The scoring model of weighting by difficulty and number right to assess professional competency among prospective science teachers based on international benchmarking survey standards

Cite as: AIP Conference Proceedings **2556**, 060002 (2023); https://doi.org/10.1063/5.0110963
Published Online: 17 March 2023

Didik Setyawarno, Dadan Rosana and Eko Widodo

View Online Export Citation

AIP Conference Proceedings **2556**, 060002 (2023); https://doi.org/10.1063/5.0110963 **2556**, 060002

The Scoring Model of Weighting by Difficulty and Number Right to Assess Professional Competency among Prospective Science Teachers Based on International Benchmarking Survey Standards

Didik Setyawarno[a)], Dadan Rosana[b)] and Eko Widodo[c)]

Department of Natural Science Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Jl Colombo No 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia

[a)] Corresponding author: didiksetyawarno@uny.ac.id
[b)] danrosana@uny.ac.id
[c)] eko_widodo@uny.ac.id

Abstract. This study aims at enhancing the quality of science learning in order to meet the standard of international benchmarking surveys. The assessment analysis in this study used Rasch Model based on the responses from the test takers of prospective science teachers towards the assessment results of their professional competency. The research population was the Undergraduate students of Science Education in the Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta. There were 67 students through cluster random sampling from 105 students who joined Biophysics courses. The scoring model to assess the students' professional competency used Weighting by Difficulty (WD) and Number-Right (NR). The results of this study showed that the average WD score was higher than the NR score, but the score was more spread than its average. The data from the normality test both WD and NR score models were normally distributed meaning that the score distribution was close to the median scores. The application of the two scoring model resulted in dissimilar results that influenced the ranking of the test-takers and the scoring consistency between the estimated score using the NR scoring model and the WD model can be categorized as fair agreement. The distribution of the WD scoring model tended to slope to the right compared to the NR model, and the distribution of the NR scoring model tended to widen down compared to the WD model.

INTRODUCTION

Science education provides great potential and strategic role in preparing qualified human resources to face the era of industrial revolution and globalization. This potential should be optimized with science education departments by producing well-rounded graduates that are able to think logically-critically-creatively, to solve problems, to master technology and to be adaptive to the current changes and development. In fact, Teacher Education Institution which is responsible to prepare prospective teachers has not optimally developed scientific literacy assessments and high order thinking skills, especially related to international benchmarking surveys such as the Program for International Student Assessment [1].

Improving the learning quality in science education requires collaboration from various parties, especially prospective science teachers. Law number 14 of 2005-chapter IV article 8 concerning teachers and lecturers mentions that teachers must have academic qualifications, competencies, teachers' certificates as well as physically and mentally healthy to realize the goals of national education. This high demands of teachers, particularly in science, urges the policymakers in Indonesia to give serious attention to international surveys or mapping since it is related to national competitiveness in this global era [2]. Actually, the escalation of Indonesia's PISA achievements in 2015 has raised the optimistic for the future of young generation, though it is still relatively lower than to the OECD average. There were three development of Indonesian competency based on PISA results in 2005 and the biggest was science

The 8th International Conference on Research Implementation and Education of Mathematics and Science (ICRIEMS 2021)
AIP Conf. Proc. 2556, 060002-1–060002-14; https://doi.org/10.1063/5.0110963
Published by AIP Publishing. 978-0-7354-4314-3/$30.00

where it obtained 382 points in 2012 and raised to 403 points in 2015. Meanwhile, based on the median value, the highest aspect was in the science achievements, from 327 points in 2012 to 359 in 2015. The high median compared to the average point can be a good indicator in terms of access distribution and quality equalization. This increase escalated Indonesia's ranking to 6th place compared to the second-lowest position in 2012. The Indonesia's position in 2019 is 41 from 84 countries [3].

PISA model demands a transformation in the assessment system within science education of which test developers (especially prospective science teachers) must be able to measure abilities, learning success, attitudes, interests or other latent traits. Since the ability to think at a higher level is latent trait or unmeasurable (not directly observed), it requires some indirect way to assess the various features among students [4]. One effort that can be done is to provide a number of stimuli, both in the form of tests or questionnaires, though it is difficult to obtain a consistent measurement instrument on one's characteristics. If the stimulus is right on target, the responses to the visible stimulus can reveal the ability, learning success, attitudes, interests, or other measured characteristics. Then, the visible response can then be interpreted in a represented score [5].

In the classical scoring system of the objective test, the score (in this case the raw score) obtained by students refers to the number of Brookhart the correct items, i.e. one point for the correct response and zero for the incorrect ones. This scoring model is called the number right or NR scoring [6]. So, when an item is dichotomized, there will be two score possibilities, 0 or 1. Meanwhile, in modern scoring, NR scoring is the sum of opportunities to answer correctly on the ability of θ to all items answered by students. With this scoring model, all items have equal weighting. Although this scoring model is quite easy to be done, problems may appear if applied on multiple-choice tests or true false model [6]. One of the ways by which such accountability is measured is by the extent to which students' performance in teacher-made tests can predict their potential performance on the standardized tests [7, 8].

A multiple-choice test is a form of objective test that is widely used for various purposes, including in international surveys like the Program for International Student Assessment. In other hand, assessment for learning is a new perspective on the assessment system in education. The traditional practice is for evaluating outcomes is an assessment of learning [9]. The extensively use of this type of test is based on its effectiveness to measure various types of knowledge and complex learning outcomes [10]. It is also appropriate to be used for a test with many participants and the results can also be gathered in a short time [11]. The high implementation of this tes has increased its reliability factor.

Since the item difficulty level is different in the multiple-choice tests, it can be estimated that test-takers who can complete the test item correctly have higher ability or knowledge than those who make incorrect responses [12]. In fact, the items are classified into certain levels of difficulty consisting of easy, medium, and difficult [13]. It makes the items that are considered "difficult" by prospective science teachers can be different to other students. It is very hard to determine the extent to which the test item is difficult before the students complete the test [14].

Based on the results of previous studies, it was revealed that test makers did not pay attention to the difficulty level on each item. As a result, test takers who answer the "difficult" items correctly will have the same score as test-takers who correctly answer the "easy" parts. This certainly does not reflect the fairness principle of test. The various reasons put forward by teachers in maintaining this kind of test, especially the ease of correction process compared to other scoring models since they do not have any incentive for correcting students' answers. It stimulates that the multiple-choice test with NR scoring model become the most economical and hassle-free alternative for competency measurement. This circumstance raises attention to enhance the testing system that really reflects the test-takers' abilities. It can be realized with the Classical Test Theory (CTT) approach and Item Response Theory (IRT). The estimated ability in CTT is based on the participant's visible score, i.e. the average visible score, and the reliability index. Meanwhile, the ability estimation using IRT is based on the opportunity to answer correctly (θ), item characteristics, and applied IRT model [14].

In practice, CTT is more commonly used because its calculations are relatively simpler since the students' ability is measured from the accumulation of the correct answers (NR classic scoring). This score is then processed by prospective science students using the Benchmark Reference Assessment or Normative Reference Assessment. Based on these weaknesses, several improvements were made through IRT with a variety of logical parameters, one of them is the 1-PL model that is developed into the Rasch model that has dissimilar system. The main purpose of Rasch model is to scale the measurement with the same interval because the raw score does not have any. It makes the score cannot be used directly to judge the students' ability [15]. It indicates that the score estimation, both based on CTT and IRT, can be done through various scoring models.

The score estimation according to IRT is based on the opportunity to answer correctly on the ability of θ in each item. It means the estimated score using IRT is the real score that can be obtained by applying scoring models [6]. The scoring model of multiple-choice forms, according to experts, can be divided into several types. There are scoring models that can be classified based on their approach, namely the explicit and the implicit approach [16]. In the implicit approach, there is a raw scores addition (number right scoring, NR) and IRT model (optimum weighting), while the explicit approach contains weighting by difficulty (WD) models, reliability weighting, and validity weighting. However, other experts do not classify the scoring model into specific. Basically, the NR scoring model indeed allows random blind guessing so that the scoring model offers to correct raw scores for the effect of guessing and right minus wrong correction [6, 17]. The reliability coefficient and the standard error of measurement in classical test theory are not properties of a specific test, but are attributed to both a specific test and a specific trait distribution. In latent trait mod els, or item response theory, the test information function (TIF) provides more precise local measures of accuracy in trait estimation than are available from the reliability coefficient.

Along with developments of the measurement field, especially the development of scoring models, guessings are no longer assumed as blind-guessing model. By using a particular scoring model, students' partial knowledge can be identified. The scoring model that can be used to identify partial knowledge among students is confident weighting, answer until correct, and option weighting [6]. In addition to the scoring model above, there are still many scoring models that can be applied in the form of multiple-choice, namely natural weight, multiple regression, equal correlation with the composite, minimum generalized variance, minimum variation, weighting by difficulty, and weighting by validity and the weighted respondent's score [17].

One alternative to estimate the test taker scores has been revealed by Lau et al. (2011) using NRET, i.e. hybrid of NR (Number Right) with ET (Elimination Testing). This model is proven able to detect guessing, partial knowledge, and misconceptions, and to increase test reliability compared to NR. Almost similar to the research of Lau et al., Hoe et al. (2009) conduct a study using the Computer-Adaptive Assessment Software (CAAS) which contains a multiple-choice test form which is also scored using the NRET scoring model [18, 19].

The results showed that 60.3% of students support the use of CAAS and NRET scoring since it is able to identify partial knowledge and misconceptions. Another study from concludes that Confidence Weighting Computerized Adaptive Testing (CWCAT) produces more reliable scores than Dichotomous Computerized Adaptive Testing (DCAT) and is able to measure more precision as evidenced by the low SEM and high-test information function [20]. Therefore, this study tries to accommodate explicit weighting to examine its effect on professional competency among prospective science teachers. Broadly speaking, in Indonesia, explicit weighting has only been applied to non-objective test forms (essays). On the other hand, theoretically, explicit weighting can also be applied to multiple-choice test forms through weighting based on its difficulty level (weighting by difficulty, WD) which is done a priori or logical judgment [16]. Based on the background described above, the researchers are trying to investigate the the professional competency test among prospective science teachers based on the estimating result with the different scoring models, i.e. Item Response Theory IRT.

Based on the problem formulation and the background of the problem, the purpose of this study is to enhance the professional competencey among prospective science teachers to develop international benchmarking survey standards in order to be globally competitive. The strategic objective is among the Science prospective teachers in the Teacher Education Institute. Therefore, this study is trying to answer the following questions:

1. What are the characteristics of the academic question items among the prospective science teachers based on the international benchmarking survey?
2. How is the suitability description between the difficulty level and its empiric difficulty distribution level? and
3. What are the distribution characteristics of the professional competency test among prospective science teachers as the result of WD and NR scoring models? What is the description results from the application of the scoring model toward the professional competency test scores among prospective science teachers?

METHOD

This research can be categorized as quantitative by obtaining the data in the form of numbers or measured qualitative data [21]. In addition, the data obtained were analyzed and compared between the scoring models of weighting by difficulty (WD) and number right (NR) based on PISA. The focus of this quantitative research was identified as a working process that took place in a concise, limited and selective on the problem into measurable parts in form of numbers.

This quantitative research employed the instruments of data collection that produced numerical data or numbers. The data analysis was done with statistical techniques to reduce and classify data, determine relationships and identify differences among groups of data. The controls, instruments, and statistical analysis were used to guarantee the accuracy of the research findings. Thus, the conclusions of the hypothesis test that were obtained through quantitative research can be generally applied.

Sample

The sample is a part of the number and characteristics of the population [21]. The population in this study was the undergraduate students of Science Education, Universitas Negeri Yogyakarta in the class 2016 consisting of 105 students. The cluster sampling technique was used based on groups that gather together naturally. If the sample number is less than 100, so it is better to take them all as the research population, but if the number is larger, it can be taken in the range of 10-15%, 20-55% or more [22]. The sample of this study was the undergraduate students of Natural Sciences A and C classes with a total of 33 and 34 students respectively or 67 out of 105 students. The number of samples was more than 50% of the total which was considered sufficient.

Data Collection

The instrument in this study was the item test on the professional competency adopting from the PISA model. This instrument focused on problem-solving skills at each level of test item as mentioned by European [23] . The test items guidelines are presented in Table 1.

TABLE 1. The guidelines of the research instruments

Stimulus	Item	Test type	International Benchmarking Survey Framework
Funhouse mirror	1	Essay	The ability to modeling problems in complex situations
	2	Essay	The ability to use a simple problem-solving strategy
	3	True-false	The ability to draw conclusions directly
	4	Multiple-choice	The ability to select, compare and evaluate problem-solving strategies
	5	Matching	The ability to work with boundaries and assumptions
Solar stove	6	Essay	The ability to solve clear and direct problems
	7	Essay	The ability to modeling problems in complex situations
	8	True-false	The ability to draw conclusions directly
	9	Multiple-choice	The ability to select, compare and evaluate problem-solving strategies
	10	Matching	The ability to work with boundaries and assumptions
Solar system	11	Essay	The ability to solve clear and direct problems
	12	Essay	The ability to modeling problems in complex situations
	13	True-false	The ability to draw conclusions directly
	14	Multiple-choice	The ability to select, compare and evaluate problem-solving strategies
	15	Matching	The ability to work with boundaries and assumptions
Digestive System in mouth	16	Essay	The ability to solve clear and direct problems
	17	Essay	The ability to modeling problems in complex situations
	18	True-false	The ability to draw conclusions directly
	19	Multiple-choice	The ability to select, compare and evaluate problem-solving strategies
	20	Matching	The ability to work with boundaries and assumptions
Bacterial Growth	21	Essay	The ability to solve clear and direct problems
	22	Essay	The ability to modeling problems in complex situations
	23	True-false	The ability to draw conclusions directly
	24	Multiple-choice	The ability to select, compare and evaluate problem-solving strategies
	25	Matching	The ability to work with boundaries and assumptions

Data Analysis

The data analysis technique was adjusted to the type of research that was quantitatively using the application of QUEST, MINISTEP, and SPSS Version 25. The complete stages of data analysis are as follows:

1. Analysing PISA test items with the QUEST and MINITEP applications to determine the compatibility of items with Rasch model. The compatibility of items with Rasch model was done by using the following references in Table 2 [15].

TABLE 2. Item interpretation

Mean-square (MNSQ) scores	Item Interpretation
> 2,0	Decreasing the quality of the measurement system
1,5 – 2,0	Not good for measurement
0,5 – 1,5	Good conditions for measurement
< 0,5	Less productive for measurement

2. The number of items that were fit the model was re-analyzed to measure the reliability level of the instrument. The instrument reliability level is seen from the value of internal consistency or reliability by referring to the following reference in Table 3 [24].

TABLE 3. Internal consistency or reliability

Reliability Scores	Interpretation
≤ 0,20	Very low
0,21 - 0,40	Low
0,41 - 0,60	Moderate
0,61 - 0,80	High
0,81 - 1,00	Very high

3. The descriptive statistical data analysis used the application of SPSS Version 25. The parameters consisted of mean, std. error of mean, median, mode, std. deviation, variance, skewness, std. error of skewness, kurtosis, std. error of kurtosis, range, minimum, maximum, sum, and percentage for both scoring model of weighting by difficulty (WD) and number right (NR).
4. To analyze the number of test-takers in the subgroup test based on the normal reference assessment, it included high, medium, and low both for weighting by difficulty (WD) and number right (NR) scoring models.
5. The normality test scores data both for weighting by difficulty (WD) and number right (NR) scoring models used the Kolmogorov Smirnov one sample test which was analyzed using the application of SPSS Version 25. The data distribution was normal if the significance value was less than 5%.
6. To visualize the relationship between the test takers' abilities and the opportunity to answer correctly both for weighting by difficulty (WD) and number right (NR) scoring models, Ms. Excell application was used.

RESULT

The items in the weighting by difficulty (WD) and number right (NR) scoring models were made based on the PISA standard. The test item was developed by the research team and it was tested on a field test among undergraduate students of Natural Sciences at Universitas Negeri Yogyakarta. There were 25 questions in this trial that combined Physics and Biology materials. The analysis of test items was done empirically using Rasch Model approach to meet fit criteria based on the following aspects. The field test data were analyzed to determine the quality of each item realted to the conformity with the Rasch model. Based on the analysis results with the QUEST application, there were 25 items that fit Rasch Model amounted to 13 items with an average value as follows in Table 4.

TABLE 4. Test item interpretation

The conformity Model Rasch	Average scores
INFIT Mean Squared (MNSQ)	0,99 ± 0,14
OUTFIT Mean Squared (MNSQ)	1,00 ± 0,17
INFIT T	0,10 ± 0,90
OUTFIT T	0,10 ± 0,70

The quality of item reliability or consistency from the aspect of classical measurements can be seen from the reliability scores. The results of the reliability analysis with the MINISTEP application are as follows in Table 5. Based on the reference level of reliability, the score of the instrument that obtained 0.93 can be categorized as high.

TABLE 5. General analysis results

Statistic	Total Score	Model S.E	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD
Mean	122.1	0.36	0.96	-0.03	0.96	0.01
SEM	22.5	0.11	0.06	0.33	0.06	0.30
P. SD	77.9	0.39	0.21	1.13	0.22	1.03
S. SD	81.1	0.41	0.22	1.18	0.23	1.07
Max	239.0	1.20	1.42	2.33	1.44	2.12
Min	1.0	0.13	0.68	-1.72	0.63	-1.08
Reliability	0.93					

The scoring model of Weighting by Difficulty (WD)

The number of valid items based on the empirical test results from 25 items was 13 items with the maximum total score of 37 if the test taker was able to perfectly answer all items. The respondent group was divided into 3 groups including the high, moderate, and low ability groups with the following details in Table 6.

TABLE 6. Distribution of participants' ability on WD model

Category	Score		Number of Respondents
	Percentage (%)	Real (score)	
High	0,67 – 100	25-37	40
Moderate	0,33 – 0,66	12-24	27
Low	0 – 0,32	0-11	0

The statistical descriptive analysis on the test takers who worked on the PISA model with the WD scoring model was analyzed with the SPSS Version 25 application with the following results in Table 7.

TABLE 7. The statistical descriptive analysis of WD scoring model

Paramaters		WD
N	Valid	67
	Missing	0
Mean		25.6119
Std. Error of Mean		.51480
Median		25.0000
Mode		25.00^a
Std. Deviation		4.21381
Variance		17.756
Skewness		-.372

Paramaters		WD	
Std. Error of Skewness		.293	
Kurtosis		-.055	
Std. Error of Kurtosis		.578	
Range		20.00	
Minimum		14.00	
Maximum		34.00	
Sum			1716.00
Percentiles	25		23.0000
	50		25.0000
	75		29.0000

The distribution of the three groups of test-takers' abilities with the WD scoring model is illustrated in the following graph relationship between the scores and the abilities.

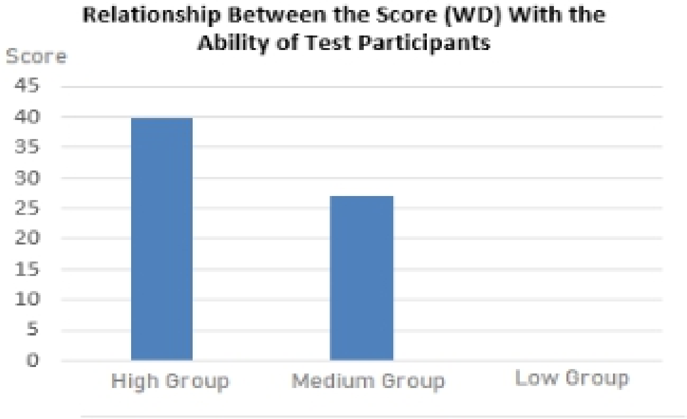

FIGURE 1. Distribution Graph of the abilities group on WD Model

The graph in Fig. 1 shows that the highest was the high group test participants. It was followed by the medium group and the low group (none). To have clear data distribution on the data distribution, all collected data were analyzed with SPSS Version 25 as presented in Fig. 2.

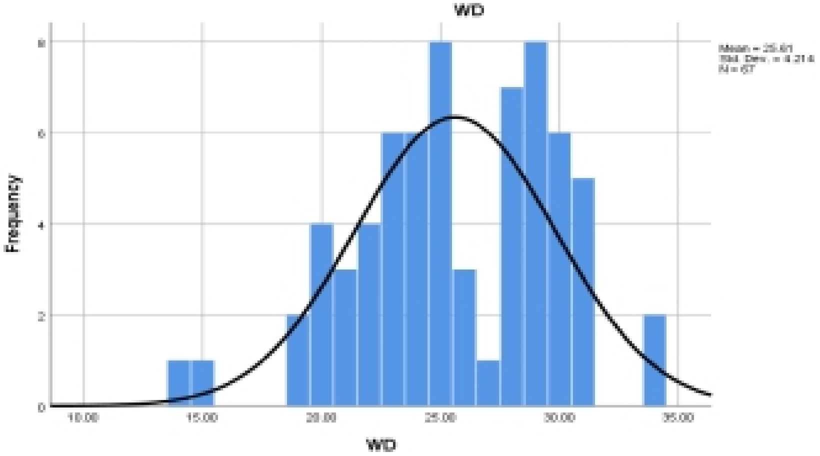

FIGURE 2. Data Distribution of WD Model

The ability of test-takers that was divided into three categories (high, medium, and low) had a direct relationship with the answering opportunities of the whole item. The relationship of the ability among test takers with the opportunity to answer correctly on the WD scoring model is illustrated in the following graph. The graph in Fig. 3 indicates that the higher the ability of test-takers, the greater the chance of answering items correctly.

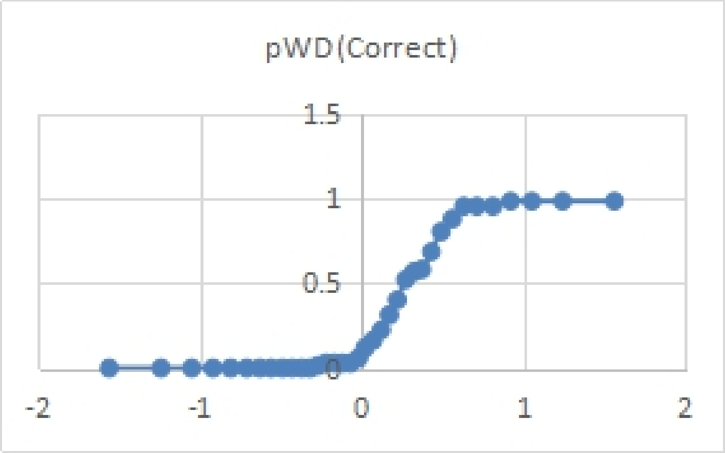

FIGURE 3. The relationship of test participant's ability with the correct answer opportunities in WD model

Scoring model of Number Right (NR)

The number of valid items based on the empirical test results from 25 items was 13 items with the maximum total score of 37 if the test taker was able to perfectly answer all items. The respondent group was divided into 3 groups including the high, moderate, and low ability groups with the following details in Table 8.

TABLE 8. Distribution of participants' ability on the NR model

Category	Score		Number of Respondents
	Percentage (%)	Real (score)	
High	0,67 – 100	9 – 13	32
Moderate	0,33 – 0,66	4 – 8	35
Low	0 – 0,32	0 -3	0

The statistical descriptive analysis on the test takers who worked on the PISA model with the NR scoring model was analyzed with the SPSS Version 25 application with the results in Table 9.

TABLE 9. The statistical descriptive analysis of NR scoring model

Paramaters		NR
N	Valid	67
	Missing	0
Mean		25.6119
Std. Error of Mean		.51480
Median		25.0000
Mode		25.00^a
Std. Deviation		4.21381
Variance		17.756
Skewness		-.372
Std. Error of Skewness		.293
Kurtosis		-.055

Paramaters		NR
Std. Error of Kurtosis		.578
Range		20.00
Minimum		14.00
Maximum		34.00
Sum		1716.00
Percentiles	25	7.0000
	50	8.0000
	75	10.0000

The distribution of the three groups of test-takers' abilities with the NR scoring model is illustrated in the following graph relationship between the scores and the abilities.

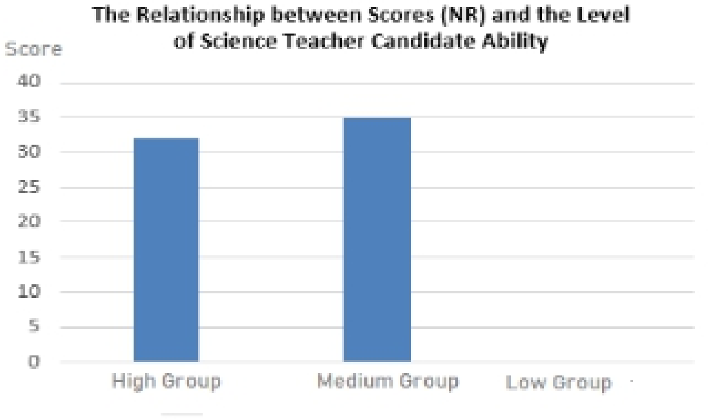

FIGURE 4. Distribution graph of the abilities group on NR Model

The graph in Fig. 4 shows that the highest was the medium group test participants. It was followed by the high group and the low group (none). To have clear data distribution on the data distribution, all collected data were analyzed with SPSS Version 25 as presented in Fig. 5.

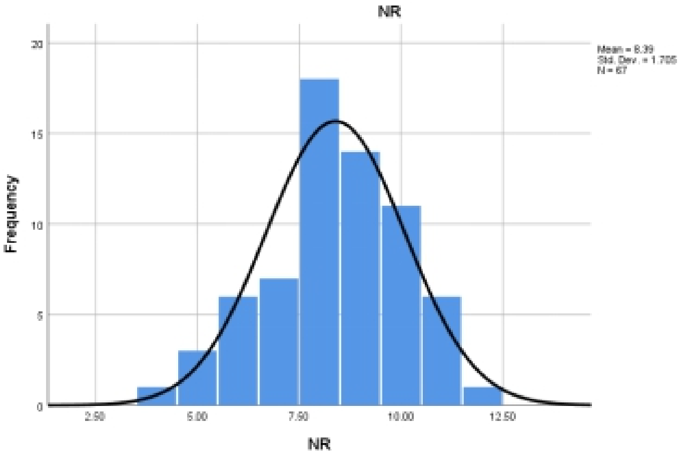

FIGURE 5. Data Distribution of NR Model

The ability of test-takers that was divided into three categories (high, medium, and low) had a direct relationship with the answering opportunities of the whole item. The relationship of the ability among test takers with the opportunity to answer correctly on the NR scoring model is illustrated in the following graph. The graph in Fig. 6 indicates that the higher the ability of test-takers, the greater the chance of answering items correctly.

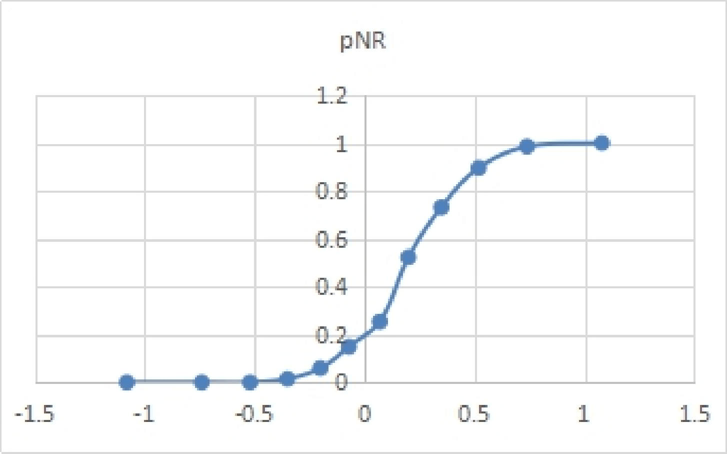

FIGURE 6. The relationship of test participant's ability with the correct answer opportunities in NR model

The comparison of scoring model between Weighting by Difficulty (WD) and Number Right (NR) on the test items with PISA standard

The statistical descriptive comparison of WD and NR Model can be seen in Table 10.

TABLE 10. The statistical descriptive comparison of WD and NR Model

Parameters		NR	WD
N	Valid	67	67
	Missing	0	0
Mean		8.3881	25.6119
Std. Error of Mean		.20832	.51480
Median		8.0000	25.0000
Mode		8.00	25.00a
Std. Deviation		1.70521	4.21381
Variance		2.908	17.756
Skewness		-.313	-.372
Std. Error of Skewness		.293	.293
Kurtosis		-.171	-.055
Std. Error of Kurtosis		.578	.578
Range		8.00	20.00
Minimum		4.00	14.00
Maximum		12.00	34.00
Sum		562.00	1716.00
Percentiles	25	7.0000	23.0000
	50	8.0000	25.0000
	75	10.0000	29.0000

a. Multiple modes exist. The smallest value is shown

The normalization test of score data from both WD and NR models using one-sample Kolmogorov Smirnov test was analyzed using the SPSS Version 25 application with the following results in Table 11.

TABLE 11. The normalization test score

One-Sample Kolmogorov-Smirnov Test

		NR	WD
N		67	67
Normal Parameters[a,b]	Mean	8.3881	25.6119
	Std. Deviation	1.70521	4.21381
Most Extreme Differences	Absolute	.156	.132
	Positive	.112	.080
	Negative	-.156	-.132
Test Statistic		.156	.132
Asymp. Sig. (2-tailed)		.000[c]	.005[c]

The comparison of Estimated Ability towards the opportunity to answer correctly between WD and NR as shown in Fig. 7.

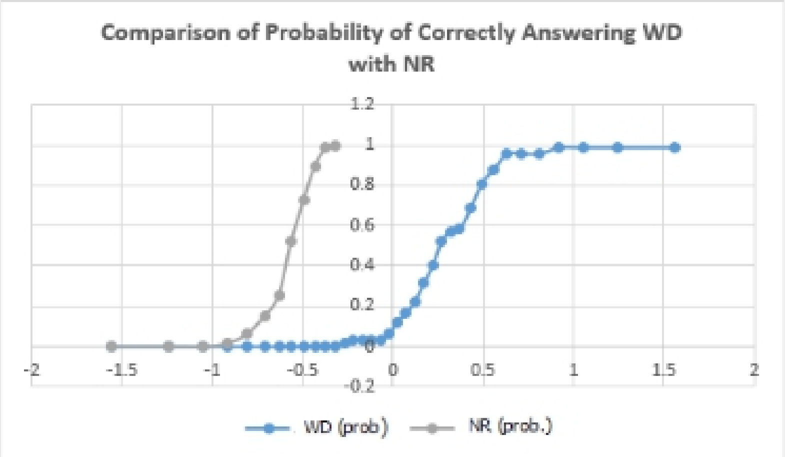

FIGURE 7. The curve of characteristic of NR and WD Test model

DISCUSSION

The ability estimation among test takers on the test item competency with international benchmarking survey standards for prospective science teachers can be done with a classical or modern approach. Based on the data description above, it can be seen that the analysis with the classical approach is simpler because the ability of the test participants is measured based on the accumulation score of the items number which is answered correctly (Number Right). The test score results are processed using the Benchmark Reference Model or Normative Reference Assessment. The estimation of test-taker scores on the international benchmarking survey model items in this study was carried out based on a modern approach (IRT) with Rasch Model. Rasch model was used to find out whether items were fit or not. Based on Table 3, the average results of the Rasch empirical test results obtained INFIT MNSQ = 0.99 ± 0.14, OUTFIT MNSQ = 1.00 ± 0.17, INFIT T = 0.10 ± 0.90, and OUTFIT T = 0.10 ± 0.70. Meanwhile, in the field test, there were 13 out of 25 items in international benchmarking survey model questions that fit with Rasch model. The results of the scores from the 13 items were further analyzed with WD and NR models.

Based on Table 10, it shows that the comparison of NR and WD models with the standard deviation of 8.33881 ± 1.70521 and 25.6119 ± 4.21381 from 67 test participants. Based on these results, it can be stated that the average WD score was higher than the average NR score. When viewed from the standard deviation, the WD score in its average was relatively more spread than the NR score. The scores from WD and NR test participants can be seen in Table 10 with one sample Kolmogorov Smirnov test with a significance value of 0.005 and 0,000 respectively. Based on the results of the two models, it shows a normal distribution of scores.

The results of the skewness analysis (sk) were used to see the slope of the data distribution which referred to the size of the data sharing or the average distribution of data in the form of a bell for the normally distributed data. If the value of skewness (sk) = 0, it means that the data is normally distributed, skewness (sk) <0 is sloping to the right, and skewness (sk)> 0 is sloping to the left. The results of skewness (sk) analysis of NR and WD scoring models were -0.313 and -0.372, so it can be stated that the WD scoring model had a right-angled distribution comparing to the NR scoring model.

The results of kurtosis analysis are interpreted as the frequency-distribution curve. The sharper the kurtosis value, the more homogeneous the data. If the value of kurtosis was 0, it shows normal data. If the kurtosis value is getting smaller, it shows the more spread of the data or heterogeneous data. If the value of kurtosis is close to zero, then the data tends to be normal. When the value of kurtosis is negative, it means the data has a blunt curve or tends to widen down, conversely if the value of kurtosis is positive then the curve is sharp or tends to be clustered (homogeneous). The results of kurtosis analysis of NR and WD scoring models were -0.171 and -0.055, respectively. It means that the NR scoring model has a distribution that tends to widen downward than WD scoring model.

Number right scoring is also called adding raw score [16]. This scoring model is the simplest scoring model because it only sums up the opportunity to answer correctly on each item and considers each item to have the same weight [6, 14]. According to Rudner, the correct number scoring model actually includes a scoring model that accommodates implicit weighting [16]. The application of NR is relatively easy because it only counts the number of opportunities to answer correctly, but it turns out to cause problems when used on multiple choice tests. There are four weaknesses in the application of the NR scoring model. First, from a psychometric point of view, the existence of an element of guessing in the NR scoring model causes a relatively large variation of errors plus this model does not consider partial knowledge possessed by students. Second, based on pragmatic reasons, that exam participants are more careful and ignore items that they cannot answer "punished" compared to exam participants who dare to take risks. Third, based on moral reasons, that the act of guessing is wrong and giving a prize value on the answer to the guessing result is an unwarranted act. Fourth, based on political reasons, that encouraging test takers to guess results in a mental lack of confidence in facing multiple choice test exams. The items should possess a significant loading, indicating a statistically valued contribution; however, an item’s conceptual significance should be examined before an item is removed from the set. Theoretical knowledge is more relevant than a statistical measure [25]. Latent component cognitive processing models such as the LLTM or component latent trait model have sporadically been used to empirically study sources of item difficulty in first and FL reading [26, 27].

In the WD scoring model, the question maker assigns a weight to each item based on his intuition about the level of difficulty of the item he has made or how valuable the item is compared to other items. This method of weighting is also called "a priori" (subjective weighting). Therefore, the scoring of this model belongs to scoring with an explicit approach [16]. Others argued that the influence of risk-taking behavior threatens the measurement of actual mastery of domain knowledge [28]. Ability estimation can be done with a classical approach (Classical Test Theory, CTT) or a modern approach (Item Response Theory, IRT). In practice, CTT is more commonly used because its calculations are relatively simpler because students' abilities are measured from scores which are the accumulation of the number of items that are answered correctly (NR classic scoring). This score is then processed by the teacher using the Benchmark Reference or Normative Reference Assessment. Based on this weakness, improvements were made through IRT with a variety of logistical parameters (PL), one of which was the 1-PL model developed into Rasch model. The approach taken through Rasch model is different. The main purpose of Rasch model is to scale the measurements at the same intervals. This is because the score (raw score) does not have the nature of intolerance, so the score cannot be used directly to provide an interpretation of students' abilities [15]. So, score estimation, both based on CTT and IRT, can be done with various scoring models. The application of the two scoring models to estimate the test takers’ scores on the international benchmarking survey model items has the effect on the obtained estimation results where there are differences in the rank of test participants based on the produced scores.

The last finding found out in the research process that the ability estimation among test takers on the test item competency with international benchmarking survey standards for prospective science teachers can be done with a classical or modern approach. Based on the data description above, it can be seen that the analysis with the classical approach is simpler because the ability of the test participants is measured based on the accumulation score of the items number which is answered correctly (Number Right).

CONCLUSION

Based on the data description and discussion that have been described, it can be concluded that 1) the average WD score was higher than the NR score, but the score was more spread than its average, 2) the data of normality test both the WD and NR score models were normally distributed meaning that the score distribution closed the median scores, 3) The different application of the scoring model resulted in dissimilar results that influenced the ranking of test-takers and 4) the scoring consistency between the estimated score using the NR scoring model and the WD model can be categorized as fair agreement, 5) the WD scoring model has a sloping distribution to the right compared to the NR scoring model, and 6) the NR scoring model shows a distribution that tends to widen downward compared to the WD scoring model.

REFERENCES

1. B. Houtz, *Teaching Science Today* (Corinne Burton, New York, USA, 2010), pp. 56-58.
2. T.M. Simatupang and Rustiadi, *Enhancing the Competitiveness of the Creative Services Sectors in Indonesia* (School of Business and Management, Bandung, Indonesia, 2015), pp. 31-32.
3. IEP, *IEP Manual and Forms* (Academic Office/Bureau of Special Education, Hartford, USA, 2019), pp. 43-47.
4. S.M. Brookhart, *How to Assess Higher-Order Thinking Skills in Your Classroom* (ASCD, Alexandria, Virgina USA, 2010), pp. 88-91.
5. P. Ferrando, U. Lorenzo, and G. Molina, Appl. Psychol. Meas. **25**, 3-6 (2001).
6. L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Reinhart, and Winston, Inc., New York, USA, 2008), pp. 178-180.
7. C.E. Notar, D.C. Zuelke, J.D. Wilson, and B.D. Yunker, J. Instr. Psychol. **31**, 115-117 (2004).
8. K. Kinyua and O.O. Okunya, Afr. Educ. Res. J. **2**, 61-63 (2014).
9. E. Akib Ghafar, M.N.A., Educ. J. **4**, 64-65 (2015).
10. M.D. Miller, R.L. Linn, and N.E. Grondlund, *Measurement and Assessment in Teaching (10th Ed)* (Pearson Education, Inc, New Jersey, 2009), pp. 72-75.
11. S. Supranata, *Panduan Penulisan Tes Tertulis (Penilaian Berbasis Kelas)* (Remaja Rosdakarya, Bandung, Indonesia, 2005), pp. 63-65.
12. T.M. Haladyna, *Developing and Validating Multiple-Choice Test Items (3rd Ed* (Lawrence Erlbaum Associates Publishers, New Jersey, 2004), pp. 57-59.
13. S. Gajjar, R. Sharma, P. Kumar, and M. Rana, Indian J Community Med **39**, 17-20 (2014).
14. F.B. Baker, *The Basics of Item Response Theory (2nd Ed)* (ERIC Clearinghouse on Assessment and Evaluation., USA, 2001), pp. 92-94.
15. B. Sumintono and W. Widhiarso, *Aplikasi Model Rasch Untuk Penelitian Ilmu Sosial* (Trim Komunikata Publishing House, Cimahi, 2015), pp. 83-86.
16. L.M. Rudner, *Informed Test Component Weighting* (Educational Resources Information Center (ERIC), US Department of Education., USA, 2000), pp. 67-69.
17. D.S. Naga, *Teori Sekor Pada Pengukuran Mental (Edisi Kedua)* (Nagarani Citarayasa, Jakarta, Indonesia, 2013), pp. 234-236.
18. L.S. Hoe, L.N. Kiong, H.K. Sam, and H.B. Usop, US-China Educ. Rev. **6**, 51-52 (2009).
19. P.N.K. Lau, S.H. Lau, K.S. Hong, and H. Usop, Educ. Technol. Soc. **14**, 99-101 (2011).
20. Y.-C. Yen, R.-G. Ho, L.-J. Chen, K.-Y. Chou, and C. Yan-Lin, Educ. Technol. Soc. **13**, 163-164 (2010).
21. Sugiyono, *Metode Penelitian Kombinasi* (Alfabeta, Bandung, Indonesia, 2012), pp. 123-127.
22. S. Arikunto, *Prosedur Penelitian Suatu Pendekatan Praktik* (Rineka Cipta, Jakarta, Indonesia, 2010), pp. 98-102.
23. OECD, *PISA 2012 Assessment Framework—Key Competencies in Reading, Mathematics and Science* (OECD, Paris, 2012), pp. 87-93.
24. D. Rosana and D. Setyawarno, *Statistik Terapan Untuk Penelitian Pendidikan* (UNY Press, Yogyakarta, Indonesia, 2016), pp. 75–78.
25. A.S. Beavers, J.W. Lounsbury, J.K. Richards, S.W. Huck, G.J. Skolits, and S.L. Esquivel, Pract. Assess. Res. Eval. **18**, 1 (2013).
26. J.S. Gorin, J. Educ. Meas. **42**, 351-352 (2005).
27. P. Sonnleitner, Psychol. Sci. Q. **50**, 345-346 (2008).

28. S. Fowell and B. Jolly, Med Educ. **34**, 785-786 (2000).