Gadjah Mada Journal of Psychology (GamaJoP)
Journal Homepage: https://jurnal.ugm.ac.id/gamajop
Email : gamajop.psikologi@ugm.ac.id
ISSN 2407 - 7798 (Online)

Volume **10**
Number 1, 2024
Page: 23–30
DOI:10.22146/gamajop.79149

Received 14 November 2022
Revised 1 February 2023
Accepted 2 March 2023
Published 31 May 2024

Keywords:
educational technology; UGM's eLOK; psychology; psychometrics

***Author for correspondence:** Email: marvianto.ramadhandwi@gmail.com

The use of UGM's eLOK as a Student Learning Outcomes Evaluation Platform

Ramadhan Dwi Marvianto,*[,1] Haryanti Mustika,[1] and Sri Suning Kusumawardani[2]

[1]Faculty of Psychology, Universitas Gadjah Mada, Indonesia
[2]Faculty of Engineering, Universitas Gadjah Mada, Indonesia

Abstract

The use of online platforms in testing is a necessity nowadays, especially in an academic context. The platform used in UGM, namely eLOK, provides a testing facility which enable the lecturers to simultaneously see the results of the test evaluations and questions directly from the platform. However, there haven't been any studies that compare the evaluation results using UGM's eLOK with other approaches. Therefore, this study compared evaluation results from UGM's eLOK with the Classical Test Theory (CTT) and Item Response Theory 2-Parameter Logistics (IRT 2-PL) approach using the Graded Response Model (GRM). This study included 22 active students who took the test using the UGM's eLOK platform in the Multivariate Statistics course during the even semester of 2020/2021 academic year. The results of the analysis showed that the evaluation using the UGM's eLOK platform had close equivalence with the CTT approach, although each parameter's value was slightly different. In addition, the results of the IRT analysis were found to have far differences with the other two methods, but these results only slightly reflect the actual parameters due to the minimal number of subjects. The results of this study can be used as a reference in using UGM's eLOK as a student academic testing platform, where lecturers are able to evaluate the quality of the tests and items given to test.

Assessment is a very important activity carried out by teachers to determine the level of development of student learning outcomes (Sumardi, 2020). This learning assessment includes tests, measurements, assessments, and evaluations of learning to determine the effectiveness of the learning process. This article's focus is to discuss the importance of measurement in knowing the effectiveness of learning, as many human activities require measurement as a way to describe the characteristics of certain people or objects (Sumardi, 2020).

The measurement itself, according to Sumardi (2020), is an activity intended to determine the size of an object in the form of numbers. However, the instruments used by each profession to take measurements are different. Allen and Yen (1979) also define measurement as a systematic procedure used to determine numbers that represent the characteristics of certain individuals or objects. In the context of learning, these numbers refer to the scores obtained by students after taking certain exams or tests. The process of determining this number should not be done arbitrarily, but carried out carefully based on predetermined procedures that should be repeatable (Sumardi, 2020).

eLOK (e-Learning: Open for Knowledge Sharing) is an e-learning system or online learning system owned by Universitas Gadjah Mada. This system is assembled using the Moodle database, which is a software package for internet-based or website-based learning activities. One of the benefits of eLOK is that it can take measurements automatically. This automatic measurement has the usefulness of being able to produce objective measurements, both in measuring student scores and measuring the effectiveness of the items. The UGM's eLOK uses the Moodle platform, which uses its own method of estimating where this platform has also been widely used in the world of education (Butcher, 2021).

In addition, the main advantage of eLOK is that it is a form of contribution from Universitas Gadjah Mada in the context of enlightening the Indonesian society by utilizing technology (for Academic Innovation & Studies, 2021). One of the uses of eLOK in measurement is as a tool to test learning outcomes. The purpose of using eLOK in the learning outcomes test is to measure the participants' mastery of the

materials that have been taught, as well as measuring the learning progress of students. The usage of eLOK in measuring students' learning outcome is beneficial, especially during this COVID-19 pandemic, as it can be taken online and thus can streamline the limited time constraint. In addition, the advantages of this measurement are that participants can be expected to know the parameters of success in the implementation of education through the scores obtained as well as determining how much of the knowledge they've learned has been retained.

Aside from discussing the results of the analytical methods obtained from eLOK, this study also perform analysis using classical and modern approaches, otherwise known as the Classical Test Theory (CTT) and Item Response Theory (IRT) approaches. CTT is a measurement model that works at the level of test scores by using a linear model in explaining the score model without discussing the relationship between items and specific abilities (**Azwar2016**). However, **Azwar2016empty citation** states that CTT has contributed a lot in the world of measurement tools; almost the entire psychometric property formula was developed under this concept. On the other hand, IRT is a model that was built so that latent psychological constructs can be expressed in the form of responses to items in order for them to be observed (**Azwar2016**). This model, according to Harvey and Hammer (1999), is very useful in the development, evaluation, and scoring of a testing instrument.

There have been many studies comparing CTT and IRT as a measurement. One of the studies regarding IRT and CTT was conducted by Ramadhaningtyas (2018) compared the two approaches with an automated measurement, namely eLOK. This research was done to see the comparison results between these 3 analytical methods. It aimed to know which method is the most effective and accurate in measuring the learning outcomes. Sumardi (2020) explained that in order for the competence of participants to be measured appropriately. Because of this, it is necessary for the measuring instrument to be accurate. Based on these explanations, it is formulated that the purpose of this research was to see how the analytical method built on the UGM's eLOK platform compared to those that use the CTT and IRT approaches.

In comparing these three methods, the parameters of difficulty level and item discriminating power are the main parameters. Azwar (2016a) explained that the level of difficulty refers to how easily the item can be answered by the population or all test participants. Meanwhile, the discriminating power of the item refers to how well the item distinguishes between individuals with high and low ability levels. These two parameters aim to see the quality of the test at the item level.

In addition to the item level, this study also compared the 3 methods from the test level. The test parameters used are reliability and Standard Error of Measurement (SEM). Reliability is a parameter which shows the level of consistency of the measurement results (Furr & Bacharach, 2013). On the other hand, SEM can be interpreted as a deviation value from the test score obtained (**Azwar2016**), so that this value will provide an overview of the confidence interval of the score generated on a test.

Method

Research Participants

This study included students of the 2020 Master of Psychology study program in the even semester of the 2020/2021 academic year. The number of test participants was 22 students who were registered as active students in the Multivariate Statistics course. This number is relatively small when compared to the general analyses that had been done using CTT and IRT, with more than 200 participants (Riswan, 2021) or above 60, especially CTT (Azwar, 2013, 2016a, 2016b). Nevertheless, the use of analysis can still be done with the consideration that this study aims to compare the psychometric properties of the three approaches with the same number of samples. In addition, the number of participants also depends on the conditions that exist for the CTT approach, so it is not appropriate to determine a certain number of participants. However, in the IRT approach, the number of participants will have an effect as the smaller it is, it will allow misleading results (Cappelleri et al., 2014).

Research Instruments

The research instrument used was a test of learning outcomes in the Multivariate Statistics course. The test is carried out using a standardized procedure, where students take the test through UGM's eLOK proctored by a supervisor who accompanied students on the Zoom meeting platform. This test consisted of 50 questions and was carried out for a maximum of two hours with supervision, and test takers were given permission to open notebooks or search on various sources as it was an open book type of test.

Data Analysis

Considering the purpose of this study, the analysis was carried out using three different approaches, namely (1) the approach used in UGM's eLOK, (2) classical-test theory (CTT), and (3) item-response theory with two parameters logistics (IRT-2PL). IRT analysis with 2 logistic parameters refers to the IRT analysis approach that considers the value of discriminating power and level of difficulty, so that these two parameters can be estimated in this analysis (de Ayala, 2009). The first approach was carried out directly on the eLOK platform while the CTT and IRT approaches were carried out in the R Studio software with the help of several main packages, namely base (Team, 2019), CTT (Willse, 2018), and mirt (Chalmers, 2012) as well as package assistance that included knitr (Xie, 2014), readxl (Wickham & Bryan, 2019), dplyr (Wickham et al., 2022) and openxlsx (Schauberger & Walker, 2021).

Results

(See Table 1)

The results of the analysis on each approach show unique results. In Table 1, the analysis results obtained from UGM's eLOK were displayed as percentage, while those from CTT and IRT used integer form. Even so, the differences in the format can be slightly ignored as there are still similarities in the concepts of these three approaches.

Table 1 showed that the level of test difficulty generated by the UGM's eLOK shows that 85.82% of test takers can answer all of the questions. This was classified as the same as the results using the CTT approach (50 items), which showed that 86% of participants can answer correctly.

In IRT analysis, items that can be analyzed were only those who had varied responses. Therefore, item number 7, 13, 21, 22, 23, 25, 26, 27, 31, 33, 34, and 35 were excluded from the analysis using the CTT (37 items) and IRT approaches. The results of CTT analysis using 50 items and 37 items were not significantly different because it was only reduced by 5%. On the other hand, the IRT analysis shows the difficulty level that was classified as difficult, which was 2.32.

In terms of discriminating power, the results of the analysis using the eLOK platform have a value of 29.94%, which was equivalent to the results of the analysis using the CTT approach, both 50 items and 37 items, as items that do not have variations cannot produce discriminating power thus do not affect the average value of discriminating power of the test. On the other hand, the results of the IRT analysis showed a relatively high value, namely 14.30.

In terms of reliability, the CTT analysis on both 50 and 37 items as well as the eLOK platform showed almost equal values, namely 0.82 and 81.64%, respectively. This result was slightly below the reliability result using the IRT empirical reliability formula which was 0.91. However, the coefficients on all approaches were considered satisfying, or it can be said that the measurement results using this test were reliable or trustworthy.

Furthermore, the Standard Error of Measurement (SEM) value from the eLOK platform showed a value of 4.26 % which was slightly different from the CTT approach which was 2.1 for both 50 and 37 items. In the IRT approach, SEM is generated for each participant so they were not excluded in the test level.

In addition, based on the mean score and Standard Deviation (SD), the eLOK platform showed a score that was double (in percentage) the 50-items CTT score. This was because the highest score on the eLOK platform was 100% while the highest score of 50-items CTT was 50. On the other hand, IRT analysis produced an estimation of individual's ability in a continuum that ranges from $-\infty$ up to ∞. Table 1 showed that the average of ability score obtained was -0.02, or 29.82 if converted into a raw score. This score was almost similar to the mean score obtained in the CTT approach with 37 items. In addition, the mean SD scores in the two approaches were quite similar, namely 4.93 for CTT and 4.03 for IRT.

Item Level

Table 2 showed that the analysis results on the UGM's eLOK platform with the CTT approach can be said to be equivalent, both in difficulty level as well as discriminating power. The difference that presented was because the CTT approach uses two numbers to be rounded after the comma, while the eLOK platform uses a percentage as well as two numbers after the comma. If the value of the UGM's eLOK platform is rounded up, the results showed the same value.

Analysis results of the two approaches above showed that almost all items were classified as easy questions (> .70) (Azwar, 2016a, 2016b) and some were classified as very easy according to UGM's eLOK, except for item number 9, 10, 15, and 20. These four items were classified into moderate items according to the CTT approach (.30 <x <.71) (Azwar, 2016a, 2016b) and the UGM's eLOK platform (35–65%) (Butcher, 2021). In terms of discriminating power, most of the items, except for items that have a difficulty level of 100% or 1.00, showed optimal values in the CTT approach (> .40) (Crocker & Algina, 1986; Ebel, 1965) and the UGM's eLOK approach (> .30) (Butcher, 2021). Some items are quite satisfactory, but need a little revision as they have values above .20 or 20% (Butcher, 2021; Crocker & Algina, 1986), namely item number 16, 18, and 20. Nevertheless, the two approaches above assessed item number 3 as having poor discriminating power as its value was below .10 or 10% (Butcher, 2021; Crocker & Algina, 1986) or 10%.

On the other hand, the IRT approach showed quite different figures. In addition, the results of the analysis with this approach are quite different in interpretation from the two previous approaches. The value of discriminating power (a) which is above 1.35 indicates that the item can discriminate against test takers optimally (Baker & Kim, 2017). The results show that there are contradictions in the results of the IRT approach with the previous approach, namely items 12, 15, and 20 have poor discriminating power according to IRT, but have good discriminating power according to the previous approach. In addition, item number 3 was found to have poor discriminating power in the CTT and UGM's eLOK approaches, but has satisfactory discriminating power in the IRT approach.

In addition, the results of the IRT approach also showed that all items were classified as difficult according to Hambleton (1991), as those were higher than .5 except for items 3, 11, and 12 which were classified as moderate because they were between -.50 to .50. This finding was in stark contrast to the results of the CTT and UGM's eLOK analysis whose items were classified as easy. Moreover, items that were classified as moderate in the CTT and UGM's eLOK approaches were classified as difficult in the IRT approach. (See Table 3)

The same finding was also shown in Table 3 where all items were classified as easy with the results of the UGM's eLOK analysis and the CTT approach except for item

Table 1
Summary of Analysis Results of Test Level with CTT, IRT, and UGM's eLOK Approaches

Methods	*N* Items	Item Difficulty	Item Discrimination	Reliability	SEM	Score Mean	SD Mean
eLOK	50	85.82a	29.94a	81.64a	4.26a	85.82a	9.95a
CTT	50	.86	.30	.82	2.13	42.91	4.98
CTT	37	.81	.30	.82	2.10	29.91	4.98
IRT	37	2.32	14.30	.91	-	29.82	4.03
						(-.02)b	(.29)b

Note. [a]in percent; [b] estimated ability or theta (θ), *SEM* = Standard Error of Measurement, *SD* = Standard Deviation of Score

Table 2
Summary of Item Level Analysis Results with CTT, IRT, and UGM's eLOK Approaches for Item Number 1 – 25

Butir	eLOK			CTT			IRT	
	b	a	DE	b	a	φ	b	a
1	86.36%	45.52%	72.90%	.86	.46	.71	5.43	6.75
2	86.36%	42.61%	68.38%	.86	.43	.67	9.70	11.61
3	95.45%	4.30%	9.59%	.96	.04	.11	0.38	3.12
4	86.36%	39.72%	61.10%	.86	.40	.62	2.54	3.58
5	81.82%	43.73%	61.22%	.82	.44	.63	3.35	3.87
6	90.91%	40.04%	75.00%	.91	.40	.73	3.70	5.52
7	100.00%			1.00				
8	90.91%	29.25%	52.94%	.91	.30	.54	2.38	4.01
9	63.64%	60.45%	70.86%	.64	.61	.79	2.48	1.40
10	68.18%	43.30%	52.01%	.68	.43	.56	2.83	2.03
11	81.82%	41.12%	57.69%	.82	.41	.59	.01	1.50
12	77.27%	53.19%	71.59%	.77	.53	.72	.35	1.27
13	100.00%			1.00				
14	90.91%	46.90%	87.50%	.91	.47	.86	.35	6.31
15	45.45%	63.39%	74.09%	.46	.63	.94	48.68	-3.00
16	95.45%	27.07%	59.82%	.96	.27	.67	3.96	6.87
17	100.00%			1.00				
18	77.27%	22.10%	29.62%	.77	.22	.30	1.03	1.54
19	81.82%	51.60%	74.42%	.82	.52	.74	1.41	2.11
20	63.64%	32.65%	38.36%	.64	.33	.43	1.20	.81
21	100.00%			1.00				
22	100.00%			1.00				
23	100.00%			1.00				
24	90.91%	26.48%	47.06%	.91	.27	.48	1.67	3.28
25	100.00%			.00				

Note [a] difficulty level; [b] discriminating power; [DE] discriminative efficiency; [1] facility index; [2] the proportion of number of participants that answered correctly; [3] point-biserial correlation (rPBis); [4] difficulty level of IRT/delta (δ); [5] IRT/alpha discriminating power (α); φ= item analysis index (phi coefficient)

Table 3
Summary of Item Level Analysis Results with CTT, IRT, and UGM's eLOK Approaches for Item Number 26 – 50

Butir	eLOK			CTT			IRT	
	b	a	DE	b	a	φ	b	a
26	100.00%			1.00				
27	100.00%			1.00				
28	95.45%	-.20%	-.46%	0.96	.00	.00	1.73	4.17
29	95.45%	-22.45%	-50.68%	0.96	-.23	-.56	.17	3.06
30	95.45%	40.95%	100.00%	0.96	.41	1.01	51.37	77.44
31	100.00%			1.00				
32	90.91%	19.78%	35.29%	.91	.20	.36	1.54	3.16
33	100.00%			1.00				
34	100.00%			1.00				
35	100.00%			1.00				
36	45.45%	16.23%	19.02%	.46	.16	.24	.46	-.18
37	86.36%	51.38%	81.93%	.86	.51	.81	7.82	9.47
38	90.91%	50.36%	93.75%	.91	.50	.92	65.75	85.58
39	95.45%	4.30%	9.59%	.96	.04	.11	-47.41	90.07
40	95.45%	4.30%	9.59%	.96	.04	.11	-47.41	90.07
41	95.45%	4.30%	9.59%	.96	.04	.11	-47.41	90.07
42	40.91%	30.57%	38.27%	.41	.31	.49	-.01	-.37
43	81.82%	43.73%	61.22%	.82	.44	.63	1.60	2.26
44	86.36%	28.25%	43.81%	.86	.28	.44	-.45	1.92
45	86.36%	28.25%	43.81%	.86	.28	.44	-.45	1.92
46	90.91%	29.85%	52.94%	.91	.30	.54	.60	2.47
47	86.36%	51.38%	81.93%	.86	.51	.81	2.63	3.67
48	50.00%	-8.39%	-9.28%	.50	-.08	-.12	-.81	-.04
49	86.36%	28.25%	43.81%	.86	.28	.44	.56	1.98
50	40.91%	-6.65%	-8.24%	.41	-.07	-.10	.01	-.37

Notes. [a]level of difficulty; bdiscriminating power; [DE]discriminative efficiency; [1]facility index; [2]the proportion of number of participants who answered correctly; [3]point-biserial correlation (rPBis); [4]IRT difficulty level/delta (δ); [5] = IRT discriminating power/alpha (α); φ = item analysis index (phi coefficient)

number 36, 42, and 50. In terms of discriminating power, the two approaches showed consistent results where both approaches displayed that item number 28, 29, 36, 39, 40, 41, 48, and 50 have unsatisfactory discriminating power.

On the other hand, in the IRT approach, there were some fundamental differences, namely item number 28, 29, 39, 40, and 41 had satisfactory discriminating power (>1.35), contrast to those in the other two approaches that have poor discriminating power. However, this approach showed consistent discriminating power in item number 36, 48, and 50, which all of them had poor discriminating power.

From the level of difficulty, it can be seen that several items have different results between the IRT approach and the two previous approaches. First, item number 39 to 41 had very easy level of difficulty (<-2) (Hambleton, 1991), but all four items were classified as easy in the other two approaches. Moreover, item number 44 and 45 which were classified as moderate in the IRT approach had easy level of difficulty in the other two approaches. Lastly, the remain items were classified as difficult in the IRT approach, even though they were classified as easy in the other two approaches.

In addition to the discriminating power and level of item difficulty, the approach in UGM's eLOK issues Discriminative Efficiency (DE) which shows how effectively the item distinguishes between upper- and lower-class individuals based on the level of difficulty (Butcher, 2021). This value was classified as good if it is above 50%. Furthermore, in the CTT approach, this value was also provided by a formula called the item analysis index or phi-coefficient (F) (Aiken, 1979) which is based on the level of difficulty and discriminating power. These two parameters were found to have equivalent values, as in the example in item number 2, eLOK has a DE of 68% and CTT has an F of .67. Meanwhile in item number 48, eLOK has a DE of -8% whereas CTT has an F of -.10. (See Table 4)

Score Correlation

Table 4

Matrix of Score Correlation using IRT Approach (q), IRT Raw-Score, CTT, and UGM's eLOK

	Score	1	2	3
1	Abilitas (θ) IRT			
2	Expected Score	.801***		
3	CTT	.825***	.689***	
4	eLOK UGM	.825***	.689***	1.00***

Note. ***significant below .001

Furthermore, correlation results between the scores generated by each approach were shown in Table 4. The IRT ability score was score obtained from the estimation of IRT 2-PL using the Expected A Posteriori (EAP) method which was a very general and easy-to-use method for unidimensional constructs (Brown & Croudace, 2015) such as the test in this course. The expected score was obtained from the expected test function of the IRT analysis feature. In contrast to IRT, the CTT score was obtained purely from the total score obtained on 37 items that have discriminating power. In addition, the UGM's eLOK score is obtained from the percentage of correct answers divided by the total score.

Then, the correlation matrix shown in Table 4 shows that the UGM's eLOK score with the CTT score has a perfect correlation, which is 1 ($p < .001$). In addition, the correlation between the θ score and the IRT raw-score and CTT and UGM's eLOK scores showed a high correlation value, which were .801 and .825, respectively ($p < .001$). These four correlations were classified as high correlation, as it was above 0.70 (Hinkle et al., 2003). On the other hand, the correlation between IRT raw-score and CTT and eLok was moderate, which was 0.689 ($p < .001$) or between .30 – .70 (Hinkle et al., 2003). These results can be interpreted that the scores generated from the three analytical approaches were equivalent.

Discussion

This study aimed to compare the analytical methods built on the UGM's eLOK platform compared to the CTT and IRT approaches. The results showed that eLOK and CTT analysis had similar results, while producing far different results compared to the IRT approach.

The reliability test result showed equality in the analysis on the UGM's eLOK platform with the CTT although those were presented in different forms, namely in the form of fractions in CTT and percentages in UGM's eLOK. Nunnally and Bernstein (1994) said that the measurement with a reliability value above 0.70 can produce a reliable or trustworthy score.

In addition, reliability is also related to the Standard Error of Measurement (SEM) where this value shows the true value range (confident interval of true score) of the test participants (Azwar, 2013, 2016a, 2016b, 2017; Furr & Bacharach, 2013). However, this value does not have a standard to be said whether it is satisfactory or not. Thus, SEM scores can be interpreted together with individual scores in this test.

For example, individual X got a score of 40 from the CTT approach, or equal to eLOK score of 80%. By using the confidence level (α) of 0.05 ($z = 1.96$ or equivalent to 2) and the SD of the test was 4.98 or equivalent to 5 (equivalent to 9.95 or equal to 10 in the SD of eLOK test), so the true value of the individual X will move between $40 \pm (2 \times 5)$ or 30 to 50. This value is considered high because the true probability range of individual X was about of 20 units. However, this test was an open book test so that it made the actual score of the individual more difficult to find because the correct answer from the individual is not necessarily from his knowledge, but from his speed in finding information instead.

In addition to these two properties, the CTT and UGM's eLOK approaches were equal in difficulty level and discriminating power even though they had different forms, namely fractions and percentages, respectively.

This means that the evaluation process carried out in UGM's eLOK which uses the guidance from Butcher (2021) was equivalent to the commonly used approach, namely CTT.

This equality can be seen from the interpretation of the difficulty level, which means the percentage (UGM's eLOK) or proportion (CTT) of items can be answered correctly by the test (Azwar, 2016a, 2016b). In addition, the analysis results of the items' discriminating power on eLOK were also equivalent to those with the CTT approach which uses a high- and low- group approach (Azwar, 2016a, 2016b). Thus, the use of eLOK for test presentation was supported by an analysis that was equivalent to CTT so that it became an evidence that testing through eLOK produce a certain quality in test evaluation.

In addition to the parameters of discriminating power and difficulty level, UGM's eLOK also produced Discriminative Efficiency (DE) parameters which are computations of the level of difficulty and discriminating power of items (Butcher, 2021). The results of this study found that DE had a value that was not much different from the item index analysis, $\varphi coefficient$ which was proposed by Aiken (1979). In his study, he suggested that This phi coefficient shortened the use of difficulty level and discriminating power to select items, thus the phi coefficient was used to combine the two parameters. His study explained that a phi coefficient with a value that exceeds .5 also indicates that the item works optimally on the level of difficulty associated with discriminating power. This report was in line with Butcher (2021) explanation which said that a DE value that exceeds 50% indicates that the item works optimally at the level of difficulty and discriminating power. Therefore, although these two parameters produce slightly different numbers, they have similar interpretations.

In terms of scores, the three approaches used in this study had a high overall correlation and there was even a perfect correlation between eLOK and CTT scores. This had implications on the three scores that appear to only have little variations.

On the other hand, overall, the IRT approach had far different results from the UGM's eLOK and CTT approaches. This was probably due to relatively small number of subjects (< 100) that prevented optimal parameters from the use of IRT (Şahin & Anıl, 2017; Setiadi, 1997). Other than that, the use of this small subject size also affected the pattern of responses, as the IRT approach also depends on the response pattern of the test subjects (de Ayala, 2009). In other words, this significant difference in the results of the IRT analysis can be ignored due to the insufficient number of subjects.

Conclusion

From the results and discussion above, it can be concluded that the evaluation of tests using the existing approach at UGM's eLOK had close equivalence with the evaluation of psychometric properties using the CTT approach. This provided an argument that the use of UGM's eLOK for student examination testing can provide a trustworthy result of the psychometric properties' evaluation. In addition, the use of UGM's eLOK will produce stable or reliable scores that comparable to scoring using the CTT approach. The implication was that the results of this study can be used as a reference in using UGM's eLOK as a student academic testing platform where teachers have the opportunity to see the quality of the tests and the items they give to test takers.

However, this study also had some limitations. The number of participants in this study was limited, with only 22 active students who took the exam. This small number of participants can affect the results of the analysis in both the CTT and IRT approaches, even though IRT is theoretically sample independent. In addition, this study used data generated from an open book type of test so that the scores obtained were mixed with other abilities that were not relevant to the measured construct, such as reading speed, skimming ability, and others.

Recommendation

For further research, it is recommended to use more data (> 200 subjects) so that later the analysis can be more comprehensive and produce more robust estimation values. Comprehensive analysis can be done, for example, is a Confirmatory test Factor Analysis (CFA) as it is able to ensure validity evidence based on internal structure on the test (American Educational Research Association and American Psychological Association, 2014). In addition, considering the type of test that is open book, further research can conduct studies with closed-book tests or compare psychometric properties and test scores between open book and closed book exams.

Declaration

Acknowledge

Thanks to Haidar Buldan Thontowi, Ph.D for the support in this research by giving permission to access and evaluate the examination of multivariate statistic.

Conflict of Interest

We declared that there is no any conflict of interest that influences this study.

Authors' Contribution

RDM contributed in designing the study and conducting data analyzing also writing discussion section. HM contributed to conduct non-systemic literature review related to this study and writing the introduction section. SSK supervised the whole research process and designed an outline of this article.

Funding

This study was funded by all authors themselves.

Orcid ID

Ramadhan Dwi Marvianto https://orcid.org/0000-0001-8531-5519
Haryanti Mustika https://orcid.org/0000-0001-9091-5412
Sri Suning Kusumawardani https://orcid.org/0000-0003-1705-3232

References

Aiken, L. R. (1979). Relationships between the item difficulty and discrimination indexes. *Educational and Psychological Measurement*, *39*(1), 821–824. https://doi.org/10.1177/001316447903900415

Allen, M. J., & Yen, M. M. (1979). *Introduction to measurement theory*. Brooks/Cole Pub. Co.

American Educational Research Association and American Psychological Association. (2014). *National council on measurement in education*. Standards for educational; psychological testing. American Educational Research Association.

Azwar, S. (2013). *Penyusunan skala psikologi* (2nd). Pustaka Pelajar.

Azwar, S. (2016a). *Dasar-dasar psikometrika* (2nd). Pustaka Pelajar. https://doi.org/979-9075-73-4

Azwar, S. (2016b). *Konstruksi tes kemampuan kognitif* (1st). Pustaka Pelajar.

Azwar, S. (2017). *Reliabilitas dan validitas* (4th). Pustaka Pelajar.

Baker, F. B., & Kim, S.-H. (2017). The basics of item response theory using r. Interdisciplinary research; perspectives, 16(3). Springer. https://doi.org/10.1080/15366367.2018.1462078

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional irt1. In *Handbook of item response theory modeling: Applications to typical performance assessment (a volume in the multivariate applications series) (issue march* (pp. 307–333). Routledge.

Butcher, P. (2021). *Quiz report statistics*. https://docs.moodle.org/dev/Quiz_report_statistics

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publishers.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.

Ebel, L. (1965). *Measuring educational achievement*. Prentice Hall.

for Academic Innovation, C., & Studies. (2021). *Elok: Tentang kami*. https://elok.ugm.ac.id/mod/page/view.php?id=18

Furr, M. R., & Bacharach, V. R. (2013). *Psychometric: An introduction* (2nd). SAGE Publisher.

Hambleton, R. K. (1991). *Fundamentals of item response theory*. SAGE Publications.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353–383. https://doi.org/10.1177/0011000099273004

Hinkle, D., Wiersma, W., & Jurs, S. (2003). *Applied statistics for the behavioral sciences*. Houghton Mifflin.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometrics theory* (3rd). McGraw-Hill. https://doi.org/10.1007/978-1-4020-9173-5_8

Ramadhaningtyas, D. A. (2018). *Item response theory (irt) vs classical test theory*. Universitas Brawijaya.

Riswan, R. (2021). Sample size and test length for item parameter estimate and exam parameter estimate. *Al-Khwarizmi : Jurnal Pendidikan Matematika Dan Ilmu Pengetahuan Alam*, *9*(1), 69–78. https://doi.org/10.24256/jpmipa.v9i1.2384

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Schauberger, P., & Walker, A. (2021). *Openxlsx: Read, write and edit xlsx files*. https://cran.r-project.org/package=openxlsx

Setiadi, H. (1997). *Small sample irt item parameter estimates [dissertations]*. University of Massachusetts.

Sumardi. (2020). *Teknik pengukuran dan penilaian hasil belajar*. Deepublish Publisher.

Team, R. C. (2019). *R: A language and environment for statistical computing*. https://www.r-project.org/

Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files*. https://cran.r-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Willse, J. T. (2018). *Ctt: Classical test theory functions*. https://CRAN.R-project.org/package=CTT

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in r. In *Implementing reproducible computational research*. Chapman; Hall/CRC.