patterns, (c) scoring based on the response patterns and the assigned weight of each stage, (d) assigning a positive score for a correct answer and a negative score for an incorrect answer, (e) calculating the total score by adding the scores of each option/step, and (f) correlating the total score equivalent to a category or level.

Data Analysis

After the responses of the test participants were categorized according to the developed model, they were then analyzed by the Quest program (Adams & Khoo, 1996) in order to estimate the participants' abilities (θ) and to obtain the standard estimation error. The Quest program was selected because it is relatively simple due to its being limited to only one parameter of items, which is the Rasch model which applies an item response theory approach. The IRT approach and the Rasch model were chosen because this a study that assumes that guessing is a part of a student's ability and that all items that fit the model have equivalent discriminations so that items are only described by a single parameter (1 PL).

Results

Before analyzing the participants' aptitude and the difficulty of the items using Quest (Adams & Khoo, 1996), the authors first conducted the so-called 'fit item analysis' using an infit and outfit mean square and an infit and outfit *t*, as shown in Table 3.

Table 3

Fit Item Analysis

Scoring	Infit mean square	Outfit mean square	Infit *t*	Outfit *T*
Dichotomous	1.00	1.04	0.10	0.27
Polytomous	1.00	1.00	-0.02	-0.04

The data or responses are said to fit within the Rasch model when the infit and outfit mean's square value is close to 1 and the infit and outfit *t* are near 0 (Adams & Khoo, 1996). Table 3 shows that the responses of the test participants used in this study fit within the Rasch model, although the *fit* of dichotomous scores was lower than that of the polytomous scores.

The Quest program's estimations for a number of the upper secondary school student participants are shown in Table 4, and the summary of the mean is presented in Table 5.

Table 4

The of Physics-Aptitude Estimation of Several Participants

Dichotomous Scoring					Polytomous Scoring				
Name	Score	Score Max	Estimate	Error	Name	Score	Score Max	Estimate	Error
14	3	15	-1.62	0.73	14	23	69	-0.46	0.22
27	3	15	-1.62	0.73	27	37	69	0.15	0.21
32	3	15	-1.62	0.73	32	36	69	0.11	0.21
33	3	15	-1.62	0.73	33	32	69	-0.06	0.21
68	3	15	-1.62	0.73	68	33	69	-0.02	0.21
76	3	15	-1.62	0.73	76	20	69	-0.60	0.23
78	3	15	-1.62	0.73	78	29	69	-0.19	0.21
81	3	15	-1.62	0.73	81	30	69	-0.15	0.21
82	3	15	-1.62	0.73	82	29	69	-0.19	0.21
83	3	15	-1.62	0.73	83	25	69	-0.37	0.21
103	3	15	-1.62	0.73	103	29	69	-0.19	0.21
105	3	15	-1.62	0.73	105	27	69	-0.28	0.21
111	3	15	-1.62	0.73	111	30	69	-0.15	0.21
112	3	15	-1.62	0.73	112	32	69	-0.06	0.21
118	3	15	-1.62	0.73	118	42	69	0.37	0.21
140	3	15	-1.62	0.73	140	27	69	-0.28	0.21

Based on the data in Table 4, it can be seen that the polytomous scoring method developed for this study yields a better capability to estimate the participants' physics ability than the estimations made using dichotomous scoring. This can be seen from the ability to estimate the ability of test participants in more detail. Of the 140 participants in the upper secondary school sample, sixteen students in the dichotomous scoring had the same estimated aptitude of -1.62 logits, but on the polytomous scoring of those sixteen students, the participants had different aptitude estimates ranging from -0.60 to 0.37 logits. On the other hand, of the 410 of undergraduate-student participants, as many as 151 participants are estimated to have the same aptitude, measured at -3.41 logits, the lowest ability. Meanwhile, when their abilities were scored using polytomous scoring, there was only 1 person who had the lowest aptitude, while the others had a higher estimation of their abilities measured by various scores.

Polytomous scoring provides a more accurate estimation of the participants' capabilities. This can be seen from the standard deviation (SD) values as shown in Table 5.

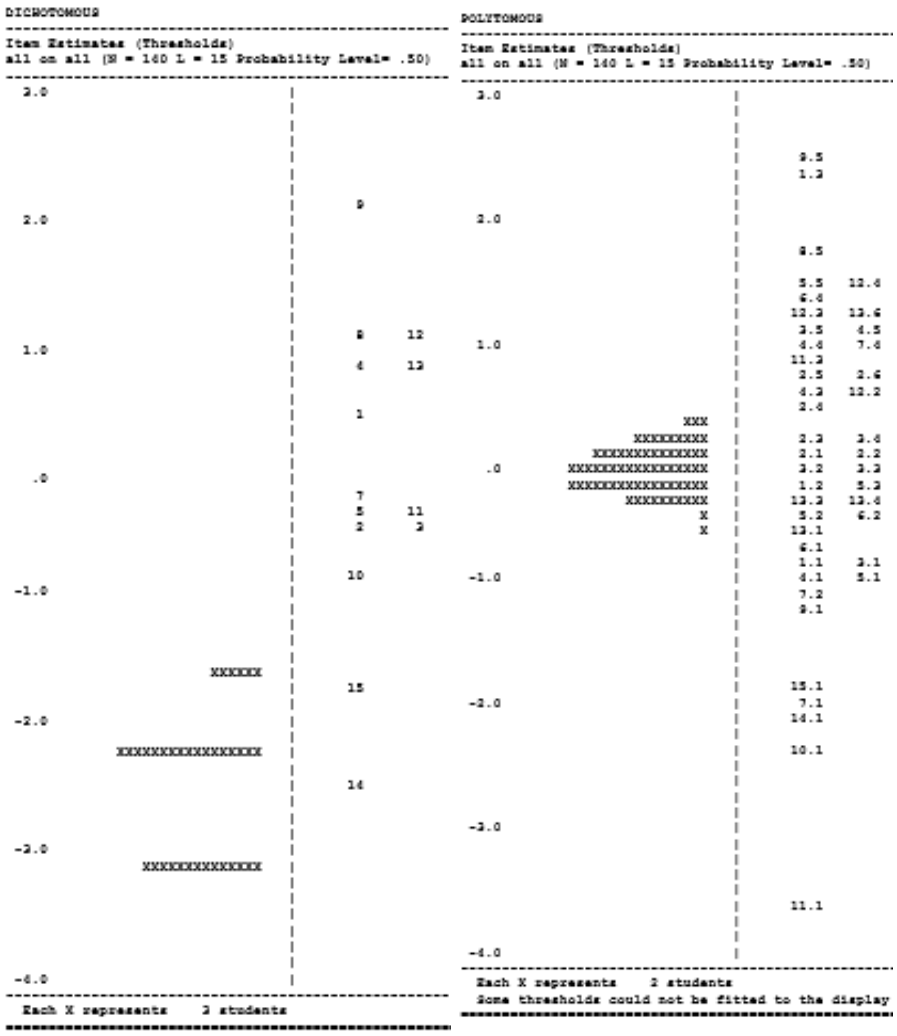
Table 5

Means of Estimation and Standard Deviation (SD) for the Participants' Ability

Participants	Dichotomous scoring			Polytomous Scoring		
	Number of categories	Mean of ability	SD	Number of categories	Mean of ability	SD
Upper secondary students	2	-2.47	0.56	3–7	-0.07	0.22
Undergraduate students	2	-2.76	0.81	5–10	-0.08	0.18

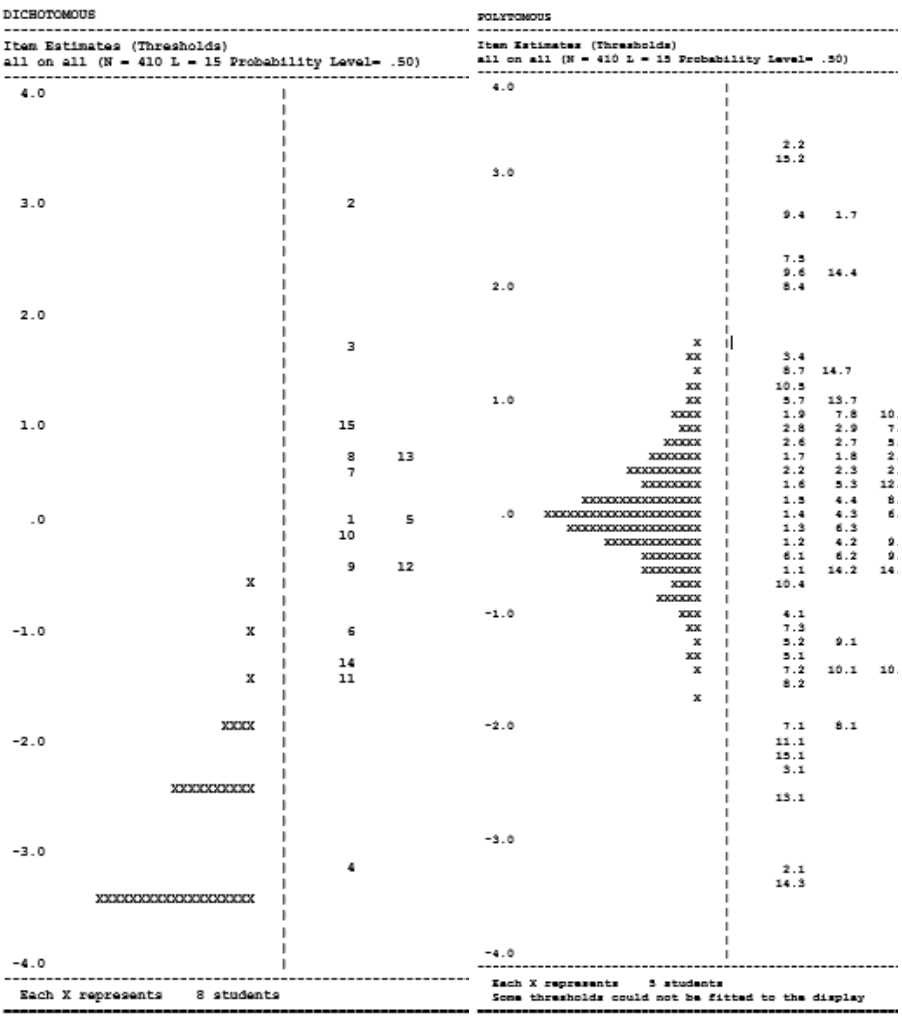
Table 5 shows that the standard deviation (SD) of ability in the polytomous scoring, which has more categories, is always smaller than the standard deviation resulting from dichotomous scoring. Those results were consistently obtained for both upper-secondary and undergraduate-student participants. Based on the above results, it can be concluded that the greater number of categories in a scoring system, the smaller the standard deviation of estimation is, meaning that the result of the assessment will contain fewer errors and be closer to the actual ability.

Quest analysis also maps the estimation of the students' ability and item difficulty for dichotomous and polytomous scoring, as shown in Figures 2 (a) and (b) for upper secondary school student and Figures 3 (a) and (b) for first-year undergraduate student participants.



(a) (b)

Figure 2. The mapping of the estimation of the upper secondary school students' ability and the item difficulty for (a) dichotomous scoring and (b) polytomous scoring



(a) **(b)**

Figure 3. The mapping of the first-year undergraduate student s' ability and level of difficulty of the items based on the results of (a) dichotomous scoring and (b) polytomous scoring

Discussion, Conclusion, and Recommendations

Overall, the results of this analysis as indicated by Table 4 show that polytomous scoring's having more categories than dichotomous scoring makes the measurement results more accurate. This is analogous to measuring tools used in everyday life: i.e. the more strips or scale lines that a measuring device has, the more thorough the

measurement. On a ruler a distance of 1 mm is represented only by two lines, while on a slide rule, a distance of 1 mm is subdivided into 10 or 20 lines. Thus,the measurement resulting from the slide rule can become more accurate than that of the ruler. Bond & Fox (2007) stated that scoring within many categories can estimate the ability of test participants better than an assessment with fewer categories. Donoghue (2005), in an article on reading scores, reported that polytomous scoring yields an average of about two to three times the average function of dichotomous scoring information. The greater the information function, the more accurate the resulting estimate (Hambleton et al., 1991, Lin, 2008).

As mentioned above, the improvement of accurary in scoring physics ability assessments can also be claimed as the contribution of the analysis of the suitability of the developed scoring models with the items formats and characteristics of the physics content. Problem solving is a fundamental part of studying physics, and the students do not solve most problems at the desired level of proficiency (Redish et al., 2006). When the stages of problem-solving are formulated as self-selected options in multiple-correct items and each answer in the option is analytically assessed based on the degree of difficulty, the student's ability can be measured in more detail than if judged only by dichotomous scoring based on the final result of their problem solving. Haladyna et al. (2002), who reviewed various forms of multiple-choice questions, stated that in the development of choice questions, it is not only important to consider the content examined in the questions developed but also the methodological viewpoints used in the development process. If multiple-choice questions are methodically designed to analytically record the students' thinking processes and record such stages, then such questions will be able to assess students more accurately.

Figures 2 and 3 show that the estimation of the participants' aptitude under polytomous scoring tends to illustrate the real condition of a group of participants because it has a normal distribution with an average value close to 0, whereas the mean estimation of participants' ability under dichotomous scoring tends to tilt to the left. The polytomous item-person maps shown in Figures 2 and 3 demonstrate how the test is well matched to the sample (Bond & Fox, 2007). It is also known that the estimation of item difficulty by using polytomous scoring is more accurate and detailed than the dichotomous scoring results although both are relatively consistent in estimating the item difficulty of the most difficult and easiest items. Figures 2 and 3 also demonstrate that the estimation of the degree of difficulty of the problem within polytomous scoring stretches further and generates a more detailed report than that of dichotomous scoring. Like a ruler, the polytomous scoring system with the analytical weighting approach developed in this study produced a more detailed ruler on a smaller-scale as opposed to the dichotomous scoring's basic ruler with its wider-scale.

For the upper secondary school student participants, Figure 2 shows that both dichotomous and polytomous scoring indicated that item number 9 was the most difficult item. However, for the undergraduate participant, Figure 3 shows the most difficult problem was item number 2. For the easiest problem, both undergraduate student and upper secondary school student respondents showed a difference between the two scoring methods. Figure 2 shows that the most difficult item in

dichotomous scoring was item number 14, while the most difficult item in polytomous scoring was item number 10, especially in regard to reaching level 1 (10.1). For undergraduate students, the easiest item in dichotomous scoring was number 4, while in polytomous scoring, the easiest item was number 14, especially 14.3. The above results show that the scoring of responses of the difficult items, often coming from the respondents with high ability, are relatively consistent both dichotomously and polytomously. The consistent responses can be interpreted as the product of an actual thinking process among the respondents, instead of simply guessing. Meanwhile, the responses to the easy problem might come from the respondents with low ability, especially since they were empirically proven to be inconsistent, demonstrating that they might be the result of guessing.

The results of this study indicate that polytomous scoring is able to estimate more the abilities of the participants in more detail, yield a smaller standard deviation than that of dichotomous scoring, and cause the distribution of the estimates of participants' ability closer to a normal distribution. The above results serve as empirical evidence that weighted polytomous scoring is able to estimate students' physics ability more accurately than dichotomous scoring. The results of this study are in keeping with the results of Jiao et al. (2012) who applied computerized adaptive tests (CAT) on a large scale and found that polytomous scoring yielded a slightly more precise estimate of ability than dichotomous scoring. Also, the results are similar to the results of research of Grunert et al. (2013), which show that polytomous scoring on a chemical exam yields a higher percentage average score than dichotomous scoring. The results of this study along with the results of research by Jiao et al. (2012) and Grunert et al. (2013) have become empirical evidence of the statements of Baker et al. (2000), Tognolini & Davidson (2003), Wu (2003), and Bond & Fox (2007) that argue that multiple-category scoring (polytomous) can estimate the test-takers' aptitude better than that of dichotomous scoring.

The new finding of the development of polytomous scoring based on analytical weighting and a penalty system developed in this study as compared to previous polytomous scoring, especially in estimating students' physics ability, is its ability to appreciate respondents' thinking processes on a step-by-step basis when they are selecting response items. Because the appreciation for every step also considers the complexity of each option, then polytomous scoring system will not only be more accurate than that of dichotomous scoring but also becomes more equitable in estimating respondents' abilities and ensures that points are awarded for ability rather than luck (Wiseman, 2012; Holt, 2006). Thus, it can be emphasized that the polytomous-scoring system based in an analytical weighting and penalizing system using multiple-correct items is able to estimate physics ability more accurately than dichotomous scoring. This conclusion is based on the empirical facts that polytomous scoring is able to estimate more detailed capabilities with smaller average standard deviations and approximate distributions closer to a normal distribution than dichotomous scoring.

In summary, the analytical-weighting approach to scoring multiple-correct items in this study was able to produce a more accurate estimation of physics ability than

the dichotomous scoring approach. This is indicated by the findings that the scoring model estimated students' physics abilities in more detailed and with greater accuracy, had an approximate distribution closer to the normal distribution, and produced a standard deviation smaller than that of dichotomous scoring.

Based on the results of this research, it is recommanded that assessments of physics ability that use selected response items, especially multiple-correct items, on a large scale can apply the analytical weighting scoring based on the complexity of content and a penalty system. However, this scoring model might be difficult to apply manually on a large scale because it requires much more time than the standard multiple-choice item. Therefore, further research is needed to develop application software that supports this scoring model.

Acknowledgements

We acknowledge the generous support of the Government of Indonesia through the Directorate of Research and Community Service, which has supported this project with funds through the competitive grants research scheme.

References

Adams, R.J., & Khoo, S.T. (1996). *Quest* (Computer software). *The interactive test analysis system*. Victoria: Acer.

Adeyemo, S. A. (2010). Students' ability level and their competence in problem-solving task in physics. *International Journal of Educational Research and Technology, 1*(2), 35 – 47.

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning*, *16* (1), 1-14.

Baghaei, P., & Dourakhshan, A. (2016). Properties of single-response and double-response multiple-choice grammar items. *International Journal of Language Testing, 6* (1), 33-49.

Baker, J.G., Rounds, J.B., & Zeron, M.A. (2000). A comparison of graded response and rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistic, 25*(3), 253-270.

Bishara, A. J., & Lanzo, L. A. (2015). All of the above: When multiple correct response options enhance the testing effect. *Journal Memory*, 23(7), 1013-1028.

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, *40*(2), 218-231.

Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences* (2nd *ed.*). Mahwah: Lawrence Erlbaum Associates, Publishers.

Donoghue, J. R. (2005). An empirical examination of the IRT information of polythomously scored reading items under the generalized PCM. *Journal of Educational Measurement*, *31*(4), 295-311.

Emaikwu, S. O. (2015). Recent issues in the construction and scoring of multiple-choice items in examinations. *The International Journal of Humanities & Social Studies*. *3*(6), 201-207.

Fenrich, P. (2004). Instructional design tips for virtually teaching practical skills. In *Proceedings of the 2004 Informing Science and IT Education Joint Conference. Rockhampton, Australia June 25–28, 2004.*

Frisbie, D.A. (1992). The multiple true-false format: A status review. *Educational Measurement: Issues and Practice, 11*(4), 21-26

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: development of a rubric for rating partial credit. *Journal of Chemical Education*, *90*(10), 1310-1315.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15* (3), 309-334.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* London: Sage Publications.

Hickson, S., Reed, W. R., & Sander, N. (2012). Estimating the effect on grades of using multiple-choice versus constructive-response questions: Data from the classroom. *Educational Assessment*, *17*(4), 200-213.

Holt, A. (2006). An analysis of negative marking in multiple-choice assessment. *The 19th Annual Conference of the National Advisory Committee on Computing Qualifications* (NACCQ 2006), Wellington, New Zealand. Samuel Mann and Noel Bridgeman (Eds)

Jiao, H., Liu, J., Hainie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement*, *72*(3), 493–509.

King, K. V., Gardner, D. A., Sasha Zucker, S., & Jorgensen, M. A. (2004). The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics. *Assessment Report*, Pearson Education, Inc, July 2004.

Klein, P., Müller, A., & Kuhn, J. (2017). Assessment of representational competence in kinematics. *Physical Review Physics Education Research 13*, 1-18.

Lin, C.J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test form. *The Journal of Technology, Learning, and Assessment. 6*(8), 1-42.

Martin, D. L., & Itter, D. (2014).Valuing assessment in teacher education-multiple-choice competency testing. *Australian Journal of Teacher Education*, 39(7), 1-14.

Merrell, J. D., Cirillo, P. F., Schwartz, P. M., & Jeffrey A. Webb, J. A. (2015). Multiple-choice testing using immediate feedback—assessment technique (IF AT) forms: Second-chance guessing vs. second-chance learning. *Higher Education Studies*. *5*(5), 50-55.

Oosterhof, A. (2003). *Developing and using classroom assessments* (*3th ed*). Upper Saddle River: Merrill Prentice Hall.

Redish, E.F., Scherr, R.E. & Tuminaro, J. (2006). Reverse engineering the solution of a "simple" physics problem: Why learning physics is harder than it looks. *The Physics Teacher*, 44(5), 293-300.

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, Summer, 3-13.

Stankous, N. V. (2016). Constructive response vs. multiple-choice tests in math: American experience and discussion (Review). *European Scientific Journal*. May 2016, 308-3016.

Tognolini, J., & Davidson, M. (Juli 2003). *How do we operationalise what we value? Some technical chalenges in assessing higher order thinking skills.* Paper presented in the Natinaonal Roundtable on Assessment Conference. Darwin, Australia.

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing, *Iranian Journal of Language Testing*, *2* (1), 59-92.

Wooten, M. M., Cool, A. M., Edward E. Prather, E. E., & Tanner, K. D. (2014). Comparison of performance on multiple-choice questions and open-ended questions in an introductory astronomy laboratory. *Physical Review Special Topics - Physics Education Research,* 10, 1-22.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Chicago: Mesa Press.

Wu, B.C. (2003). Scoring multiple true-false items: A comparison of sumed scores and response pattern scores at item and test level. *Research report*, Educational Resources International Center (ERIC) USA, 1-40.