REiD (Research and Evaluation in Education), 3(2), 2017, 152-162
Available online at: http://journal.uny.ac.id/index.php/reid
Research Article

Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum

Dian Normalitasari Purnama
Graduate School of Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
Email: diannsp@gmail.com

Submitted: 23 January 2018 | Revised: 12 February 2018 | Accepted: 12 February 2018

Abstract

This study is aimed at: (1) understanding the characteristics of Accounting Vocational Theory trial test items using the Item Response Theory and (2) determining the horizontal equation of Accounting Vocational Theory trial exam instruments. This was explorative-descriptive research, observing the subject of the eleventh-grade students. The research objects were test instruments and responses of students from six schools selected through the stratified random sampling technique. The data analysis employed review sheets and BILOG program for the Item Response Theory 2PL. The findings were as follows. (1) The test item review of test packages A and B found 37 good quality items, the Item Response Theory using 2PL showed that Package A Test generated 27 good questions, Package B Test contained 24 good questions. (2) The question equating using the Mean/Sigma method resulted in the equation of b_2^*= 1.168bx + 0.270, with the Mean/Mean method resulting in the equation of b_2^* = 0.997bx - 0.250, the Mean/Mean method at 0.250, while Mean/Sigma method at 0.320.

Keywords: *accounting questions, vocational high school, horizontal equating, Item Response Theory*

How to cite item:

Purnama, D. (2018). Characteristics and equation of accounting vocational theory trial test items for vocational high schools by subject-matter teachers' forum. *REiD (Research and Evaluation in Education), 3*(2), 152-162. doi:http://dx.doi.org/10.21831/reid.v3i2.18121

Introduction

Nitko and Brookhart (2011, p. 3) define assessment as a broad term referring to a process for obtaining information used for making decisions about students; curricula, programs, and schools; and educational policy. Assessment and evaluation of learning outcomes are among the efforts made to monitor the students' competency following the learning process. In accordance with Article 57 Paragraph (1) of Law No. 20 of 2003 on National Education System, evaluation is performed in the national education quality control framework to show education provider's accountability to interested parties such as students and educational institutions and programs. The evaluation, for instance, is implemented by the government through National Examination (*Ujian Nasional* or UN).

National Examination is held annually and simultaneously across Indonesia. Regulation of the Minister of Education No. 20 of 2007 on the educational assessment standard explains that National Examination is an activity which measures students' competency in certain science and technology subjects to appraise their achievements in National Education Standards. The outcomes of the National Exam are further used by the government to establish policies pertaining to education. Article 68 of Government Regulation No. 19 of

2005 on National Education Standard mentions that the outcomes of the National Exam are used as a consideration in mapping the quality of educational program and/or unit. The mapping has the purpose to understand the quality of education in each region.

Before National Examination is held, the Provincial and Regency Education Offices hold trial exams (nationally known as 'try-outs')- as a preparation for students in facing the exam. In an interview between the researcher and an accounting teacher at a vocational high school, the teacher said that he chaired the Accounting Subject-Matter Teachers' Forum (*Musyawarah Guru Mata Pelajaran* or MGMP) of Sleman Regency. The interview revealed that the test used in the accounting trial exam for vocational high schools held by the Education Office of Sleman Regency, particularly for the Productive Accounting subject, was prepared by the Accounting Subject-Matter Teachers' Forum. The questions were given in two packages (A and B), with the same exam content outline and materials to avoid cheating during the trial exams.

Both packages for the Accounting Vocational Theory trial exam for vocational high schools in Sleman Regency can be used as a collection of questions with good characteristics. A good test instrument is composed of good items (Retnawati, 2014, p. 62). Therefore, an analysis of test items contained in a test instrument is necessary to help finding out the quality of the instrument. Mardapi (2012, p. 128) suggests that an item analysis can observe the difficulty level, discrimination index, and distractor's effectiveness of test items. The analysis also helps in observing the validity and reliability of a test.

In addition to test item characteristics, the parallelism of both trial test packages is unproven. This means that the difficulty level and discrimination index of both test packages may or may not be the same. This can cause a student's scores to be higher than his ability, and thanks to the easier test package he received. This situation may result in the inaccurate measurement in students' competency achievement. For this reason, although both packages for the Accounting Vocational Theory trial exam prepared by the Accounting Subject-Matter Teachers' Forum are provided with the same exam content outline and materials, the equation between package A and B still becomes a subject of attention.

When the parallelism of the two test packages is proven, an equation process is the next step to be taken. Kolen and Brennan (2014, p. 2) define equation or equating as a statistical process in order to adjust the scores of a test so that they can be used interchangeably. Sukirno (2007) explains that equating can compare the scores earned by students albeit using different test packages. In that way, test participants will not be disadvantaged by easier or harder test packages they receive. There are two approaches that can be used for test equating: Classical Test Theory (CTT) and Item Response Theory (IRT). In CTT, the test to be equated must have the same reliability index. The Item Response Theory, which utilizes the mathematical model, determines that the probability of test participants in giving the right answer to a question depends on the ability they possess and also the characteristics of the question (Hambleton, Swaminathan, & Rogers, 1991, p. 9).

Test equating using IRT is more representative than that using CTT, since IRT has invariance characteristics in its parameter. The ability parameter is invariance with the test parameter and vice versa (Aminah, 2012). The same measurement scale in the scores obtained by students during a trial exam will make education quality monitoring easier. The test outcomes will show the students' competence mastery in facing the National Exam, while serving as a consideration for making decisions for improving the quality of graduates.

Hambleton and Swaminathan (1985, p. 197) explain that horizontal equating is performed between two different versions of a test, and vertical equating is performed on tests across the difficulty levels. Horizontal equating can also be defined as determining the equal score for differences (Crocker & Algina, 2008, p. 456). Horizontal equating is proper when it is used for the security of a test, so that several forms of tests are needed. These forms are not the same, but it is expected that they are similar in their content and difficulty. When the difficulty, reliability, and content of tests are so different from one form to another, few methods of equating can properly work (Cook & Eignor, 1991). Dorans, Moses, and Eignor (2010) mention that in an equivalent group design, two tests are administrated to two equivalent groups chosen randomly from the same population (they are assumed to have equivalent ability). Moghadamzadeh, Salehi, and Khodaie (2011) also explain that the equivalent group design might reduce the effect of exercise and boredom, but it might also cause a bias since they might not have equivalent distribution of ability. To reduce the possibility of bias, the use of a big sample is suggested. In addition, Liao and Livingston (2012) present three approaches that could be considered as alternatives to a common-item equating design. In their paper, the randomly equivalent form approach assembles the test forms of equal difficulty by stratified random sampling of items from the item pool. Previous study which was conducted by Miyatun and Mardapi (2000) also introduces the non-anchor item equating using the equivalent group design.

The above description illustrates the significance of equating both test packages of Accounting Vocational Theory trial exam for vocational high schools prepared by the Accounting Subject-Matter Teachers' Forum of Sleman Regency. The question analysis and test instrument equating will realize objective information and show the actual competency of students in preparing for the National Examination.

Method

This descriptive-quantitative research tries to equate the test instruments of Accounting Vocational Theory trial exam for vocational high schools that were prepared by the Accounting Subject-Matter Teachers' Forum of Sleman Regency in the academic year of 2015/2016 in two packages, A and B. The research was conducted at vocational high schools in Sleman Regency, Yogyakarta Special Region. The subjects of this research are grade XII students of vocational high schools in Sleman Regency who took the Accounting Vocational Theory trial exam in the academic year of 2015/2016. The research objects were test instruments and also 650 students' Package B participants in the form of answer sheets from six vocational high schools selected through the stratified random sampling technique based on the National Exam rank for Accounting Vocational Theory subject in the academic year of 2014/2015.

Kolen and Brennan (2014, p. 13) state that there are two ways to do an equivalent group design: (1) by giving single test to measure students' ability, and (2) by doing a structure test administration, for example, X test for the first student, Y test for the second student, X test for the third student and so on. In reference to the theory, the accounting competency try out test is considered to be suitable with the equivalent group design since the students with odd number of students identity were working with Package A test, while those with even number working with Package B test.

The data were collected through documentation. They were reviewed by experts to see the characteristics of the test items qualitatively. The review of the test items was made to material, construction, and language to see their qualitative characteristics. The trial exam answer sheets or responses were used for the quantitative analysis. The test instruments were analyzed using the Item Response Theory with the assistance from the BILOG-MG program to generate three-phase output. In the first phase, it revealed the number of test participants correctly answering test items, ratio of correct answer probability divided by wrong answer probability, and biserial coefficient. The second phase obtained the data on item parameter according to the Item Response Theory model used. The 1-PL model covers the data on the difficulty level, the 2-PL model covers information on the difficulty leve, and discrimination index, and the 3-PL model covers the difficulty level, discrimination index, and guessing factor. In estimating the parameter, the logistic model with the highest number of fit items was used. Fit items are items with calculated Chi-Square value smaller than table Chi-Square value or *p-value* above 5%. The goodness of fit test aims at knowing whether or not the items used are in accordance with the model applied.

The level of difficulties is an item category, easy or uneasy item to students. It can be understood by calculating the number of students who answer correctly. It is considered good when the scores range from -2 to 2, the discrimination index is considered good when the scores range from 0 to 2, and guessing factor is considered good when the score is lower than 0.2 (1/total answer alternatives).

The testing of the equation of the two test packages is aimed at observing whether or not Packages A and B tests were parallel. In the presence of any evidence of non-parallelism, both packages need to be equated. Allen and Yen (1979, p. 59) suggest that two test instruments are considered parallel when both have the same mean and variance. The parallelism testing of the test instruments was carried out using the SPSS Program.

Equating was carried out based on the result of parameter estimation from BILOG which generates information on the equated test instrument conversion constant. Equating was held using equivalent group design since, as shown by the data, the students' responses were sourced from two different test instruments and answered by two different student groups with equivalent ability. There were no anchor items in both test instruments.

Findings and Discussion

Findings

Validity and Reliability of Questions

This research involved five raters to estimate the validity using Aiken formula. The validity of the test items in both Packages A and B according to Aiken formula is relatively good. Package A contains 26 questions with good validity index (minimum 0.87) and also 14 questions with poor content validity. Package B contains 27 questions with good content validity index. There are 13 questions with very poor content validity index.

Characteristics of Accounting Vocational Theory Trial Test items based on Question Item Review Criteria

Table 1 shows the characteristics of trial test items based on the outcome of expert review. In the material aspect, test Packages A and B have 37 good questions and three poor questions. This is due to the reason that the prepared questions are not in accordance with the exam content outline. In the material and language aspects, 40 items in both Packages A and B are in a good category.

Characteristics of Accounting Vocational Theory Trial Test items based on Item Response Theory

The quantitative analysis using Item Response Theory requires an assumption test as a prerequisite. A unidimensional assumption test was carried out to observe whether or not the Accounting Vocational Theory trial exam instruments measure one's ability (trait). The unidimensional test was performed with the factor analysis using SPSS 20. As presented in Figure 1, the result of the factor analysis shows that 40 test items form 11 factors that explain 55.063% of the total variance. The result also shows that the first factor is dominating, with Eigen value of 9.439 which is five times bigger than the second factor. Therefore, it is safe to say that Package A of the Accounting Vocational Theory trial exam instrument is unidimensional.

Table 1.Outcome of trial test items review

Aspect	Package	Question Criteria					
		Good		Poor		Very Poor	
		Qty	%	Qty	%	Qty	%
Material	A	37	92.5	3	7.5	-	-
	B	37	92.5	3	7.5	-	-
Construction	A	40	100	-	-	-	-
	B	40	100	-	-	-	-
Language	A	40	100	-	-	-	-
	B	40	100	-	-	-	-

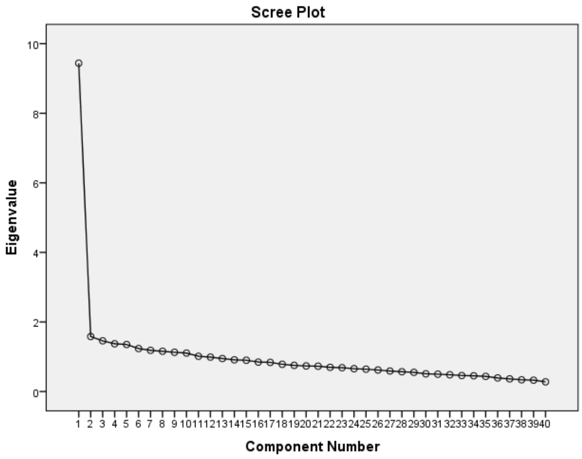

Figure 1. Scree plot of Package A

As presented in Figure 2, in Package B Test, 40 test items form 13 factors which explain 56.740% of the total variance. The result also shows that the first factor is dominating, with the Eigen value of 7.595 which is four times larger than the second factor. Therefore, it can be assumed that Package B of the Accounting Vocational Theory trial exam instrument is unidimensional.

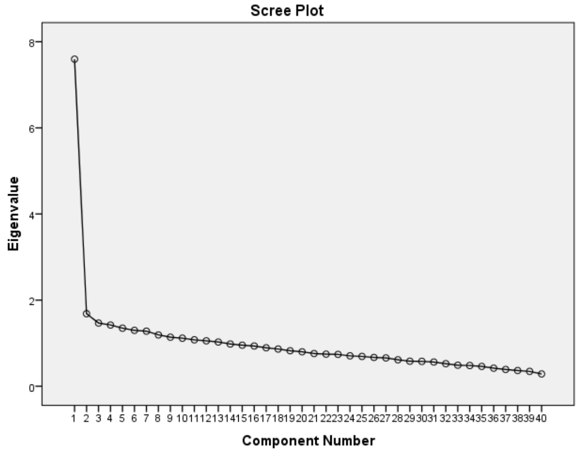

Figure 2. Scree plot of Package B

Local independence assumption test for Package A is proven with variance-covariance matrix and students' ability in doing Package A test, where the students were divided into 15 groups. The classification was carried out by listing the students' rank from the highest to lowest ability. The classification was held using the 2-parameter ability estimation model. The result shows that the elemental value is outside the diagonal approaches, meaning that the test instruments have passed the local independence assumption test.

The parameter invariance assumption test came in two types. The first was question item parameter invariance test which is aimed to observe whether or not the test questions changed when answered by different student groups. The second was parameter invariance test on participants' abilities to see whether or not the estimated students' abilities changed when the test items were changed. The test was performed using scree plots as presented in Figure 3, 4, and 5.

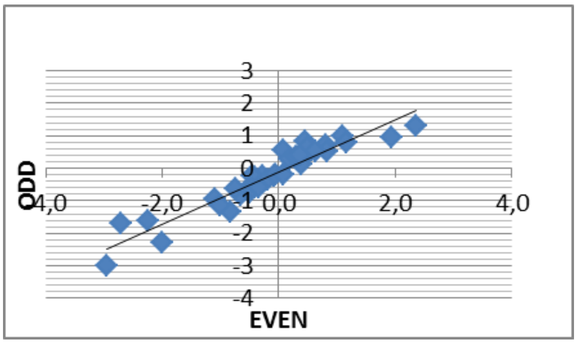

Figure 3. Scree plot of parameter invariance for the difficulty level in Package A test

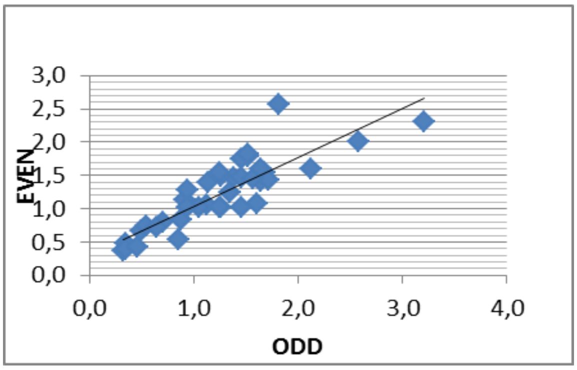

Figure 4. Scree plot of parameter invariance for discrimination index in Package A test

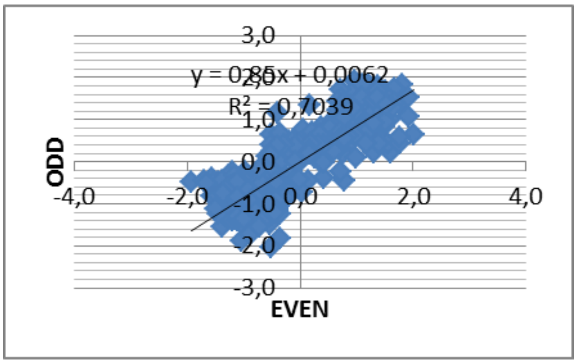

Figure 5. Scree plot of parameter invariance for participants' abilities in Package A test

Figure 3, 4, and 5 show that in general, all of the plots are relatively close to the diagonal line, which can be read that the parameter invariance in Package A Test is met.

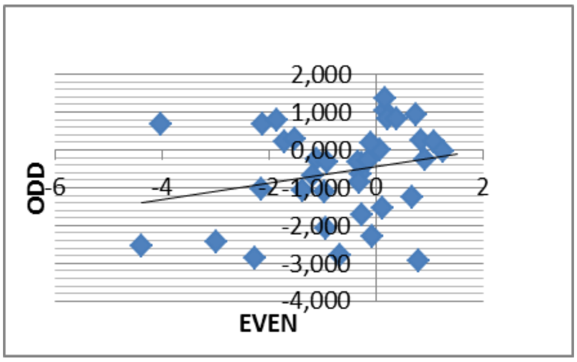

Figure 6. Scree plot of parameter invariance for the difficulty level in Package B test

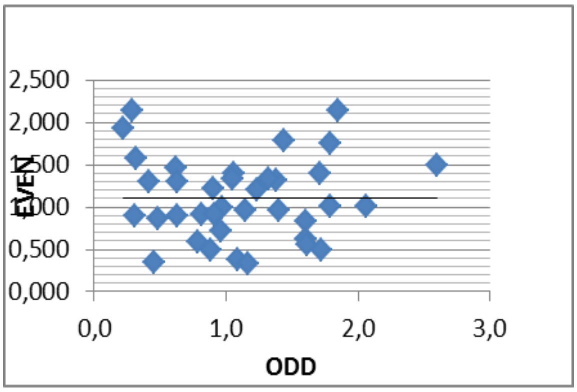

Figure 7.Scree plot of parameter invariance for discrimination index in Package B test

Meanwhile, figure 6 and 7 show that in general, all plots are scattered, away from the diagonal line. Scattered plots away from diagonal line show that the invariance parameter of the difficulty level and the discrimination index of Package B test are not met.

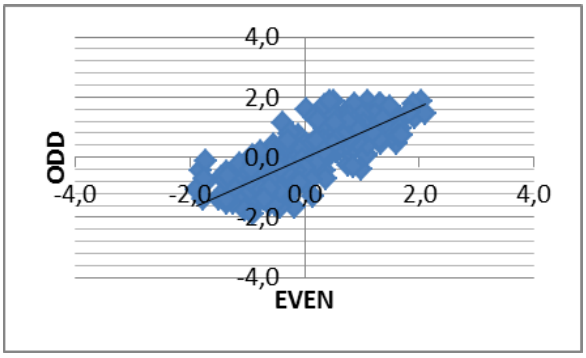

Figure 8. Scree plot of parameter invariance for participants' abilities in Package B test

Figure 8 shows that, in general, all plots are relatively close to the diagonal line. Therefore, it can be inferred that the assumption for invariance parameter for students' abilities in Package B Test is met.

The Result of Model Fitness. In order to determine the model that is fit to the items, data analysis under the three parameter logistics was conducted (1PL, 2PL, and 3PL). The fit-model analysis was assisted by BILOG software version 3.0. The fit-items were the items with Chi-Square value bigger than 5%. The fit-model analysis was beneficial to the determination of the model fitness test to this modern approach by using BILOG version 3.0 program.

Table 2. Goodness of fit test of model by *p-value*

Category	Model		
	1PL	2PL	3PL
Fit	15	32	31
Unfit	25	8	9

Table 2 shows that the item analysis based on the Item Response Theory fits the 2PL model. The result of question analysis based on 2PL model in Package A Test found 27 good questions and 13 poor questions. Such poor questions were caused by the difficulty level and discrimination index that exceeded the criteria.

Table 3. Goodness of fit test of model by *p-value*

Category	Model		
	1PL	2PL	3PL
Fit	11	30	28
Unfit	29	10	12

Table 3 shows that the item analysis based on IRT fits the 2PL model. The result of the item analysis based on 2PL model in Package B Test found 24 good questions and 16 poor questions.

Information Function (IF). The item information function helps determining the quality of a test instrument. To observe the information function of Package A and B tests, 2PL model was used. In the 2PL model, the highest plot information function will be reached when a student who responds to an item has an ability that is equivalent to the difficulty level and discrimination index of the item.

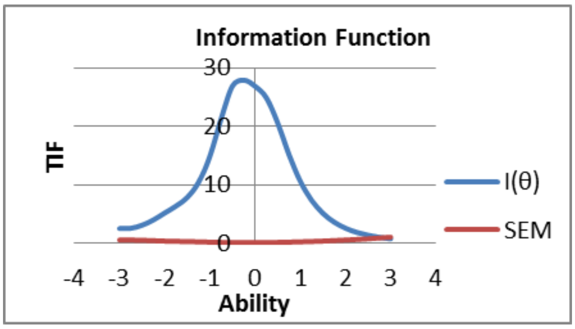

Figure 9. Chart of function information of Package A

Figure 9 shows that the maximum information function value is 27.884 with -0.250 logit (theta). The Estimated Standard Error of Measurement for Package A is 0.189 or inversely proportional with the information function of the test. This means that the participants of Accounting Vocational Theory trial Package A Test will give good information with the smallest measurement error if answered by the participants with -0.250 ability.

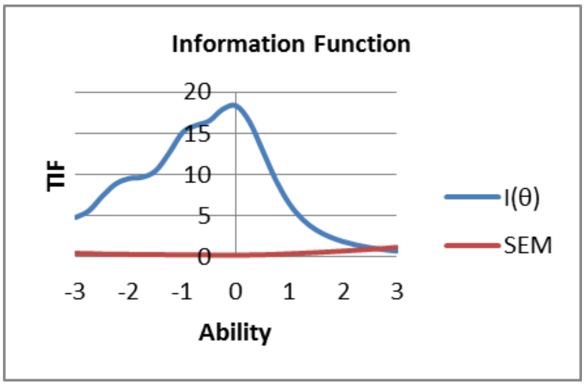

Figure 10. Chart of function information of Package B

Figure 10 indicates that the maximum information function value at 18.362 is reached with 0 logit (theta). The test's SEM is 0.2337 or inversely proportional with the test function. This means that the participants of Accounting Vocational Theory trial Package B Test will give good information with the smallest measurement error if the test was done by the participants with zero (0) ability.

Accounting Vocational Theory Trial Exam Equating Test. Verification of the equation of the Accounting Vocational Theory trial test of both Package A and B must be held in order to see whether or not both packages are parallel. The test for the test instruments' equation can be done using the t-test. The result of the t-test shows the significance value at equal variances assumed at 0.000 < alpha 0.05. This means that the average score in Package A and B differs (with the average difference of 3.092), and therefore, equating is necessary.

Equating

When the Accounting Vocational Theory trial exam instruments were proven unparallel, equating was necessary. During equating test, one needs to determine which package will be used as the benchmark. This research equated Package A to Package B, as presented in Table 4. Based on the result of analysis using BILOG 3.0, it is found that the items with good characteristics and the mostly fit are in the 2PL model.

Mean/Sigma Method. In the mean/sigma method, the calculation of α and β constants using the mean and standard deviation of the difficulty level resulted in constants α = 1.168 and β = 0.270. From the constants α and β, it is found the equation of Package A (*x*) to Package B (*y*) as follows:

$$\theta_2^* = 1.168\theta x + 0.270$$
$$b_2^* = 1.168bx + 0.270$$
$$a_2^* = \frac{a_1}{1.168}$$

Using α and β, item parameter transformation was carried out, which resulted in the equating item parameter as presented in Table 5. The Package A Test shows that there are 17 test items whose average difficulty level is -0.113 and standard deviation 0.641, and after equation, the mean changes to 0.138 and standard deviation changes to 0.749. Further, the average discrimination index of Package A test is 1.285 with the standard deviation of 0.386, and after equation the mean changes to 1.100, and the standard deviation changes to 0.330.

Table 4. Summary of question parameter

No	Package A		Package B	
	The difficulty level	Discrimination index	The difficulty level	Discrimination index
5	-0.320	1.364	0.843	1.084
8	-0.608	1.444	-0.677	1.719
9	-0.136	1.842	-0.282	1.793
12	-0.369	1.634	-0.949	1.606
13	-0.484	1.548	-0.307	1.709
16	-0.262	1.197	-0.154	1.167
17	-0.743	1.508	0.927	0.628
20	0.297	1.199	0.066	1.165
21	1.511	0.573	1.529	0.630
23	-0.103	1.552	0.169	1.787
24	-1.116	0.606	-0.270	1.848
27	-0.035	1.591	0.229	1.619
30	0.624	0.981	0.682	1.049
33	0.602	1.272	0.738	0.910
34	0.427	1.601	0.791	1.402
36	-0.468	1.317	0.391	0.887
37	-0.730	0.611	-1.375	0.905
μ	**-0.113**	**1.285**	**0.138**	**1.289**
σ	**0.641**	**0.386**	**0.749**	**0.422**

Table 5. Conversion of Package A to Package B using mean/sigma method

No	Package A		Package B	
	b Initial	aInitial	(b_2^*)	(a_2^*)
5	-0.320	1.364	-0.104	1.168
8	-0.608	1.444	-0.440	1.236
9	-0.136	1.842	0.111	1.577
12	-0.369	1.634	-0.161	1.399
13	-0.484	1.548	-0.295	1.325
16	-0.262	1.197	-0.036	1.025
17	-0.743	1.508	-0.598	1.291
20	0.297	1.199	0.617	1.026
21	1.511	0.573	2.035	0.491
23	-0.103	1.552	0.150	1.329
24	-1.116	0.606	-1.033	0.519
27	-0.035	1.591	0.229	1.362
30	0.624	0.981	0.999	0.840
33	0.602	1.272	0.973	1.089
34	0.427	1.601	0.769	1.371
36	-0.468	1.317	-0.277	1.128
37	-0.730	0.611	-0.583	0.523
μ	**-0.113**	**1.285**	**0.138**	**1.100**
Σ	**0.641**	**0.386**	**0.749**	**0.330**

Mean/Mean Method. In mean/mean method, the calculation of constants α and β uses the mean of difficulty level and discrimination index, which resulted in constants α = 0.997 and β = 0.250. From the constants α and β, it is found that the equation of Package A (*x*) to Package B (*y*) is as follows:

$$\theta_2^* = 0.997\theta x - 0.250$$

$$b_2^* = 0.997bx - 0.250$$

$$a_2^* = \frac{a_1}{0.997}$$

Table 6 shows the conversion of the result of equation to the difficulty level and discrimination index parameters. Package A test shows that there are 17 test items whose average difficulty level is -0.112 and standard deviation is 0.641, and after equation, the mean changes to 0.138 and standard deviation changes to 0.639. The parameter of discrimination index of Package A is 1.285 with the standard deviation of 0.385, and after equation, the mean changes to 1.289 and standard deviation changes to 0.387.

Table 6. Conversion of Package A to Package B using mean/mean method

No	Package A		Package B	
	b Initial	aInitial	(b_2^*)	(a_2^*)
5	-0.320	1.364	-0.069	1.368
8	-0.608	1.444	-0.356	1.448
9	-0.136	1.842	0.114	1.847
12	-0.369	1.634	-0.118	1.639
13	-0.484	1.548	-0.232	1.553
16	-0.262	1.197	-0.011	1.201
17	-0.743	1.508	-0.491	1.512
20	0.297	1.199	0.546	1.203
21	1.511	0.573	1.756	0.575
23	-0.103	1.552	0.147	1.557
24	-1.116	0.606	-0.863	0.608
27	-0.035	1.591	0.215	1.596
30	0.624	0.981	0.872	0.984
33	0.602	1.272	0.850	1.276
34	0.427	1.601	0.676	1.606
36	-0.468	1.317	-0.217	1.321
37	-0.730	0.611	-0.478	0.613
μ	**-0.112**	**1.285**	**0.138**	**1.289**
σ	**0.641**	**0.385**	**0.639**	**0.387**

Accuracy of Equating Result Based on Root Mean Square Difference. Kim and Cohen (1996, p. 17) explain the formula to calculate the equating accuracy as follows.

$$\text{RMSD } (a) = \sqrt{\frac{\sum_{i=1}^{N}(a_2^* - a_1)^2}{N}}$$

$$\text{RMSD } (b) = \sqrt{\frac{\sum_{i=1}^{N}(b_2^* - b_1)^2}{N}}$$

$$\text{RMSD } (\theta) = \sqrt{\frac{\sum_{i=1}^{N}(\theta_2^* - \theta_1)^2}{N}}$$

Note:
RMSD = Root Mean Square Difference
a_2^* = Differentiator power of the first test after being equated to the second test
a_1 = Differentiator power of the first test
b_2^* = The difficulty level of the first test after being equated to the second test
b_1 = The difficulty level of of the first test
θ_2^* = the ability of the test participants of the first test after being equated to the second test
θ_1 = the ability of the test participants of the first test

Table 7. Summary of RMSD calculation result for mean/sigma and mean/mean methods

Parameter	RMSD	
	Mean/Sigma Method	Mean/Mean Method
The difficulty level (b)	0.272	0.251
Discrimination index (a)	0.192	0.004
Ability (θ)	0.320	0.250

Table 7 shows that the RMSD value in the mean/mean method is lower than that of the RMSD value in mean/sigma method. It can be assumed that equation with the mean/mean method is more accurate compared to that with the mean/sigma method.

Discussion

Characteristics of Trial Exam Question Item Based on Question Item Review

Both Package A and B tests in the material aspect have 3 test items that require revision as they do not fit the exam content outline. For the construction and language aspects, both Package A and B are 100% in good criteria.

Characteristics of Test items

The result of the analysis of Package A test shows that 15 test items fit the 1PL model, 32 test items fit the 2PL model, while 31 test items fit the 3PL model. The characteristics of questions in Package A based on the 2PL model show that there are 27 good questions that fit the model. Thirteen items are poor as their difficulty level and discrimination index do not meet the criteria (above +2).

The result of the analysis of Package B Test shows that 11 test items fit the 1PL model, 30 test items fit the 2PL model, and 28 test items fit the 3PL model. This shows that the 2PL model has the largest number of fit test items. If seen based on the 2PL model, 24 items are good and fit the model, while 16 items are poor.

Trial Exam Question Equating

The questions used in the trial exam of the Accounting Vocational Theory in Sleman Regency were given in Packages A and B. If both packages were used unequally, one of the student groups would be disadvantaged, particularly for students working on harder test packages. The result of the *t*-test on the scores in the two packages shows that both packages are non-parallel, and therefore, equating is necessary. The result of the question equating using the Mean/Sigma method resulted in the equation b_2^*= 1.168*bx* + 0.60, while the Mean/Mean method resulted in the equation b_2^* = 0.997*bx*– 0.250.

Kilmen and Demirtasli (2012) conduct similar equating research by using four methods in the IRT approach. Those four methods use the least RMSD value to determine the accuracy. The RMSD value in the mean/mean method is smaller than the RMSD value in the mean/sigma method. The mean/mean method resulted in the RMSD for parameter b at 0.251, parameter a at 0.004, and ability parameter at 0.250 whereas the mean/sigma method resulted in the RMSD for parameter b at 0.272, parameter a at 0.192, and ability parameter at 0.320. The lower RMSD value shows more accurate equating result, in this case, it is shown that the mean/mean equating method shows better result than the mean/sigma method.

Conclusion

The results of expert review of the test items are as follows. (1) In terms of the material, construction, and language aspects, the test items in the test instruments of Accounting Vocational Theory trial exam prepared by Accounting Subject-Matter Teachers' Forum of Sleman Regency are in a good category. (2) The content validity of Package A and B test items according Aiken formula is satisfactory. (3) The reliability coefficient of test instruments of Accounting Vocational Theory trial exam for both Package A and B is in a good category, at 0.887 for Package A and 0.856 for Package B. (4) The analysis based on the Item Response Theory using the 2PL model to Package A test shows that 32 items fit the model, whereas 30 items fit the model of Package B. The discrimination index of Package A shows that there are 27 good items and 13 poor items. In Package B test, 24 items are in a good category while the remaining 16 items are in a poor category. Poor items are resulted from the difficulty level and discrimination index which exceed the criteria. (5) Equation using the Mean/Mean method shows smaller result compared to the RMSD value found using the Mean/Sigma method.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Montery, CA: Cole Publishing.

Aminah, N. S. (2012). Karakteristik metode penyetaraan skor tes untuk data dikotomos. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *16*(Special Issue for UNY's 48th Dies-Natalis), 88–101. https://doi.org/10.21831/pep.v16i0.1107

Cook, L. L., & Eignor, D. R. (1991). An NCMF instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, *10*, 37–45.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating*. Princeton, NJ: Educational Testing Service.

Government Regulation No. 19 Year 2005, on National Education Standard (2005). Republic of Indonesia.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

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

Kilmen, S., & Demirtasli, N. (2012). Comparison of test equating methods based on Item Response Theory according to the sample size and ability distribution. *Procedia - Social and Behavioral Sciences*, *46*, 130–134. https://doi.org/10.1016/J.SBSPRO.2012.05.081

Kim, S.-H., & Cohen, A. S. (1996). A comparison of linking and concurrent calibration under Item Response Theory. In *American Educational Research Association Annual Meeting* (pp. 1–52). New York, NY: American Educational Research Association.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Law No. 20 of 2003 of Republic of Indonesia on National Education System (2003).

Liao, C.-W., & Livingston, S. A. (2012). A search for alternatives to common-item equating. In *paper presented at the annual meeting of the National Council on Measurement in Education*. Vancouver, British Columbia, Canada.

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

Miyatun, E., & Mardapi, D. (2000). Komparasi metode penyetaraan tes menurut teori respons butir. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *2*(3), 1–18. https://doi.org/10.21831/pep.v2i3.2083

Moghadamzadeh, A., Salehi, K., & Khodaie, E. (2011). A comparison method of equating classic and Item Response Theory (IRT): A case of Iranian study in the university entrance exam. *Procedia - Social and Behavioral Sciences*, *29*, 1368–1372. https://doi.org/10.1016/j.sbspro.2011.11.375

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston, MA: Pearson Education.

Regulation of the Minister of Education No. 20 of 2007 on the educational assessment standard (2007). Republic of Indonesia.

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.

Sukirno, S. (2007). Penyetaraan Tes UAN: Mengapa dan Bagaimana? *Cakrawala Pendidikan*, *26*(3), 305–321. https://doi.org/10.21831/cp.v3i3.3983