REID (Research and Evaluation in Education), 10(1), 2024, 50-63
Available online at: http://journal.uny.ac.id/index.php/reid

Construction of an instrument for evaluating the teaching process in higher education: Content and construct validity

Risky Setiawan*[1]; Wagiran[1]; Yasir Alsamiri[2]
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Islamic University of Madinah, Saudi Arabia
*Corresponding Author. E-mail: riskysetiawan@uny.ac.id

ARTICLE INFO

Article History
Submitted:
1 July 2023
Revised:
14 May 2024
Accepted:
14 June 2024

Keywords
teaching; validity; evaluation; construct; content

Scan Me:

ABSTRACT

This study aims to reveal the content validity, construct validity, and reliability of the instrument for evaluating the teaching process in higher education. This research is development research applying the ADDIE model from Molenda. The indicators evaluated consist of context, inputs, processes, and products. The sample consisted of 1200 students from eight faculties, each represented by three study programs. Data analysis uses three stages: content validity test analysis using the V-Aiken method involving six panellists or experts; construct validity test using Confirmatory Factor Analysis (CFA). Quantitative descriptive analysis and interpretive qualitative analysis used the Miles and Huberman method. The results showed that the developed evaluation instrument had good proof of the validity of the content, with an average V-Aiken score of 0.752, which was in the high category. Universitas Negeri Yogyakarta's evaluation instrument, which was developed through the instrument, already meets the validity of an exemplary construct of a good loading factor value (> 0.3). It has a composite reliability score above 0.7 and Cronbach's alpha above 0.6. The analysis results show that all empirical test criteria indicate the data is fit against the developed model.

To cite this article (in APA style):
Setiawan, R., Wagiran, W., & Alsamiri, Y. (2024). Construction of an instrument for evaluating the teaching process in higher education: Content and construct validity. *REID (Research and Evaluation in Education), 10*(1), 50-63. doi:https://doi.org/10.21831/reid.v10i1.63483

INTRODUCTION

The 21st century is making a big difference in the world of education. Today's technological developments have an impact on the world of education. Learning in the era of Industrial Revolution 4.0. requires this to be done by maximizing information and communication technology. The advantage of online learning is that it provides both teachers and students with flexibility and convenience. Meanwhile, the weakness is the infrastructure constraints both from students' geography and the devices used to support learning.

Evaluation is a process of collecting information comprehensively in deciding a particular policy (Ebert-May et al., 2011). Along with the rapid advancement of information and communication technology, distance education has also developed. Utilizing technology makes its reach wider, and its effectiveness in delivering learning materials is also increasing. The distance education system has also integrated various media types whose interactive capabilities are increasing. It is based on the separation between students and teachers in time and space, the use (package) of learning materials systematically designed and produced, the existence of non-continuous communication between students and students, tutors, and organisations through various media, as well as the existence of the intensive provision and monitoring of an educational organization.

The integration of digital skills in higher education is crucial for the future employability of graduates (Sokhanvar et al., 2021). E-learning can enhance the quality of learning and teaching, but it must meet certain requirements (Jara & Mellar, 2010; Wu & Lin, 2012). To manage the changes in learning and teaching, universities need to identify what needs to change and how to do it effectively (Kuo et al., 2017). The development of digital competencies is essential for students to succeed in the digital economy (Javaid et al., 2024).

Evaluation is the determination of the value of something, used in assessing the value of a program, product, process, or goal or the potential usefulness of alternative approaches to achieving a particular goal (Stufflebeam & Shinkfield, 2012). It is a process of collecting information to make a systematic decision. It is needed to interpret the collected data in the learning process. The data collected includes three aspects: planning, process, and evaluation. The teaching and learning at Universitas Negeri Yogyakarta (UNY) so far prioritizes aspects of character. Therefore, an objective and thorough evaluation process is needed. Evaluation is preceded by the process of preparing valid and reliable instruments. This study aims to develop trustworthy and reliable evaluation instruments based on the content's and construct's validity. The implementation of teaching with the blended learning method must be comprehensively evaluated (Divayana et al., 2017; Yangari & Inga, 2021; Zampirolli et al., 2018). The process of evaluating teaching processes in higher education is an integral part of its quality assessment (Qi et al., 2022).

Allen and Seaman (2010) research, which developed the system that was formed, was tested on a small group of Undiksha students majoring in informatics management. This online evaluation system has successfully developed the design and implementation. The functions that this system can accommodate are managing student data, managing exam question data, and setting some necessary settings. This online evaluation system is tested in small groups of students, and it is found that these online evaluations tend to be well received by students. Qomari (2015) developed an effective realm learning evaluation instrument. The development of this instrument tends to be more complicated than the instrument test (Qomari, 2015). Therefore, an in-depth study is needed to derive and elaborate the affective realm to certain aspects to develop valid and reliable instruments. The research conducted by Aprilia on the item trial (empirical) was analyzed using product moment, and it was obtained that there were several items/statements smaller than the table correlation coefficient for a 5% signification level, but after revision, all statement/items were valid (Aprilia, 2021). As for the results of reliability analysis using Alfa Cronbach, the calculated value of the reliability coefficient of 0.715 was obtained, which is classified as high reliability. For the quality of use/effectiveness of the instrument, obtained from 10 appraisers, six appraisers stated good, with a percentage of 60%, and four persons stated poor, with a percentage of 50%.

Evaluation provides information that can be used to determine the design, implementation, and impact for price and performance (value and benefit) of goals achieved, decision-making, accountability, and understanding of improvement phenomena. According to this formulation, the essence of evaluation is to provide information that can be used in exchange for decision-making. Mertens (2000) defines evaluation as the systematic application of social research procedures in assessing the conceptualization, design, implementation, and utility of social intervention programs. Evaluation is a systematic application of social research procedures in assessing program interventions' concept and design, implementation, and usefulness. In other words, evaluation research concerns the use of social research techniques to assess and improve the planning, monitoring, effectiveness, and efficiency of health services, education, welfare, and other program services (Aman et al., 2021).

According to Saaty (2007), measurement can be defined as assigning numbers to the individual or their characteristics according to specified rules, namely the quantification or determination of numbers about the characteristics or circumstances of individuals according to certain rules. This state of the individual can be cognitive, affective, and psychomotor abilities (Widoyoko, 2009). Assessment, according to TGAT (Mardapi, 2008), includes all the means used

to assess an individual or group's performance. The assessment process includes collecting evidence about learners' learning achievement.

Luo et al. (2023) and Murad et al. (2024) explain that measurement and evaluation are hierarchical. Measurement compares an observation to a reference. Evaluation describes and interprets measurement results. Thus, evaluation is a determination of the value or impact of an action. It can be the behavior of an individual or an institution. This hierarchical nature suggests that any evaluation activity involves measurement and assessment. There are three reasons for conducting an evaluation: (1) to decide and determine the organizer of the training by showing how the evaluation results can contribute to the goals and objectivity of the organization; (2) to decide whether the training program is continued or not; and (3) to obtain information on how to develop training programs in the future (Kirkpatrick, 1994). From the various understandings above, evaluation is a systematic action in assessing a program's concept, design, implementation, and usefulness to provide information that can be used as a consideration in making decisions. Program evaluation aims to collect information and determine the existence of deviations and shortcomings in a program by assessing the context, input, process, and outcomes. The primary purpose of preparing this paper is to develop and construct a valid and reliable instrument for evaluating the teaching process at UNY. A good instrument meets the criteria of validity and reliability (Fajardo et al., 2020; Pan et al., 2021). Thus, the developed instrument to evaluate the teaching process will be tested to prove the validity of the content and its construct.

METHOD

Research Design

This research is research and development (R&D) to develop an instrument for evaluating the teaching process at UNY. The model used is ADDIE with steps including (1) needs analysis with literature studies and field studies on the main needs of teaching evaluation at UNY, (2) designing instrument drafts through reference studies and previous research, (3) developing instrument by testing the model properness through content validity, (4) implementing results with construct validity tests, and (5) evaluating by looking at the measurement results and conducting in-depth interpretations. The research was carried out at UNY with as many as 1,200 students for a broad trial. As for content validation, it used six experienced people, three education experts, and three evaluation experts. The study was conducted from March to August 2022.

Sample and Data Collection

Data collection for the trial of 1,200 students at UNY was done with a sample from six faculties: education, economics, sports, mathematics and sciences, social science, engineering, and graduate school. Comprehensive data collection was conducted by linking the instrument to the UNY survey system so that the data obtained were representative and good.

Data Analysis

A descriptive analysis was conducted to provide an overview of data from the test subjects, consisting of 1,200 student respondents at UNY. The analysis used described the average data, standard deviation, and achievement percentage. Besides, the calculation for the content validity test used the V-Aiken formula. The formula used to prove the content validity is in Formula (1). It explains that the V-Aiken formula was used to see the construct quality of the instrument being developed (Anculle-Arauco et al., 2024), in which S = r -lo; Lo= lowest validity assessment number; C = highest validity assessment number; r = number divided by an expert.

$$v = \frac{\Sigma s}{[n(c-1)]} \quad \text{..... (1)}$$

Inference analysis was carried out through a construct validity test. The validation of the construct was carried out using the Confirmatory Factor Analysis (CFA) method, which calculates the degree of relationship of each item to the indicator using the second-order technique. The analysis was carried out using the smart computer program smart-PLS 3.0. The result of the construct testing showed that the minimum covariance value was 0.3.

FINDINGS AND DISCUSSION

Findings

The Construct of the Evaluation Instruments

The first stage in evaluating hybrid teaching was to construct the right instruments so that the data collected represents the results that follow the empirical data. First, at the stage of preparing the draft instrument, the researcher carried out the following: (1) identifying aspects and indicators of hybrid teaching evaluation through theoretical studies carried out, (2) compiling and constructing specifications and forms of instruments with the help of judgment as many as three measurement and evaluation experts, and (3) validating the instruments that had been made with Focused Group Discussion (FGD).

Testing the Content Validity of the Teaching Evaluation Instrument

Data collection was carried out from August to September 2022. The process began with instrument validation to see the validity of the contents. Validation was carried out using the FGD method, which presents six experts. The validation process produced (1) the observation instrument entirely used and (2) the test instrument from 64 items summarized to only 15 items. The simplification was based on input from five experts, resulting in a more streamlined instrument. The result of Aiken's validity analysis with four scales is presented in Table 1.

Table 1. Teaching Evaluation Instrument

No.	Aspects	Indicators	V-Aiken
1.	Planning	Lesson plan	0.72
2.	Planning	Meaningfulness	0.78
3.	Planning	Teaching materials	0.67
4.	Planning	Media	0.78
5.	Planning	Technology	0.78
6.	Process	Collapse	0.72
7.	Process	Motivation	0.78
8.	Process	Time	0.78
9.	Process	Response	0.67
10.	Process	Feedback	0.78
11.	Process	Mastery of the material	0.78
12.	Process	Clarity	0.72
13.	Evaluation	Assignment	0.78
14.	Evaluation	Test	0.78
15.	Evaluation	Concern	0.78
	Average		**0.75**

Unidimensional Tests

One-dimensional tests were performed by factor analysis using the SPSS 25 program. Before factor analysis was performed, a feasibility test was performed using the KMO-MSA test and Bartlett test for each instrument. The requirement for factor analysis is Kaiser-Meyer Olkin (KMO)–MSAU > 0.5, and the critical one-dimensional ballet test means that each test item measures only one of his abilities (Sánchez et al., 2016). For testing one-dimensional factor analysis, KMO and Bartlett's analyses gave results of less than 0.05. The KMO-MSA test was used to

check the validity of the samples, and the Bartlett test was used to check the normality of the data used. The result of the experiment is presented in Table 2.

Table 2. Kaiser-Meyer-Olkin Measure

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.951
Bartlett's Test of Sphericity	Approx. Chi-Square	1658 .283
	Df	105
	Sig.	.000

Table 2 shows the results of the empirical analysis, which are KMO-MSA values of 0.951 or greater than 0.5 and Bartlett's test signal of 0.000. From this, we can conclude that the analysis results are significant. In short, this tool is worthy of factor analysis. An extraction process produces several factors to obtain items with the exact dimensions. Each factor formed has an eigenvalue, and factors with eigenvalues greater than 1.00 are retained (Santoso, 2017).

Table 3. Eigen Values

Components	Eigen			Loadings			Rotation Loadings		
	Total	% Variance	%	Total	% of Variance	%	Total	% Variance	Cumulative %
1	8.77	58.49	58.49	8.77	58.49	58.49	8.63	57.54	57.54
2	1.05	7.00	65.49	1.05	7.00	65.49	1.19	7.95	65.49

Extraction Method: Principal Component Analysis.

The Variance in Table 3 is 57.54% in the first component and 65.49% in the second component. This means that one aspect of ability is dominant in the instrument. Then, Scree Plot exploratory factor analysis was done to see if any other factors could contribute to ability.

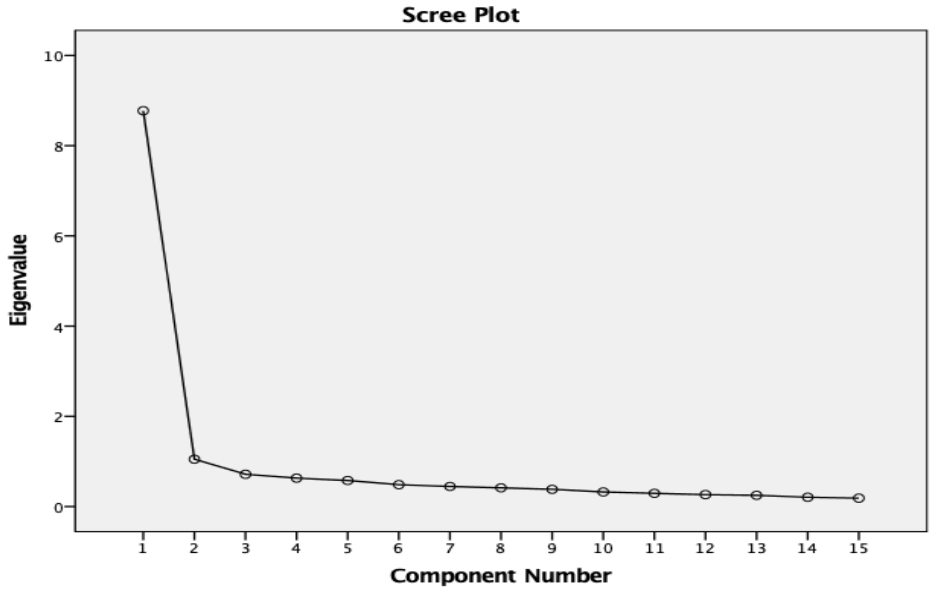

Figure 1. Screen plots Unidimensional Test Instrument

Figure 1 shows that the distance from Component 1 to Component 2 is very far or several times the distance between the other components. The steep scree plot indicates the presence of a dominant component, meaning that religious instruments measure only one factor or one dimension. In the evaluation instrument, the teaching model at UNY illustrates that the instrument measures only one dimension or aspect developed.

Reliability

The instrument's reliability level was assessed from the Cronbach Alpha coefficient value (Table 4). The instrument is reliable if Cronbach's Alpha value is more significant than 0.7 ($\alpha > 0.7$). Alpha coefficient results in a limited scale trial show a value of 0.865 (> 0.7), meaning that the instrument is very reliable.

Table 4. Reliability

Alpha	Alpha on Standardized Items	N of Items
.820	.945	15

Meanwhile, the reliability coefficient of each component obtained a value of more than 0.6, so it can be concluded that the reliability value of each component has a good category. The reliability value of each component is presented in Table 5.

Table 5. Reliability of Cronbach Instruments

Category	Cronbach's Alpha
Evaluation	0.860
Evaluation of Learning	0.945
Planning	0.857
Process	0.862

Reliability of Composite Scores

The composite reliability value or "Average Variance Extracted" (AVE) can tell us how reliable each latent variable is. This figure shows how the individual latent variables explain much variation in the indicator. The composite reliability of an instrument is a measure of how closely the different measurements of the instrument accord with one another. This is important because it means that the different indicators in the instrument are likely to be measuring the same thing. The formula for composite reliability is explained in Formula (2) (Li & Dolman, 2023).

$$\frac{(\Sigma\lambda)^2}{(\Sigma\lambda)^2 + \Sigma_i \operatorname{var}(\varepsilon_i)} \quad \text{..... (2)}$$

The reliability of the indicator is related to the component's reliability (see Table 6). The closer the parameter estimates are to being accurate, the more reliable the indicator will be.

Table 6. Composite Reliability Score for each Component

Category	Composite Reliability
Evaluation	0.914
Evaluation of Learning	0.953
Planning	0.897
Process	0.899

Validity of Teaching Evaluation Instrument Constructs

The first stage is to look at the main model of the instrument developed by proving the construct validity of the learning evaluation instrument. The developed instrument consists of 15 items with three aspects: planning, process, and evaluation. The following is an overview of the conceptual model of teaching evaluation instruments in higher education.

The next stage is to conduct a CFA analysis to see the magnitude of the factor loading in each component and item of the instrument developed (Table 7). The estimation results show that the teaching process evaluation instrument has a good reliability (above 0.7). Likewise, the

entire teaching monitoring and evaluation model amounts to 15 items and has a loading factor value of > 0.3, as many as 14 items. At the same time, one item has a loading factor below 0.3, so the item is not good or thrown away. The conceptual evaluation model in this study is shown in Figure 2.

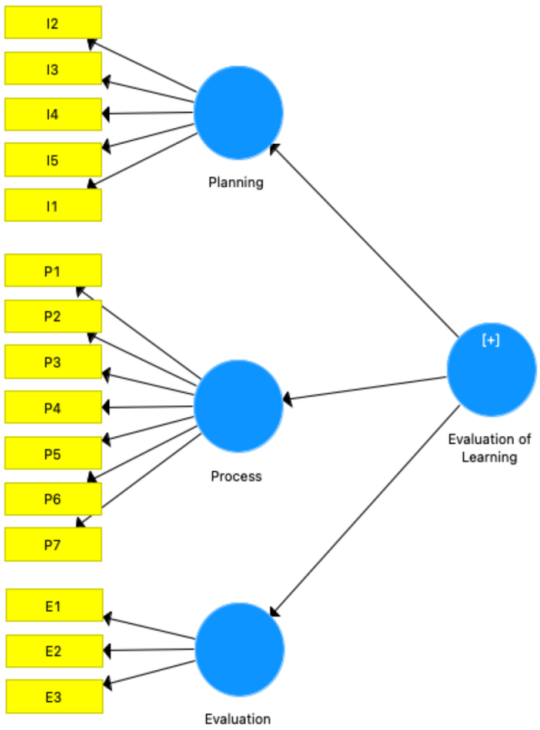

Figure 2. Conceptual Model of Evaluation

Table 7. Results of Factor Loading with CFA

No.	Category	Item
1.	Valid	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16
2.	Invalid	7

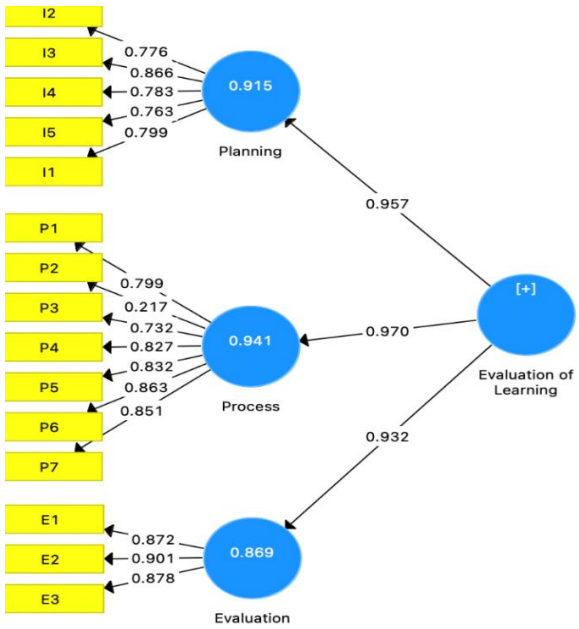

Figure 3. Path Diagram of Instrument Loading Factor

The initial analysis showed that out of 15 items, all met the loading factor requirements. Moreover, after modifications were obtained, fit models with a parsimony fit type. The results of the CFA analysis and factor loading recapitulation on the hybrid teaching evaluation instrument with the CIPP model can be seen in Figure 3.

Figure 3 shows the path graph result of the CFA analysis based on the structural model. Analysis of the structural model shows that all potential components or variables have high loading factors (0 > 0.3). The results summarizing the loading factor for the structural model are presented in Table 8. This is the same as Hayden et al. (2014) opinion about the validity criteria of instruments with a loading factor above 0.3 having a high level of covariance as well.

Walton et al. (2013) argue that the validity of constructs gives the idea that a good and valid instrument gives the meaning of the instrument worthy of use. Arslan et al. (2020) and Huda et al. (2022) suggest that the estimation results of the covariance on the path diagram provide a decision result that determines the degree of validity of the instrument (Setiawan et al., 2024). The value of a significant loading factor coefficient will provide a definition that the developed instrument has good construct validity (Reitz, 2014; Setiawan et al., 2020).

Table 8. Results of Factor Loading with CFA

No.	Component	Loading Factor	Category
1.	Planning	0.957	Excellent
2.	Process	0.970	Excellent
3.	Evaluation	0.932	Excellent

Table 8 shows that the loading factor in the path diagram shows the covariance between latent and observed variables with a coefficient above 0.5, meaning that all structural models analyzed through CFA fit empirical data.

Fit Model Proof

A step to demonstrate the fitness of the developed model was to examine the goodness-of-fit (GoF) test. It was used to validate the performance of a measurement model (external model) and a structural model (internal model). The value range is 0-1, and the interpretation is 0-0.25 (small GoF), 0.25-0.36 (moderate), and > 0.36 (large). Table 9 shows that the saturated models and estimated models have SRMR (square root) differences in terms of the tested data and the model.

Table 9. Model Fit Summary

	Saturated Model	Estimated Model
SRMR	0.095	0.096
d_ULS	4.185	4.242
d_G	N/A	N/A
Chi-Square	Infinite	Infinite
NFI	N/A	N/A

Discussion

After the V-Aiken analysis was conducted to test the validity of the contents, the results obtained showed that the highest V-Aiken value was 0.77, and the lowest value was 0.66. Thus, the value of the average V-Aiken coefficient of the entire instrument is 0.75, meaning it falls into the excellent category. This is relevant to the research results by Divayana et al. (2020). The validity of the content of the developed instrument provides clarity and strength to the instrument, as seen from its external appearance. Thus, the constructed outer appearance has the power that comes from a strong theory as well. According to Bliss et al. (2018), Boita et al. (2021), and Mensch et al. (2016), a good instrument gives small measurement error results. Thus, the data collected with valid and reliable instruments will be representative. Relevant to this,

Ward et al. (2015) provide an overview of the measurements made for instrument testing using the kappa coefficient method and interclass correlation.

According to Hambleton et al. (1991), the one-dimensional assumption is satisfied if the test contains a dominant factor measuring a person's ability. Sihombing et al. (2019) also writes the same thing, stating that if a measurement finds a dominant dimension, then that dominant dimension, on the particle's response or properties, will be 1 or 1 dimensional. Also, if the eigenvalues of the first factor have values up to multiples of the eigenvalues, then the second factor and so on are nearly identical. It can be said that it satisfies the one-dimensional condition (Cedzich et al., 2018).

The results of the composite reliability recapitulation in Table 6 show that all evaluation components, namely, contexts, inputs, processes, and products, have good internal consistency above 0.65. Then, it can be concluded that the reliability of composite hybrid teaching evaluation instruments is good, while the composite reliability of all devices is 0.865 or included in the good category. Remijn et al. (2014) state that the reliability of instruments estimated using diverse methods gives the meaning of the power of its consistency. Meanwhile, the reliability of psychological measurement instruments gives an idea of how high many people use the instrument's consistency (Faddar et al., 2017; Saito et al., 2016). Meanwhile, instrument tests with proof of construct validity reinforce the covariance of each item with its indicators (Marsh et al., 2011; Roldán-Merino et al., 2019).

The data show that the value between the data and the model is 0.001 apart, meaning there is very little difference between the data and the model. Thus, the data tested fit the model. The following criterion is d_ULS which is a measure that measures how strongly the empirical correlation matrix differs from the implied model correlation matrix. The output data show that the difference between the empirical matrix and the model is 0.057, so it can be concluded that the difference is very small, with a score above 2.00, which is very good. The analysis results show that all empirical test criteria indicate that the data fit against the developed model. Maynard et al. (2017) explain that planning for learning evaluation is of great strength. The small difference between saturated and estimated models indicates that the model fits the theory used (Huber-Carol et al., 2012). It can be concluded that a well-developed instrument model will provide validity and reliability values that fit the criteria (Setiawan, 2019).

The development of teaching process evaluation instruments in higher education has been shown to improve the quality of education (Pate et al., 2022). This can be further enhanced by updating methods and media to optimize innovation in teaching (del Rio et al., 2024). Quality management principles, such as focusing on customer needs and continuous improvement, can also be applied to improve the teaching process (Olayiwola et al., 2024). Effective evaluation methodologies are crucial in this process, particularly in the context of rapid technological advancements (Pandiyan et al., 2023).

CONCLUSION

The conclusions of the development of the instrument for evaluating the teaching process show that: (1) the monitoring and evaluation instruments developed have suitable proof of the validity of the content with an average V-Aiken score of 0.752, which is in the high category; (2) UNY's monitoring and evaluation instrument developed has met the validity of a good construct of a good loading factor value (> 0.3) and has composite score reliability above 0.7 and Cronbach alpha above 0.6; and (3) the analysis results show that all empirical test criteria show that the data fit against the developed instruments. The developed instruments provide a good contribution to the university in terms of representing the actual measurement results. The instrument's validity and reliability can be appropriately proven by conducting instrument trials.

Suggestions for the development of the construction of teaching monitoring and evaluation instruments are as follows: (1) the results of instrument development illustrate that good and valid instruments can be developed better in evaluating teaching in universities, especially at

UNY; (2) instrument development can be carried out by providing validity and reliability test results so that reliable instruments are reflected; and (3) the measurement and trial of instrument development results can be followed up by reviewing the instruments suitability and up-to-date curriculum in higher education. Finally, the advice for internal stakeholders is to adjust the indicators in the instrument to match the competence of graduates and the independent learning curriculum currently implemented in Indonesia.

The limitation of the research is that further analysis of the factors that affect lecturer performance has yet to be carried out. Moderate variables cannot be included to determine the influence between factors because the instruments only focus on performance results and teaching quality.

ACKNOWLEDGMENT

This research was funded by Universitas Negeri Yogyakarta from the Centre for Auditing, Monitoring, and Evaluation of Higher Education under the Institute for Quality Assurance and Education Development of UNY in 2022.

DISCLOSURE STATEMENT

The authors have no conflicts of interest to disclose.

REFERENCES

Allen, I. E., & Seaman, J. (2010). *Class differences: Online education in the United States, 2010*. Babson Survey Research Group.

Aman, A., Setiawan, R., Prasojo, L. D., & Mehta, K. (2021). Evaluation of hybrid learning in college using CIPP model. *Jurnal Penelitian dan Evaluasi Pendidikan*, *25*(2), 218–231. http://dx.doi.org/10.21831/pep.v25i2.46348

Anculle-Arauco, V., Krüger-Malpartida, H., Arevalo-Flores, M., Correa-Cedeño, L., Mass, R., Hoppe, W., & Pedraz-Petrozzi, B. (2024). Content validation using Aiken methodology through expert judgment of the first Spanish version of the Eppendorf Schizophrenia Inventory (ESI) in Peru: A brief qualitative report. *Spanish Journal of Psychiatry and Mental Health*, *17*(2), 110–113. https://doi.org/10.1016/j.rpsm.2022.11.004

Aprilia, N. H. (2021). *Perbedaan tingkat disiplin belajar siswa antara yang ikut dengan yang tidak ikut ekstrakurikuler paskibra pada siswa MAN 11 Jakarta Selatan DKI Jakarta*. Universitas Muhammadiyah Jakarta.

Arslan, S. S., Demir, N., & Karaduman, A. A. (2020). Turkish version of the Mastication Observation and Evaluation (MOE) instrument: A reliability and validity study in children. *Dysphagia*, *35*(2), 328–333. https://doi.org/10.1007/s00455-019-10035-8

Bliss, D. Z., Gurvich, O. V., Hurlow, J., Cefalu, J. E., Gannon, A., Wilhems, A., Wiltzen, K. R., Gannon, E., Lee, H., Borchert, K., & Trammel, S. H. (2018). Evaluation of validity and reliability of a revised Incontinence-Associated Skin Damage Severity Instrument (IASD.D.2) by 3 groups of nursing staff. *Journal of Wound, Ostomy & Continence Nursing*, *45*(5), 449–455. https://doi.org/10.1097/WON.0000000000000466

Boita, J., Bolejko, A., Zackrisson, S., Wallis, M. G., Ikeda, D. M., Van Ongeval, C., van Engen, R. E., Mackenzie, A., Tingberg, A., Bosmans, H., Pijnappel, R., Sechopoulos, I., & Broeders, M. (2021). Development and content validity evaluation of a candidate instrument to assess image quality in digital mammography: A mixed-method study. *European Journal of Radiology*, *134*, 109464. https://doi.org/10.1016/j.ejrad.2020.109464

Cedzich, C., Geib, T., Grünbaum, F. A., Stahl, C., Velázquez, L., Werner, A. H., & Werner, R. F. (2018). The topological classification of one-dimensional symmetric quantum walks. *Annales Henri Poincaré*, *19*(2), 325–383. https://doi.org/10.1007/s00023-017-0630-x

del Rio, A., Serrano, J., Jimenez, D., Contreras, L. M., & Alvarez, F. (2024). Multisite gaming streaming optimization over virtualized 5G environment using Deep Reinforcement Learning techniques. *Computer Networks*, *244*, 110334. https://doi.org/10.1016/j.comnet.2024.110334

Divayana, D. G. H., Sappaile, B. I., Pujawan, I. G. N., Dibia, I. K., Artaningsih, L., Sundayana, I. M., & Sugiharni, G. A. D. (2017). An evaluation of instructional process of expert system course program by using mobile technology-based CSE-UCLA Model. *International Journal of Interactive Mobile Technologies (IJIM)*, *11*(6), 18–31. https://doi.org/10.3991/ijim.v11i6.6697

Divayana, D. G. H., Suyasa, P. W. A., & Adiarta, A. (2020). Content validity determination of the countenance-tri kaya parisudha model evaluation instruments using lawshe's CVR formula. *Journal of Physics: Conference Series*, *1516*(1), 012047. https://doi.org/10.1088/1742-6596/1516/1/012047

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience*, *61*(7), 550–558. https://doi.org/10.1525/bio.2011.61.7.9

Faddar, J., Vanhoof, J., & Maeyer, S. De. (2017). School self-evaluation instruments and cognitive validity. Do items capture what they intend to? *School Effectiveness and School Improvement*, *28*(4), 608–628. https://doi.org/10.1080/09243453.2017.1360363

Fajardo, Z. I. E., Ramírez, R. A. N., & Álvarez, M. D. G. (2020). Instrumento alternativo para la evaluación del proceso enseñanza- aprendizaje en la educación básica general. *PUBLICACIONES*, *50*(2), 121–132. https://doi.org/10.30827/publicaciones.v50i2.13948

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publication.

Hayden, J., Keegan, M., Kardong-Edgren, S., & Smiley, R. A. (2014). Reliability and validity testing of the Creighton Competency Evaluation Instrument for use in the NCSBN National Simulation Study. *Nursing Education Perspectives*, *35*(4), 244–252. https://doi.org/10.5480/13-1130.1

Huber-Carol, C., Balakrishnan, N., Nikulin, M. S., & Mesbah, M. (2012). *Goodness-of-fit tests and model validity*. Springer Science & Business Media.

Huda, N., Sukarmin, Y., & Dimyati, D. (2022). Multiple Intelligence-based basketball performance assessment in high school. *Jurnal Penelitian dan Evaluasi Pendidikan*, *26*(2), 201–216. https://doi.org/10.21831/pep.v26i2.54981

Jara, M., & Mellar, H. (2010). Quality enhancement for e-learning courses: The role of student feedback. *Computers & Education*, *54*(3), 709–714. https://doi.org/10.1016/j.compedu.2009.10.016

Javaid, M., Haleem, A., Singh, R. P., & Sinha, A. K. (2024). Digital economy to improve the culture of industry 4.0: A study on features, implementation and challenges. *Green Technologies and Sustainability*, *2*(2), 100083. https://doi.org/10.1016/j.grets.2024.100083

Kirkpatrick, D. L. (1994). *Evaluating training program—The four levels*. Berrett-Koehler Publishers.

Kuo, Y.-C., Belland, B. R., & Kuo, T. T. (2017). Learning through blogging: Students' perspectives in collaborative blog-enhanced learning communities. *Educational Technology & Society, 20*(2), 37-50.

Li, L., & Dolman, A. J. (2023). On the reliability of composite analysis: an example of wet summers in North China. *Atmospheric Research*, *292*, 106881. https://doi.org/10.1016/j.atmosres.2023.106881

Luo, L., Ai, D., Qiao, H., Peng, C., Sun, C., Qi, Q., Jin, T., Zhou, M., & Xu, X. (2023). Evaluation of systematic frequency shift and uncertainty of an optical clock based on Bayesian hierarchical model. *Optics Communications*, *545*, 129745. https://doi.org/10.1016/j.optcom.2023.129745

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes* (A. Setiawan (ed.)). Parama Publisihing.

Marsh, H. W., Nagengast, B., Morin, A. J. S., Parada, R. H., Craven, R. G., & Hamilton, L. R. (2011). Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modeling. *Journal of Educational Psychology*, *103*(3), 701–732. https://doi.org/10.1037/a0024122

Maynard, B. R., Solis, M. R., Miller, V. L., & Brendel, K. E. (2017). Mindfulness-based interventions for improving cognition, academic achievement, behavior, and socioemotional functioning of primary and secondary school students. *Campbell Systematic Reviews*, *13*(1), 1–144. https://doi.org/10.4073/csr.2017.5

Mensch, S. M., Echteld, M. A., Evenhuis, H. M., & Rameckers, E. A. A. (2016). Construct validity and responsiveness of Movakic: An instrument for the evaluation of motor abilities in children with severe multiple disabilities. *Research in Developmental Disabilities*, *59*, 194–201. https://doi.org/10.1016/j.ridd.2016.08.012

Mertens, D. M. (2000). Institutionalizing evaluation in the United States of America. In R. Stockmann (Ed.), *Evaluationsforschung* (pp. 41–56). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-92229-8_2

Murad, M. H., Chu, H., Wang, Z., & Lin, L. (2024). Hierarchical models that address measurement error are needed to evaluate the correlation between treatment effect and control group event rate. *Journal of Clinical Epidemiology*, *170,* 111327. https://doi.org/10.1016/j.jclinepi.2024.111327

Olayiwola, R. K., Tuomi, V., Strid, J., & Nahan-Suomela, R. (2024). Impact of Total quality management on cleaning companies in Finland: A Focus on organisational performance and customer satisfaction. *Cleaner Logistics and Supply Chain*, *10*. 100139. https://doi.org/10.1016/j.clscn.2024.100139

Pan, G., Shankararaman, V., Koh, K., & Gan, S. (2021). Students' evaluation of teaching in the project-based learning programme: An instrument and a development process. *The International Journal of Management Education*, *19*(2), 100501. https://doi.org/10.1016/j.ijme.2021.100501

Pandiyan, P., Saravanan, S., Usha, K., Kannadasan, R., Alsharif, M. H., & Kim, M.-K. (2023). Technological advancements toward smart energy management in smart cities. *Energy Reports*, *10*. 648–677. https://doi.org/10.1016/j.egyr.2023.07.021

Pate, K., Powers, K., Coffman, M. J., & Morton, S. (2022). Improving self-efficacy of patients with a new ostomy with written education materials: A quality improvement project. *Journal of PeriAnesthesia Nursing*, *37*(5), 620–625. https://doi.org/10.1016/j.jopan.2021.11.020

Qi, S., Liu, L., Kumar, B. S., & Prathik, A. (2022). An English teaching quality evaluation model based on Gaussian process machine learning. *Expert Systems*, *39*(6), e12861. https://doi.org/10.1111/exsy.12861

Qomari, R. (2015). Pengembangan instrumen evaluasi domain afektif. *INSANIA : Jurnal Pemikiran Alternatif Kependidikan*, *13*(1), 87–109. https://ejournal.uinsaizu.ac.id/index.php/insania/article/view/287

Reitz, O. E. (2014). The job embeddedness instrument: An evaluation of validity and reliability. *Geriatric Nursing*, *35*(5), 351–356. https://doi.org/10.1016/j.gerinurse.2014.04.011

Remijn, L., Speyer, R., Groen, B. E., van Limbeek, J., & Nijhuis-van der Sanden, M. W. G. (2014). Validity and reliability of the Mastication Observation and Evaluation (MOE) instrument. *Research in Developmental Disabilities*, *35*(7), 1551–1561. https://doi.org/10.1016/j.ridd.2014.03.035

Roldán-Merino, J., Farrés-Tarafa, M., Estrada-Masllorens, J. M., Hurtado-Pardos, B., Miguel-Ruiz, D., Nebot-Bergua, C., Insa-Calderon, E., Grané-Mascarell, N., Bande-Julian, D., Falcó-Pergueroles, A. M., Lluch-Canut, M.-T., & Casas, I. (2019). Reliability and validity study of the Spanish adaptation of the "Creighton Simulation Evaluation Instrument (C-SEI)." *Nurse Education in Practice*, *35*, 14–20. https://doi.org/10.1016/j.nepr.2018.12.007

Saaty, T. L. (2007). The analytic hierarchy and analytic network measurement processes: Applications to decisions under risk. *European Journal of Pure and Applied Mathematics*, *1*(1), 122–196. https://doi.org/10.29020/nybg.ejpam.v1i1.6

Saito, T., Izawa, K. P., Omori, Y., & Watanabe, S. (2016). Functional independence and difficulty scale: Instrument development and validity evaluation. *Geriatrics & Gerontology International*, *16*(10), 1127–1137. https://doi.org/10.1111/ggi.12605

Sánchez, D., Chala, A., Alvarez, A., Payan, C., Mendoza, T., Cleeland, C., & Sanabria, A. (2016). Psychometric validation of the M. D. Anderson Symptom Inventory–Head and neck module in the Spanish Language. *Journal of Pain and Symptom Management*, *51*(6), 1055–1061. https://doi.org/10.1016/j.jpainsymman.2015.12.320

Santoso, S. (2017). *Menguasai statistik dengan SPSS 24*. Elex Media Komputindo.

Setiawan, R. (2019). A comparison of score equating conducted using haebara and stocking lord method for polytomous. *European Journal of Educational Research, 8*(4), 1071-1079. https://doi.org/10.12973/eu-jer.8.4.1071

Setiawan, R., Hadi, S., & Aman, A. (2024). Psychometric properties of learning environment diagnostics instrument. *Journal of Education and Learning (EduLearn)*, *18*(3), 690–698. https://doi.org/10.11591/edulearn.v18i3.21310

Setiawan, R., Mardapi, D., Aman, A., & Budi, U. (2020). Multiple intelligences-based creative curriculum: The best practice. *European Journal of Educational Research*, *9*(2), 611–627. https://doi.org/10.12973/eu-jer.9.2.611

Sihombing, R. U., Naga, D. S., & Rahayu, W. (2019). A Rasch model measurement analysis on science literacy test of Indonesian students: Smart way to improve the learning assessment. *Indonesian Journal of Educational Review*, *6*(1), 44–55. https://journal.unj.ac.id/unj/index.php/ijer/article/view/14071

Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, *70*. 101030. https://doi.org/10.1016/j.stueduc.2021.101030

Stufflebeam, D. L., & Shinkfield, A. J. (2012). *Systematic evaluation: A self-instructional guide to theory and practice*. Springer Science & Business Media.

Walton, M. B., Cowderoy, E., Lascelles, D., & Innes, J. F. (2013). Evaluation of construct and criterion validity for the 'Liverpool Osteoarthritis in Dogs' (LOAD) clinical metrology instrument and comparison to two other instruments. *PLoS ONE*, *8*(3), e58125. https://doi.org/10.1371/journal.pone.0058125

Ward, D. S., Mazzucca, S., McWilliams, C., & Hales, D. (2015). Use of the environment and policy Evaluation and Observation as a Self-Report Instrument (EPAO-SR) to measure nutrition and physical activity environments in child care settings: validity and reliability evidence. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 124. https://doi.org/10.1186/s12966-015-0287-0

Widoyoko, S. E. P. (2009). *Evaluasi program pembelajaran (instructional program evaluation)*. Pustaka Pelajar.

Wu, H.-Y., & Lin, H.-Y. (2012). A hybrid approach to develop an analytical model for enhancing the service quality of e-learning. *Computers & Education*, *58*(4), 1318–1338. https://doi.org/10.1016/j.compedu.2011.12.025

Yangari, M., & Inga, E. (2021). Educational innovation in the evaluation processes within the flipped and blended learning models. *Education Sciences*, *11*(9), 487. https://doi.org/10.3390/educsci11090487

Zampirolli, F. A., Goya, D., Pimentel, E. P., & Kobayashi, G. (2018). Evaluation process for an introductory programming course using blended learning in engineering education. *Computer Applications in Engineering Education*, *26*(6), 2210–2222. https://doi.org/10.1002/cae.22029