

ANALYSIS OF THE QUALITY OF TEST INSTRUMENT AND STUDENTS' ACCOUNTING LEARNING COMPETENCIES AT VOCATIONAL SCHOOL

Nur Ichsanuddin Achmad Kurniawan
Universitas Negeri Yogyakarta

Sudji Munadi
Universitas Negeri Yogyakarta

Abstract


The study is aimed at describing: (1) characteristics of the items of the national examination try-out test of the accounting subject matter in the 2015/2016 academic year on classical test theory and modern test theory; and (2) classification of students' masteries in the learning of accounting. The study is explorative research. Analyses are conducted using the classical and modern test theories for item characteristics and descriptive quantitative for students' masteries in accounting using the test set for the national examination try-out in the 2015/2016 academic year. A total of 414 students do the Package A test. Results show that (1) based on the classical test analyses, a number of 11 items (27.5%) belong to the "easy" category, 22 items (55%) "medium" category, and 7 items (17.5%) "difficult" category allowing a total of 19 (47.5%) to be categorized as good items; meanwhile, on the modern-theory analyses, a total of 34 items (85%) belong to the "good" category. (2) Around 38% of the students have competencies of the medium and low categories. Most students have difficulty in answering questions of the higher-order thinking levels.

Keywords: *test item characteristics, accounting, learning competencies, Rasch Model*

Permalink/DOI: <http://dx.doi.org/10.21831/pep.v23i1.22484>

Contact *Nur Ichsanuddin Achmad Kurniawan*

 *nur.ichsanuddin@gmail.com*

 *Department of Educational Research and Evaluation, Graduate School of Universitas Negeri Yogyakarta*

Jl. Colombo No. 1, Depok, Sleman, 55281, Yogyakarta, Indonesia

Introduction

Education takes an important role in the development of human resources of a country and nation. In the Law of Republic of Indonesia No. 20 of 2003, it is mentioned that education is a conscious and planned effort to provide learning conditions and processes so that learners actively develop their potentials to acquire spiritual, religious strengths, personalities, intellects, decent traits, and skills needed by themselves, the society, the nation, and the country. National education is to function in developing the learners' awareness of their potentials for the sake of the good of the nation in the frame of educating the life of the nation.

In the frame of educating the nation, the government issued the Regulation of the Minister of National Education No. 19 of 2005 about the national standard of education (NSE). The NSE is a minimal criterion of education systems in the legal area under the state of the Republic of Indonesia. The NSE becomes the basis for the planning, implementation, and control of education to realize qualified national education. The NSE also functions to ensure the quality of national education in intellectualizing the nation and building a civilization of the nation. With NSE, it is expected that the quality of education improves.

The NSE consists of eight standards that must be achieved by all education units. These are graduate's competency, content competency, process competency, teacher's competency, facility, management, funding, and evaluation. These standards must be obeyed by teachers and school personnel in running educational programs to develop students' competencies and forming the characters and civilization of the nation. The graduate's competency standard (GCS) becomes the main reference in developing the other standards. This way, evaluation of the instructional processes must be oriented to the GCS.

Outcomes of instructional processes can be seen from the results of the students' scores in examinations. Learning outcomes are interpreted through standardized evalua-

tion processes. Many education systems still use the results of exams as an indicator of students' progress and mastery of knowledge. As a consequence, society tends to look at students' achievements, mainly from final scores of the instructional activities. This view has caused students to have a burden to acquire the highest possible scores (Manoppo & Mardapi, 2014). The magnitude of students' learning outcomes obtained through evaluation processes is then regarded as a judgment for the instructional processes. Such evaluation processes cannot be separated from the assessment processes that are done using particular measuring instruments.

Evaluation is an important component in the running of education programs. Education evaluation is the quantification of phenomena or objects involved in the education process. It is expected that, through a good evaluation system, teachers can devise appropriate learning strategies, and that will motivate students to learn better. Evaluation is a tool that can be used to obtain information on the students' learning achievement.

In the Regulation of the Minister of National Education No. 19 of 2005 Chapter 63 Item (1), it is mentioned that evaluation of learning outcomes at the primary and secondary school levels consists of (a) evaluation of learning outcomes by teachers, (b) evaluation of learning outcomes by the school, and (c) evaluation of learning outcomes by the Government. In line with the advancement in the world of education, the evaluation system that is presently used is the criterion-referenced evaluation. Criterion-referenced evaluation is aimed at knowing a person's competencies on a certain criterion (Mardapi, 2012, p. 186). The criterion-referenced evaluation compares examinee's test scores with an absolute criterion determined by the teacher. So far, results of the criterion-referenced examination are pass or fail. An examinee is regarded as passing if his score is the same with or higher than the given minimal limit and failing if it is lower.

The minimal limit, more familiarly referred to as the minimal passing criterion

(MPC), is the minimum level of competency that a student has in order to be able to be declared as passing a particular education level. MPC is used to know the level of competency a student achieves. The passing label means that a student has achieved the required level of competency and failing means that a student has not.

In Sleman regency, Yogyakarta Special Region, the Business and Management Study Program of the vocational schools have had good graduates. It is shown by the fact that, between 2013 and 2015, the passing percentage of the graduates is 100%. However, the criterion is mainly passing, without information of the extent to which the graduates have the competencies of the subject matters. During the pre-survey with teachers, members of the subject matter professional group, it is found that no empirical review has been done on the levels of graduates' competencies. It is important to know the classification of students' competencies to be used as consideration in developing learning outcome evaluation. The present study is an effort to do just that.

In order to know the nature of the competency of students who take the examination, an initial effort must be made to look at the test instrument. It is a fact that, up to the present time, the test instrument that is used for the national examination in accounting is not well reviewed. A teacher in the interview stated that the test items that had just been developed for trial examination were administered right away, before being tried out first. A good test item must first go into analyses of differentiating power, difficulty level, and distractor function. This way, a student's competency can be classified into very low, low, medium, high, or very high.

Based on the preceding background, the researchers are interested in empirically attempting to look at the quality of test items of the exam and classification of the competencies of students of the accounting study program of all the vocational schools in Sleman regency. Empirical evidence is obtained by collecting responses of students

taking the try-out of the national examination of the three subject matters of the accounting subjects of the 2015/2016 academic year developed by members of the Accounting Teachers' Association of Sleman Regency.

Research Method

The study employs quantitative research approach of the descriptive explorative method. Results of the study are expected to be able to describe the quality of the test items and students' competencies in the accounting subject matter. Data were taken from students' responses in the regional test trial of the national examination developed by accounting teachers, members of the accounting teachers' professional association of Sleman Regency.

Findings and Discussion

The classical-based item analyses conducted in this study produce levels of item difficulty, discriminating powers, test reliability, and standard errors of measurement. There are 19 items accepted as good items.

Characteristics of the Test on the Classical Theory

Level of Item Difficulty

Results of the item analyses show that the levels of difficulty of the test items are found to range between 0.075 and 0.971 with a mean score of 0.556. Referring on the criterion by Crocker & Algina (1986, p. 313) and Wright & Masters (2008, p. 227), 27.5% or 11 items are of the easy category, 55% or 22 items are of the medium category, and 17.5% or 7 items are of the difficult category.

Discriminating Power

All items have positive discriminating powers, although of various degrees. It means that all the correct answer has functioned well. Most of the items can differentiate high-achieving students from low. Most of the high-achieving students choose the correct answers, while the low-achieving

students choose the distractors. Scores of the discriminating powers range from 0.032 to 0.698 with a mean score of 0.412.

The point-biserial correlation of the item analyses shows that ten items (25 %) are weakly discriminating. These ten items have a discriminating power of lower than 0.3 (Kartowagiran, 2012; Reynolds, Livingston, & Willson, 2009).

Reliability and Standard Error

The reliability index (alpha) is 0.880 with a standard error of measurement (SEM) of 2.582. It means that the test can be categorized as reliable since the alpha index satisfies the minimum line of 0.7 (Linn, 1989, p. 106; Mardapi, 2014). It is in agreement with Safrudin Amin’s study that finds a reliability index of 0.874. Meanwhile, the SEM score of 2.582 means that, on the confidence level of 95%, a student with a raw score of X will have his real score on the interval of $X \pm 2 \text{ SEM} = X \pm 5.164$.

Characteristics of the Test on the Modern-Theory Approach

IRT Pre-requisite Test

Uni-dimensionality

For the requirement of the factor analysis, the analysis sample of the study can be categorized as “good” since it is higher than 300. Williams, Onsmann, & Brown, 2003 suggest that an analysis sample should minimally be 100 or over. More specifically, it is stated that a sample of 100 is poor, 200 is fair, 300 is good, 500 is very good, and 1000 or more is excellent. Feasibility of test samples can be determined by KMO-MSA and Barlett’s test of sphericity (see Table 1).

Table 1. KMO-MSA and Barlett’s Test of Sphericity

KMO-MSA and Barlett’s Test			
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.878	
Barlett’s Test of Sphericity	Approx. Chi-Square Df. Sig.	4075.252 780 0.000	

A KMO-MSA value is regarded adequate if it passes 0.5 (Field, 2009, p. 660). Results of the study show that the KMO-MSA is 0.878. Calculations show a Barlett’s Test of the Sphericity significance level of 0.000. It means that the requirement is fulfilled since the significance level obtained is lower than 0.05.

According to Reckase (1979), Smits, Cuijpers, & van Straten (2011), and Wu et al. (2013), the uni-dimensionality assumption is fulfilled if “the first factor should account for at least 20 percent of the test variance”. The variance that can be explained amount to 57.351% and the contribution of the first factor is 19.860% (see Table 2). Since the first factor accounts for almost 1/5 of the variance test, it can be concluded that the unidimensionality assumption is satisfied.

Table 2. Eigenvalue and Component of Variance (13 Components)

Component Number	Eigen Value	Proportion	Cumulative
1	7.944	19.860	19.860
2	1.788	4.471	24.331
3	1.506	3.764	28.095
4	1.445	3.613	31.708
5	1.312	3.279	34.987
6	1.257	3.141	38.129
7	1.180	2.949	41.078
8	1.146	2.866	43.944
9	1.139	2.846	46.790
10	1.083	2.707	49.497
11	1.074	2.685	52.182
12	1.062	2.655	54.837
13	1.006	2.514	57.351

Egan, Sireci, & Swaminathan (1998) add that “if a data set is unidimensional, then the first eigenvalue should explain a relatively large proportion of the variance”. Results of the factor analysis in Table 2 show that more than 13 eigenvalues have a score higher than 1, where the first factor is the most dominant, 7.944; almost five times higher than those of the following factors which are almost equal. Since the variance scores have a linear comparison with the eigenvalue (Field, 2009, p. 652; Johnson & Wichern, 2002, p. 441) and the first factor accounts for a bigger

contribution than the other factors, then the assumption of uni-dimensionality is fulfilled.

The results of the uni-dimensionality test presented graphically in a scatter plot can be seen in Figure 1. According to Hambleton & Rovinelli (1986), as cited by Stage (2003), the number of significant factors is usually shown by the appearance of an “angle” in the plot. The scree plot in Figure 1 signifies that an angle has been formed at a point on the left side. It means that the uni-dimensionality assumption is fulfilled.

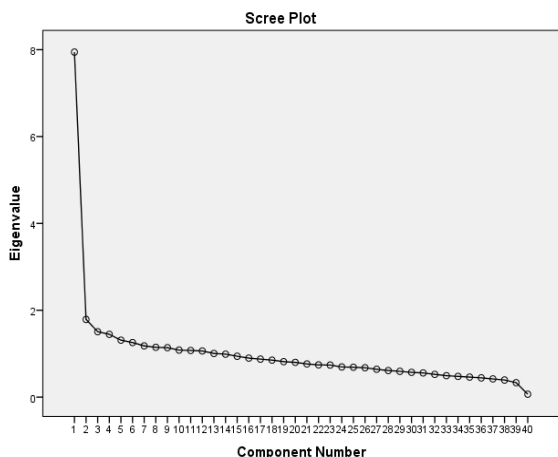


Figure 1. Eigenvalue Scree Plot

Local Independence

In general, all the elements outside the main diagonal matrix are too small, closing to zero. It can show that the local independence assumption has been fulfilled.

Parameter Invariance

Results of the item parameter estimation (in this case, levels of difficulty since the analysis uses the Rasch model) of each sample are presented in a scatter plot and correlated. Positive high correlation shows that parameter invariance is satisfied (Retnawati, 2014, p. 8). Figure 2 presents an estimation plot for item parameter invariance. From Figure 2, it can be seen that the estimate values are located relatively close to the straight line with a high correlation score (0.9881). It can be concluded then that the assumption for the item parameter invariance is fulfilled.

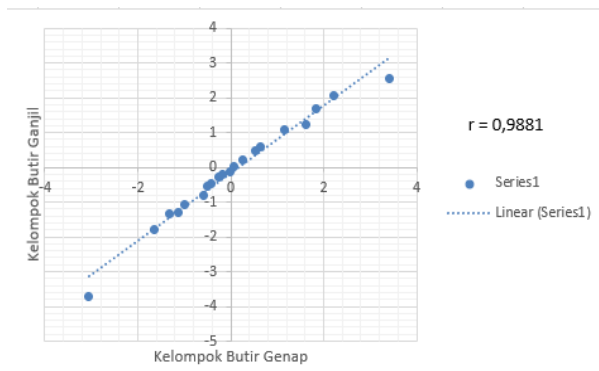


Figure 2. Parameter Invariance Plot of Item Difficulty Levels

To test the competence parameter invariance (θ), the forty test items are divided into two groups of subtests according to the results of the item parameter estimation of each sample are presented in a scatter plot and correlated. Positive high correlation shows that parameter invariance is satisfied. to the item numbers, subtest I consisting of odd numbers and subtest II even numbers (Retnawati, 2014, p. 9). Figure 3 presents the scatter plot of the competence parameter in accordance with the item groups done by the students. In Figure 3, the estimate values are located close to the straight line with a high (substantial) correlation score of 0.9989.

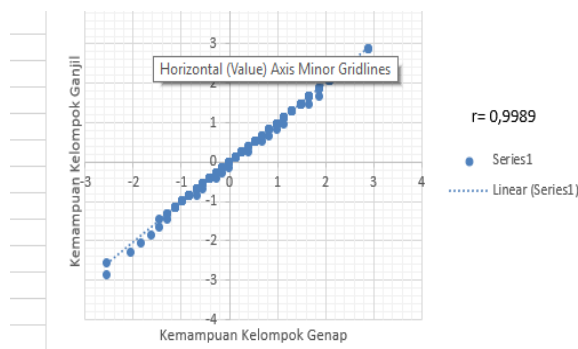


Figure 3. Parameter Invariance Plot of Students' Competencies

Instrument Characteristics

Analyses on the characteristics of the test under study include model fits, item parameter, and testees' characteristics, TIF, and SEM. Each characteristic is elaborated as follows.

Model Fit

The analysis carried out in the study makes use of the WINSTEPS program of IRT of the Rasch model. A test is regarded as fit with the item difficulty and the testees if the outfit MNSQ value is in the range of 0.5-1.5 (Linacre, 2002). Results of the study show that five items are found to be not fit with the model. These are 6, 11, 18, 31, and 40. In term of the testees, 59 students are found to be not fit with the Rasch model since they are outside the MNSQ outfit range.

Item Parameter and Testees' Characteristics

A total of 40 items and 414 students are subjected to the analyses. An item is categorized as "good" if it fulfills two requirements, namely: it has a good difficulty level ($-2 \text{ logit} \leq b_i \leq +2 \text{ logit}$) (Hambleton & Swaminathan, 1985) and it has a model fit. In the study, six items (15%) of the total 40 are not in the "good" category. They are items 1, 2, 18, 31, 37, and 40. Item 18 has the highest difficulty level (+3.4 logit), and item 2 has the lowest difficulty level (-3.72 logit). Meanwhile, testee number 146 has the highest competency (+2.89 logit) and the testee number 89 the lowest (-2.87 logit). Figure 4 presents the distribution of the item difficulty levels. From Figure 4, it can be seen that 34 items can be accepted.

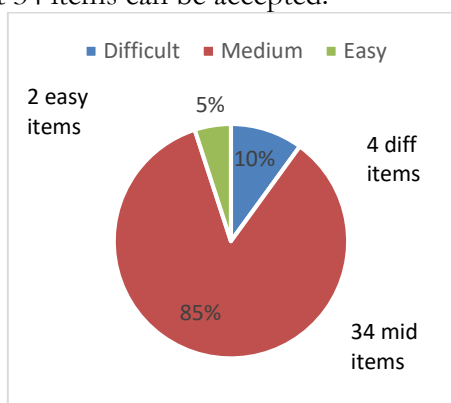


Figure 4. Distribution of Item Difficulty Levels

Test Information Function (Tif) and Sem

Results of the analyses using the Rasch model show that the test set has a maximum

information function (TIF) amounting to 16.737 on competencies around -0.2 logit. According to Hambleton (in Wiberg, 2004), a reliable test has a TIF value of ≥ 10 . In the study, the test instrument can be regarded as reliable in measuring the testees' competencies in accounting. Meanwhile, SEM values have a reverse criterion from TIF. It means that the test will have a good TIF if it has the lowest SEM value (0.2444) and answered by testees with a competence level of around -0.2 logit (of the mid-high category). Using the SEM value and what has been calculated, the interval of testees' competencies can be obtained using this equation (Hambleton & Swaminathan, 1985, p. 90):

$$\theta - z \frac{\alpha}{2} [I(\theta)]^{-\frac{1}{2}} \leq \theta \leq \theta + z \frac{\alpha}{2} [I(\theta)]^{-\frac{1}{2}}$$

Since $[I(\theta)]^{-\frac{1}{2}} = SEM$ at the confidence level 95%, the formula becomes:

$$\theta - 1.96 SEM \leq \theta \leq \theta + 1.96 SEM$$

Based on this equation, it can be stated that the test will give good information (TIF) if taken by testees of the interval range of $-0.678 \text{ logit} \leq \theta \leq 0.278 \text{ logits}$. Visualization of the test TIF and SEM is presented in Figure 5.

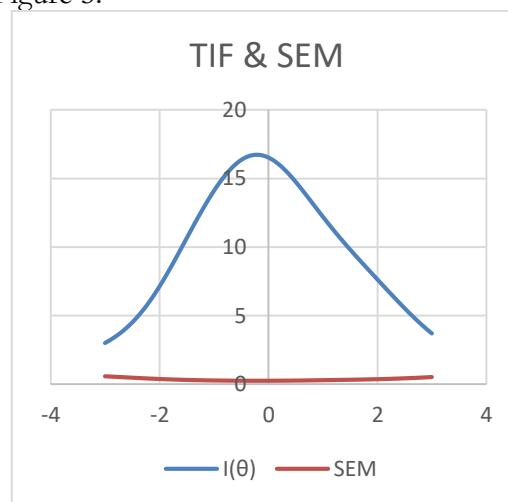


Figure 5. Relation between TIF and SEM of the Test

Classification of Students' Competencies in Accounting

Students' competencies can be graded into five categories: (1) very high, (2) high, (3) medium, (4) low, and (5) very low. Students' competencies can be seen from the

tetha measure in the analysis using Winstep software program. Prior to this, to use the Winstep, the test items that are not fit for the Winstep program are not included. Items that are not used in the analysis are numbers 6, 11, 18, 31, and 40. Results of the data analysis of the 414 students can be seen in Table 3. From Table 3, it can be seen that the very high category of students' competencies is occupied by 39% of the students, the high category 3%, the medium category 6%, the low category 7%, and the very low category 45%.

Table 3. Learning Competency Categories of the Accounting Students

Category	Number of Students	Percentage (%)
Very high	159	38
High	14	3
Medium	24	6
Low	30	7
Very low	187	45
Total	414	100

Conclusion

By the classical theory approach, it is found that the average measure of the item difficulty level is in the "medium" category, the test items have a good measure of distractor functions, and the test is reliable. Concerning the difficulty level, discriminating power of, and distractor functioning, a total of 19 items (47.5%) of the test are in the "good" description. By modern-approach analyses, it is found that the average of the difficulty level is in the "medium" category. Given the difficulty level and model fit, a total of 34 items (85%) are in the "good" category. Based on the measures of the test information function (TIF) and SEM, the test allows the best for students with a competency range of $-0.678 \text{ logit} \leq \theta \leq 0.278 \text{ logit}$. In view of the item response theory (IRT), students' competencies can be grouped into five categories; namely very high with 159 students (38%), high with 14 students (3%), medium with 24 students (6%), low

with 30 students (7%), and very low with 187 students (45%).

One implication that can be given is for the results of the study to be an input to teachers of the Accounting Teachers Professional Association in Sleman Regency in developing test items. Since students' accounting competencies in the 2015/ 2016 academic year cannot be measured maximally because of the low quality of the test, training is needed for teachers to develop and analyze test items. Use of the IRT and classical-theory analysis gives different results; caution is therefore needed in reviewing the existing tests.

References

- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.
- Egan, K. L., Sireci, S. G., & Swaminathan, H. (1998). Effect of item bundling on the assessment of test dimensionality. In *the paper presented at the annual meeting of the National Council on Measurement in Education*. San Diego, CA.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd 3d.). London: Sage Publications.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Kartowagiran, B. (2012). Penulisan butir soal. In *the paper presented in Training on Writing and Analysis of Items for the Civil Servant-Rekinpeg Resource*. Hotel Kawanua Aerotel, Jakarta.
- Law of Republic of Indonesia No. 20 of 2003 on National Education System (2003).

- Linn, R. L. (1989). *Educational measurement*. New York, NY: Macmillan.
- Manoppo, Y., & Mardapi, D. (2014). Analisis metode cheating pada tes berskala besar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 18(1), 115–128. Retrieved from <https://journal.uny.ac.id/index.php/jpep/article/view/2128/1773>
- Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.
- Mardapi, D. (2014). Authentic assessment. In *the paper presented at HEPI Conference*. Denpasar, Bali.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4(3), 207–230. <https://doi.org/10.3102/10769986004003207>
- Regulation of the Minister of National Education No. 19 of 2005, on National Standard of Education (2005). Republic of Indonesia.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika.
- Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.
- Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1), 147–155. <https://doi.org/10.1016/j.psychres.2010.12.001>
- Stage, C. (2003). *Classical test theory or item response theory: The Swedish experience*. Santiago, Chile: Centro de Estudios Públicos.
- Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test*. Stockholm: Umea Universitet.
- Williams, B., Onsmann, A., & Brown, T. (2003). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 8(3), 1–13. Retrieved from <https://ajp.paramedics.org/index.php/ajp/article/view/93/90>
- Wright, B. D., & Masters, G. N. (2008). *Rating scale analysis: Rasch measurement*. Chicago, IL: Mesa Press.
- Wu, Q., Zhang, Z., Song, Y., Zhang, Y., Zhang, Y., Zhang, F., ... Miao, D. (2013). The development of mathematical test based on item response theory. *International Journal of Advancements in Computing Technology*, 5(10), 209–216.