Studies in Learning and Teaching (SiLeT)
Homepage: https://scie-journal.com/index.php/SiLeT
Email: silet@scie-journal.com

p-ISSN: 2722-399X; e-ISSN: 2722-1857
SiLeT, Vol. 4, No. 2, August 2023
Page 329-340

Antithesis of Human Rater: Psychometric Responding to Shifts Competency Test Assessment Using Automation (AES System)

***M Idhom[1], I G P A Buditjahjanto[1], Munoto[1] , Trimono[2], P A Riyantoko[2]**
[1]Doctoral Program of Vocational Education, Universitas Negeri Surabaya, Indonesia
[2]Data Science Study Program, UPN "Veteran" Jawa Timur, Indonesia

Article Info

Article history:

Received June 12, 2023
Revised July 15, 2023
Accepted August 6, 2023
Available Online August 31, 2023

Keywords:

Automated essay scoring;
Competency;
Human rater;
Information and communication technology;
Vocational education;

ABSTRACT

This research is part of proof tests to a combination of statistical processing methods, collecting assessment rubrics in vocational education by comparing two systems, automated essay scoring and human rater. It aims to analyze the final assessment score of essays in *Akademi Komunitas Negeri (AKN) Pacitan* (Pacitan's State Community College) and *Akademi Komunitas Negeri* (AKN) *Blitar* (Blitar's State Community College) in East Java, Indonesia. The provisional assumption is that the results show an antithesis to the assessment of human feedback with an automated system due to the conversion of scores between the rubric and the algorithm design. As the hypothesis, algorithm-based score conversion affects automated essay scoring and human rater methods, which led to antithesis feedback. The validity and reliability of the measurement maintain the scoring consistency between the two methods and the accuracy of the answers. The novelty of this article is comparing between AES system and Human Rater using statistical methods. The research shows that there is a similar result using the psychometrics approach, which indicates different metaphor expressions and language systems. Thus, the objective of this study is to provide assistance in the advancement of an information technology system that utilizes a scoring mechanism merging computer and human evaluations, employing a psychological approach known as psychometric leads.

https://doi.org/10.46627/silet

INTRODUCTION

Interaction between human and computer carry one connecting relationship to develop two-way communication; human as a system user and subject of computer management system, also computer as the object. In general, the relationship between human and computer is relatively used in daily life to solve tasks effectively and efficiently. Computer also brings positive impacts to learning and assessment process in educational field. There is a common saying that human has the same intelligence level as with computer system advances. This perspective leads us to figure of metaphorical process in human cognition, as the findings about automatic metaphors of thought, language, and communication (Tong et al., 2021).

The development of computer system activation can be co-opted into an education system which frames learning and assessment problems, although one of the interpretations still comes from human cognition as knowledge, and humans are still users of the computer object itself. The application of computer systems assembly a resemblance to human knowledge, and is often used in learning assessment. For instance, the system only provides one-sided feedback in competency assessment, while humans allow more due to cognitive and psychological knowledge. However, competency assessment using humans (human rater) has currently undergone a shift towards

automation assessment which is believed to be more effective and efficient; technology as the basis for the development of intelligent management information systems (Safiullin et al., 2019).

This study reviews the interesting things about the development of information technology using Automated Essay Scoring (AES) in competency testing and assessment in vocational education. By comparing the results of the assessment tests on essay questions between automation and human raters, the problem frame can be seen whether there are significant differences, as the human rater can be labor-intensive and time-consuming (Zhang, 2013), whereas automation process brings otherwise results (Dong et al., 2017; Wong & Bong, 2019). By using the AES, statistical proof of the automation assessment results at the validity and reliability level becomes a benchmark, whether the consistency of the results supports the resulting accuracy value, or it still requires other proof from the human rater; the prediction results show the same/not much difference (Buditjahjanto et al., 2022). This phenomenon is still a topic of conversation in vocational education regarding the quality of outcomes of students at *Akademi Komunitas Negeri* (AKN) Pacitan and Blitar (research locus), because currently the competency test assessment still uses a human rater, and has an impact on recognizing ownership of a competency certificate as a competency in mastering knowledge, skills, and attitudes (Watkins, 2020).

The study between AES and human rater can be an appealing point to be discussed, in which the predicted value evaluation considers the effect of the actual value. Where, evaluation of essay answers is unique besides using multiple choice/short answers (Burrows et al., 2015). Interpretation in the descriptive analysis requires the conversion of automation scores into assessment rubrics. As an example, the automation assessment system examines pertinent knowledge mixed with the designed method, while human judgment is separated; the rubric answers a holistic critical interpretation of results and evaluation in decision-making (Facione, 2015; Nanni & Wilkinson, 2015). There are differences in conception conversion between AES systems and humans. The following is a descriptive analysis that describes the differences between the two (Table 1).

Table 1. Differences of conceptions of AES vs human rater

AES System	Human Rater
Using data	Using knowledge and psychology
Algorithmic	Heuristic in nature
Effectively manipulate the database	Effectively manipulate the knowledge base

Source: *Analysis of empirical and theoretical approaches*

According to (Wang, 2012), human rater obtains a bias effect on differences in the quality and consistency of essay scoring substantially. Therefore, a new paradigm of assessment raises another alternative using the AES system; automation using an algorithmic design (Puñal et al., 2014), aligns with the aim of examining the development of AES from the information perspective of user-centered computer science which has been introduced since 1999 (Navarro et al., 1999). The first hypothesis has occurred to test and analyze:

H1: Do the validity and reliability level have significant differences for using two assessment tools, AES and human rater?

Linguistics approach in assessing essay questions on competency tests focuses on information features or specific vocabulary to evaluate the performance of generic scores on differences in the results of automation scores with human raters. The correlation between the two systems requires an accuracy test in each of the available and different essay questions. To combine and consider features into essay scores requires a scoring consensus framework (Williamson et al., 2012); graphical and statistical evaluation of verbally constructed responses between the AES system and the human rater (Wang et al., 2018). Thus, the second hypothesis occurs:

H2: Do the score results in AES need to be compared with human rate assessment?

The relationship between automatic scores and human scores provides important information about validity as a default standard. However, considering the construction of assessments using the AES system found one condition for shifting human raters, a computer-based formative assessment (Mao et al., 2018); analysis and prediction standard for evaluation(Atteveldt et al., 2021); and using statistical methods to analyze data reliability and planning reliability tests (Meeker et al., 2022). This change will later be proven by the relationship between the two where the antithesis results of the score produced as a consistency of responses in competency testing with essay answers; technology with ideas (Almeida & Buzady, 2023). Theoretically, a human rater with the assessment rubric from the conversion results of the AES system explains the correlation of knowledge and psychology using a statistics approach; psychometric answers the level of validity, reliability, and accuracy of the score results; as a standard point of competency assessment (Collier-Sewell et al., 2023). On the other hand, the results of the calculations have a tendency and proving that the AES system is categorized as effective and efficient for competency assessments.

Based on the explanations in H_1 and H_2, the analysis of the disparity in the assessment of competency tests using the IT multimedia graphic design system with a Human Rater yielded similar results. This is despite the hypothesis in H_1 suggesting that AES is more likely to be valid, albeit not significantly. Consequently, the results obtained from both the Human Rater and AES indicate a psychological relationship between the IT system and humans. Furthermore, the assessment of competency tests using the IT multimedia graphic design system with a Human Rater shows comparable outcomes, despite the H1 hypothesis favoring the greater validity of AES, albeit insignificantly. Thus, this suggests a psychological connection with humans. The Human Rater assessment always involves an evaluation rubric with the aim of measuring competency test results based on the assessment interval range presented in Table 2. As a result, the H_0 is rejected because it can be inferred that the Human Rater assessment is influenced by the knowledge and development of the assessor. Therefore, this analysis of disparities provides psychometric leads that address the balance between the analysis performed by the Human Rater and AES.

THEORETICAL FRAMEWORK

Validity Test

Moses & Yamat (2021) argue that the validity test indicates how the tool is able to measure in certain extend. There are four types of validity test, which are face validity, content validity, criterion validity, and construct validity. Face validity illustrates how the face structure of a research instrument can be measured (Purnama, 2015). It refers to the shape and appearance of the instrument. Content validity is an ability to measure a particular concept or variable of the instrument. Criterion validity is to measure and correlate the instrument validity with other reliable and valid instruments. Construct validity refers to the ability of an instrument to measure. According to (Surucu & Maslakci, 2020), a research instrument can be considered valid if sig ≤ α.

Hypotheses:

H_0 : invalid items

H_1 : valid items

This is the formula that can be used to test construct validity by using Pearson's correlation method (Zhou et al., 2016):

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}} \quad (1)$$

Notes:

r : item correlation coefficient
n : total number of observations
X_i : item score
Y_i : total score of each variable
If sig. ≤ α, the value is null and invalid.

Reliability Test
A reliability test is conducted to test the consistency of measurement. Empirically, the consistency of reliability is indicated by the reliability coefficient (Heale & Twycross, 2015). The reliability coefficient technique used is the Cronbach alpha (α) reliability coefficient. The Cronbach alpha coefficient is greater than 0.7, so the questionnaire is considered reliable (Kennedy, 2022). However, if the value of the alpha variable is less than 0.7, then the questionnaire is considered less reliable.

$$\text{Cronbach Alpha} = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_{total}^2}\right) \quad (2)$$

Notes:
k : Numbers of questionnaire item in one dimension
σ_i^2 : Score variant of a question item
σ_{total}^2 : Total variance

Normal Multivariate Distribution
A population distribution can be considered normal if the data distribution is concentrated around the mean value symmetrically (Liang et al., 2022). A normal distribution that has a symmetrical shape, has the same mean, median, and mode so that the median value in the symmetric normal distribution is the average (Grover et al., 2014). Random variable $X = (X_1, X_2, \ldots, Xp)$ with vector mean $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]'$ and covariance matrix $\boldsymbol{\Sigma} > 0$ with a multivariate normal distribution of order p with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, if it has a density function multivariate normal as follows (Cassidy, 2016):

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad (3)$$

In which $-\infty < x_i < \infty$, $i = 1, 2, \ldots, p$, annotated by $X \sim$Normal ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$).

MAPE
According to (Abidin & Jaffar, 2014), MAPE (Mean Absolute Percentage Error) is a method commonly used to evaluate score forecasting by considering the magnitude of the actual score. The calculation of the MAPE value is as follows (Shekar & Dagnew, 2019):

$$\text{MAPE} = \left(\frac{1}{n}\times\sum_{p=1}^{n}\left|\frac{Y_p - F_p}{Y_p}\right|\right)\times 100\% \quad (4)$$

Notes:
Y_p : actual value on p time.
F_p : estimated value on p time
n : number of observations

Assessment Rubric

https://doi.org/10.46627/silet.v4i2.291

An assessment rubric is used as a tool/instrument for lecturers to set criteria for assignments. the score rubric approach can reduce the mean absolute error (Hasanah et al., 2019). Assessment rubric is a guide used to determine the human rater essay score. By using an assessment rubric, it is easier to determine the assessor's score and to overview the assessment of cognitive aspects and performance aspects as well; in addition to the need for consistency over the consequences of errors when assessing, for example fatigue (Haley et al., 2017). The following Table 2 assessment rubric scale is presented below.

Table 2. Scoring scale of assessment rubric

Score	Criterion
5 (excellent)	The percentage of correct answers is > 80%
4 (very good)	The percentage of correct answers is 60% - 79%
3 (good)	The percentage of correct answers is 40% - 59%
2 (average)	The percentage of correct answers is 20% - 39%
1 (poor)	The percentage of correct answers is < 19%

Source: processed data

RESEARCH METHOD

Research and Development is part of this research methodology. By using the Automated Essay Scoring (AES) approach, this can be divided into five (5) stages, namely, (1) preliminary investigation, (2) design, (3) realization/construction, (4) test, evaluation, and revision, and (5) implementation. Statistical analysis methods are used to test validity and reliability, and MAPE is used to test the accuracy of sample data (testing).

The sample data used were 38 respondents participating in the IT Multimedia Graphic Design competency test, and carried out in the data science laboratory for 2 days. Variable indicator measurements of competency unit titles and question materials included (1) K3- work safety, (2) software and hardware, and (3) creating, manipulating, and combining 2D and digital images (according to Competency Test Materials of vocational education at diploma degree). The technique of the data analysis of AES system uses the cosine similarity approach (training data), and text regression (data testing), while the human rater uses a scoring rubric scale, which is then converted to intervals/range of answer scores. The following are the conversion categories for the answer score ranges in Table 3 below.

Table 3. Range conversion of grading system

Similarity score	Human rater score	Grade
0.01-0.10	10	Poor
0.11-0.20	20	
0.21-0.30	30	
0.31-0.40	40	Average
0.41-0.50	50	
0.51-0.60	60	Good
0.61-0.70	70	
0.71-0.80	80	Very Good
0.81-0.90	90	Excellent
0.91-1.00	100	

RESULTS AND DISCUSSION

This study will demonstrate the proof of the human rater method as the antithesis of the AES system through validity and reliability tests. The data used for this research were the of 38 participants' competency test scores at the Pacitan and Blitar *Akademi Komunitas Negeri* (AKN) who were assessed using the AES system while still considering the human rater method of assessment. As for the antithesis, the human rater has an accuracy value which is slightly different from the AES system. The testing procedure will begin by analyzing the descriptive scores to obtain information about the characteristics of the test takers' scores. The results of the descriptive analysis are presented in Table 4 below.

Table 4. Descriptive analysis of score assessment

	Min	Max	Mean	St. dev	Skew	Kurt
AES	68.20	85.90	78.22	2.85	-0.15	0.86
HR	69.59	86.26	78.21	2.72	-0.07	0.21

Source: processed data

Based on the value of the descriptive analysis (Table 4), it is known that the AES system and the HR method have similar characteristics. Referring to the average value, AES and HR only have a difference in the value of 0.01. By statistics, there is no significant number difference. It can be concluded that the test score data has a relatively small standard deviation value (data range 0-100). Hence, it will occur, and no outlier values. The distribution of test score data is normally distributed on the average score (the AES and HR skewness values are not too far from 0). The distribution of data is presented through a histogram in Figure 1 and Figure 2 below.

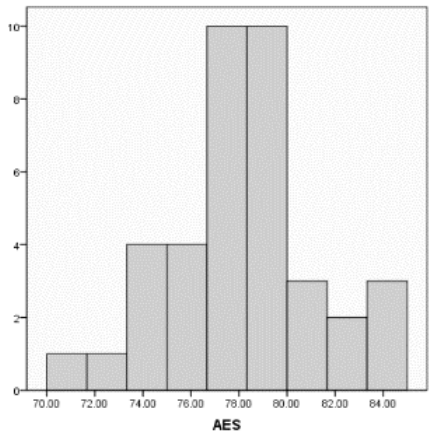

Figure 1. Histogram chart of AES score

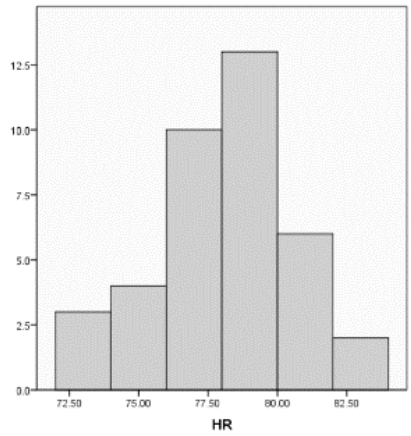

Figure 2. Histogram chart of Human Rater score

The figure above illustrates the scores of the AES system and the HR method have a normal distribution, with the average value being the center of the data distribution. Furthermore, the validity and reliability tests require that the sample data has a normal multivariate distribution (Yoo et al., 2019). It is considered normal if the Mahalanobis distance and the chi-square quintile distribute evenly around a straight line (Ghorbani, 2019). The following chart is the data plots as shown in Figure 3 below.

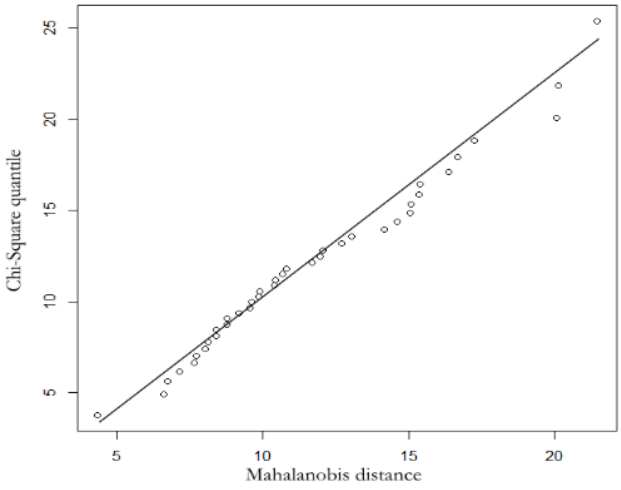

Figure 3. Scatterplot of Mahalanobis distance and Chi-square quantile

As referred to in the Figure 3 above, the data are spread around a straight line. This indicates that the assumption of multivariate normality has been met. The conclusion of the statistical test is considered weaker than the usual formal test. The formal test will be presented using the Kolmogorov-Smirnov test. The test results are presented in Table 5 below.

Table 5. K-S test for Normal Multivariate distribution test

Variable	*D*	p-value
AES and HR score	0.098	0.82

The hypothesis tested on the K-S test to check the normality of the data is:

H_0 : Distributed normal multivariate data

H_1 : Undistributed normal multivariate data

Significance level:

$\alpha = 5\%$

Statistic test:

D = 0.098

p-value = 0.82

Critical value:

H_0 is accepted if p-value < α

Decision:

Accept H_0 because p-value (0.82) > α (0.05) so it is concluded that the data is normally distributed multivariate.

The reason for proving the human rater method as the antithesis of the AES system is based on the results of validity and reliability tests. According to the Karl Pearson correlation value, the validity value of the AES system:

Table 6. Validity test result of AES system

Variable	Question	Corr (r_i)	p-value
K3	P1	0.90	0.00
	P2	0.91	0.00
	P3	0.91	0.00
	P4	0.93	0.00
Software and Hardware	P5	0.93	0.00
	P6	0.88	0.00
	P7	0.91	0.00
	P8	0.86	0.00
Create, manipulate, and combine 2D & digital images	P9	0.91	0.00
	P10	0.93	0.00
	P11	0.89	0.00
	P12	0.89	0.00

Source: processed data

As the Table 6 indicates, each question's correlation score is between 0.86 – 0.93. This score is included in the very strong correlation category. The validity of each answer score that is assessed using the AES system will be tested using the following hypotheses:

H_0 : invalid question instrument

H_1 : valid question instrument

Significant level:

$\alpha = 5\%$

Test statistics:

Test statistics using r_i and p-value obtained from Table 6

Critical value:

H_0 is rejected if p-value < α

Decision:

https://doi.org/10.46627/silet.v4i2.291

Based on Table 6, for each question indicator, the decision is rejecting H_0 and accepting H_1 because the p-value < a, so all question instruments assessed by the AES system are valid.

By using the same correlation measurement method as the AES system, the validity test results for competency test answer scores assessed using the human rater method are given in Table 7.

Table 7. Validity test result using Human Rater method

Variable	Question	Corr	p-value
K3	P1	0.89	0.00
	P2	0.89	0.00
	P3	0.91	0.00
	P4	0.88	0.00
Software and Hardware	P5	0.85	0.00
	P6	0.85	0.00
	P7	0.84	0.00
	P8	0.84	0.00
Create, manipulate, and combine 2D & digital images	P9	0.92	0.00
	P10	0.91	0.00
	P11	0.89	0.00
	P12	0.86	0.00

Based on Table 7, the correlation score for each question indicator is in the range of 0.84 – 0.91. This value is included in the very strong correlation category. The procedure for examining the hypothesis to measure the validity of the competency test score assessed using the human rater method is:

H_0 : invalid question instrument

H_1 : valid question instrument

Significance level:

$\alpha = 5\%$

Test statistics:

Test statistics using r_i and p-value obtained from Table 4

Critical value:

H_0 is rejected if p-value < α

Decision:

For each question indicator, the conclusions obtained are rejecting H_0 and accepting H_1 because the p-value < α, so all question instruments assessed by the human rater method are valid.

The hypothesis test concludes that the answer score of each question indicator assessed using the AES system and the human rater method is valid and has a very strong correlation value. Comparison of the average correlation value as a parameter of model validity is shown in Table 8.

Table 8. Comparison average of correlation value for AES and Human Rater method

Method	Average of correlation value
AES	0.90
Human Rater	0.88

Source: processed data

As we can see in Table 8, the average correlation value between the AES system and the Human Rater method has a 0.02 value difference. Statistically, considering the sample size and normality of the data, this difference is not significant (can be considered the same). This condition proves that the human rater is the right antithesis for the AES system.

Besides, the AES system and human rater method must also have good assessment consistency when used repeatedly. The reliability test to test the reliability of the two methods gives results as shown in Table 9.

Table 9. The result of the reliability test for AES and Human Rater

Method	Cronbach Alpha	Critical Value
AES	0.97	0.70
Human Rater	0.98	0.70

Source: processed data

Based on the Table 9 above, the reliability coefficient for the AES system and the human rater method is greater than the critical value (0.7) so it can be concluded that the two methods can be declared reliable (Surucu & Maslakci, 2020). This means that AES and human rater results are consistent when used to assess exam answers at different times.

Besides, prediction accuracy is an important indicator that must be met by an automated scoring system. The system is said to be accurate if the resulting score has an error close to 0. In this study, the human rater score acts as the actual value, and the AES system score acts as the predictive value. By using the MAPE method, the predicted accuracy value is included in the very accurate category (MAPE = 3.5%) (Nabillah & Ranggadara, 2020).

CONCLUSION

The shift in competency test assessment from the human rater to the AES system provides its own characteristics in the validity and reliability calculations which refer to the consideration of converting essay answer scores at *Akademi Komunitas Negeri* (AKN) Pacitan and Blitar. The consistency of the results of the human rater assessment score was able to show that the difference in results was not significant with the AES system. By using the assessment rubric, reviewing feedback on questions and answers as a basis of competence in general. This is caused by the correlation of knowledge with computer automation is biased to be used as a general competency assessment, yet can be used as a reference for mastery of cognitive knowledge. In spite of that, the AES system is more capable of giving a faster response (effective and efficient) compared to human raters. This category is not enough to prove that the results of the human rater's assessment are not representative in the field of cognitive knowledge alone, but psychology plays an important role in management decision-making in assessing competency tests for essay answers. To conclude this, the antithesis of the human rater was found when the role of psychology was combined with statistical test methods, namely psychometrics responding to shifts in the assessment of competence test for essay answers by using the AES system automation. Due to the relatively insignificant differences between the evaluations made by Human Raters and AES (Automated Essay Scoring), it is possible to conduct a rigorous analysis to provide empirical evidence for supporting or refuting existing theories. To ensure more precise outcomes, it is recommended to include supplementary trial data, competency assessment materials covering both skills and attitudes, as well as a wider range of educational levels beyond the scope of Diploma 1 and Diploma 2. This is highly intriguing, as the current outcomes in the form of psychometric leads can be stored as futuristic data. Therefore, for future research, the development of data psychology will be pursued, analyzing it through a mathematical approach that focuses on confusion matrix analysis.

ACKNOWLEDGEMENTS

This research is part of the dissertation and supported by the development of the AES system at *Akademi Komunitas Negeri* (AKN) Pacitan and Blitar with its own funding sources.

REFERENCES

Abidin, S. N. Z., & Jaffar, M. M. (2014). Forecasting share prices of small size companies in bursa Malaysia using geometric Brownian motion. *Applied Mathematics & Information Sciences*, *8*(1), 107–112. https://doi.org/10.12785/amis/080112

Almeida, F., & Buzady, Z. (2023). exploring the impact of a serious game in the academic success of entrepreneurship students. *Journal of Educational Technology Systems*, *51*(4), 436–454. https://doi.org/10.1177/00472395231153187

Atteveldt, W. v., Velden, M. A. C. G. v. D., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Buditjahjanto, I. G. P. A., Idhom, M., Munoto, M., & Samani, M. (2022). An automated essay scoring based on neural networks to predict and classify competence of examinees in community academy. *TEM Journal*, *11*(4), 1694–1701. https://doi.org/10.18421/TEM114-34

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Cassidy, D. T. (2016). A multivariate student's t-distribution. *Open Journal of Statistics*, *6*(3), 443–450. https://doi.org/10.4236/ojs.2016.63040

Collier-Sewell, F., Atherton, I., Mahoney, C., Kyle, R. G., Hughes, E., & Lasater, K. (2023). Competencies and standards in nurse education: The irresolvable tensions. *Nurse Education Today*, *125*, 105782. https://doi.org/10.1016/j.nedt.2023.105782

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 153–162. https://doi.org/10.18653/v1/K17-1017

Facione, P. A. (2015). *Critical thinking: What it is and why it counts*. Insight Assessement. https://www.insightassessment.com/wp-content/uploads/ia/pdf/whatwhy.pdf

Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis Series: Mathematics and Informatics*, *34*(3) 583-595. https://doi.org/10.22190/FUMI1903583G

Grover, G., Sabharwal, A., & Mittal, J. (2014). Application of multivariate and bivariate normal distributions to estimate duration of diabetes. *International Journal of Statistics and Applications*, *4*(1), 46–57.

Haley, B., Heo, S., Wright, P., Barone, C., Rao Rettiganti, M., & Anders, M. (2017). Relationships among active listening, self-awareness, empathy, and patient-centered care in associate and baccalaureate degree nursing students. *NursingPlus Open*, *3*(2017), 11–16. https://doi.org/10.1016/j.npls.2017.05.001

Hasanah, U., Permanasari, A. E., Kusumawardani, S. S., & Pribadi, F. S. (2019). A scoring rubric for automatic short answer grading system. *Telkomnika (Telecommunication Computing Electronics and Control)*, *17*(2), 763-770. https://doi.org/10.12928/telkomnika.v17i2.11785

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing*, *18*(3), 66–67. https://doi.org/10.1136/eb-2015-102129

Kennedy, I. (2022). Sample size determination in test-retest and cronbach alpha reliability estimates. *British Journal of Contemporary Education*, *2*(1), 17–29.

Liang, Y., Coelho, C. A., & von Rosen, T. (2022). Hypothesis testing in multivariate normal models with block circular covariance structures. *Biometrical Journal*, *64*(3), 557–576. https://doi.org/10.1002/bimj.202100023

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570

Meeker, W. Q., Escobar, L. A., & Pascual, F. G. (2022). *Statistical methods for reliability data.* John Wiley & Sons.

Moses, R. N., & Yamat, H. (2021). Testing the validity and reliability of a writing skill assessment. *International Journal of Academic Research in Business and Social Sciences*, *11*(4). https://doi.org/10.6007/IJARBSS/v11-i4/9028

Nabillah, I., & Ranggadara, I. (2020). Mean absolute percentage error untuk evaluasi hasil prediksi komoditas laut. *JOINS (Journal of Information System)*, *5*(2), 250–255. https://doi.org/10.33633/joins.v5i2.3900

Nanni, A. C., & Wilkinson, P. J. (2015). Assessment of ELLs' critical thinking using the holistic critical thinking scoring rubric. *Language Education in Asia*, *5*(2), 283–291. https://doi.org/10.5746/LEiA/14/V5/I2/A09/Nanni_Wilkinson

Navarro, G., Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. ACM Press.

Puñal, O., Aktaş, I., Schnelke, C. J., Abidin, G., Wehrle, K., & Gross, J. (2014). Machine learning-based jamming detection for IEEE 802.11: Design and experimental evaluation. *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*, 1–10. https://doi.org/10.1109/WoWMoM.2014.6918964

Purnama, I. A. (2015). Pengaruh skema kompensasi denda terhadap kinerja dengan risk preference sebagai variabel moderating. *Jurnal Nominal*, *4*(1), 129–145.

Safiullin, R., Marusin, A., Safiullin, R., & Ablyazov, T. (2019). Methodical approaches for creation of intelligent management information systems by means of energy resources of technical facilities. *E3S Web of Conferences*, *140*, 10008. https://doi.org/10.1051/e3sconf/201914010008

Shekar, B. H., & Dagnew, G. (2019). Grid search-based hyperparameter tuning and classification of microarray cancer data. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 1–8. https://doi.org/10.1109/ICACCP.2019.8882943

Surucu, L., & Maslakci, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal*, *8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Tong, D. H., Uyen, B. P., & Quoc, N. V. A. (2021). The improvement of 10th students' mathematical communication skills through learning ellipse topics. *Heliyon*, *7*(11), e08282. https://doi.org/10.1016/j.heliyon.2021.e08282

Wang, Z. (2012). Investigation of the effects of scoring designs and rater severity on students' ability estimation using different rater models. *Conference: 2012 Annual Meeting of the National Council on Measurement in Education*.

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101-120.

Watkins, S. C. (2020). Simulation-based training for assessment of competency, certification, and maintenance of certification. In J. T. Paige, S. C. Sonesh, D. D. Garbee, L. S. Bonanno (Eds.), *Comprehensive Healthcare Simulation: InterProfessional Team Training and Simulation* (pp. 225–245). Springer. https://doi.org/10.1007/978-3-030-28845-7_15

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wong, W. S., & Bong, C. H. (2019). A study for the development of automated essay scoring (AES) in Malaysian English test environment. *International Journal of Innovative Computing*, *9*(1). https://doi.org/10.11113/ijic.v9n1.220

Yoo, K., Rosenberg, M. D., Noble, S., Scheinost, D., Constable, R. T., & Chun, M. M. (2019). Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage*, *197*, 212–223. https://doi.org/10.1016/j.neuroimage.2019.04.060

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2), 1–11.

Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on pearson correlation coefficient. *Neurocomputing*, *216*, 208–215. https://doi.org/10.1016/j.neucom.2016.07.036

https://doi.org/10.46627/silet.v4i2.291

Author (s):

*Mohammad Idhom (Corresponding Author)
Doctoral Program of Vocational Education,
Universitas Negeri Surabaya,
Jl. Lidah Wetan, Surabaya 60213, Indonesia
Email: mohammadidhom.19009@mhs.unesa.ac.id

I Gusti Putu Asto Buditjahjanto
Doctoral Program of Vocational Education,
Universitas Negeri Surabaya,
Jl. Lidah Wetan, Surabaya 60213, Indonesia
Email: asto@unesa.ac.id

Munoto Munoto
Doctoral Program of Vocational Education,
Universitas Negeri Surabaya,
Jl. Lidah Wetan, Surabaya 60213, Indonesia
Email: munoto@unesa.ac.id

Trimono Trimono
Data Science Study Program,
UPN "Veteran" Jawa Timur,
Jl. Raya Rungkut Madya, Surabaya 60294, Indonesia
Email: trimono.stat@upnjatim.ac.id

Prismahardi Aji Riyantoko
Data Science Study Program,
UPN "Veteran" Jawa Timur,
Jl. Raya Rungkut Madya, Surabaya 60294, Indonesia
Email: prismahardi.aji.ds@upnjatim.ac.id