Jurnal Penelitian dan Pengkajian Ilmu Pendidikan: e-Saintika
https://journal-center.litpam.com/index.php/e-Saintika/index

July 2023 Vol. 7, No. 2
e-ISSN: 2615-6881
pp. 144-159

Quality of Performance Assessment Instruments for Educators in Higher Education: Implementation of Factor Analysis And Generalizability Theory

***[1] Hasan Djidu, [2] Edi Istiyono, [2] Widihastuti**

[1] Faculty of Teachers Training and Education, Universitas Sembilanbelas November Kolaka, Jl. Pemuda, Tahoa, Kolaka, Sulawesi Tenggara 93561, Indonesia
[2] Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Yogyakarta 55281, Indonesia

* Corresponding Author e-mail: hasandjidu@gmail.com

Received: June 2022; Revised: June 2023; Published: July 2023

Abstract

Assessing the quality of learning in higher education is one of the efforts to ensure its standards. Typically, the assessment of the quality of learning implementation involves observation by multiple raters. This study aims to provide construct validity evidence and estimate the reliability of performance assessment instruments for educators in higher education. 225 second-year and third-year students from the Faculty of Education participated as raters, evaluating the performance of educators in their teaching practices. Forty assessment items were used to evaluate the performance of 19 instructors. Exploratory Factor Analysis (EFA) and Generalizability Theory (G-Theory) were employed to examine the quality of the performance assessment instruments. The EFA analysis resulted in the identification of five factors that contribute to educators' performance in teaching: (1) readiness and planning, (2) pedagogy and professionalism, (3) personality, (4) social relationships within the classroom, and (5) social relationships beyond the classroom, collectively explaining 67.671% of the variance. Of the 40 assessment items, 37 demonstrated construct validity, while three required revisions. These findings indicate the alignment between the instrument's factors and the formulated theory of teaching competence. The reliability of the measurements was estimated using G-Theory in RStudio, yielding a relative G coefficient of 0.88 for three raters. The D-Study results indicated that the instrument could be used to assess performance, with an estimated generalizability coefficient of 0.738, requiring a minimum of five raters for each person (educator) being evaluated. We recommend employing G-Study and D-Study to determine the number of raters involved in performance assessment as a means of cost and time efficiency in the evaluation process.

Keywords: Performance assessment, Higher education, Exploratory factor analysis, Generalizability theory

How to Cite: Djidu, H., Istiyono, E., & Widihastuti, W. (2023). Quality of Performance Assessment Instruments for Educators in Higher Education: Implementation of Factor Analysis And Generalizability Theory. *Jurnal Penelitian Dan Pengkajian Ilmu Pendidikan: E-Saintika*, *7*(2), 144-159. https://doi.org/10.36312/esaintika.v7i2.716

https://doi.org/10.36312/esaintika.v7i2.716

INTRODUCTION

There is a significant body of literature focusing on the assessment of educational quality. One of the key factors that greatly influence the quality of education is the performance of educators, whether teachers in schools or lecturers in higher education (Rahardja et al., 2020). Many researchers have highlighted the crucial role of educators in shaping the overall learning experience and educational outcomes (Gil-Flores et al.,

2017; Hu et al., 2017; Retnawati et al., 2019; Stylianides, 2007). The recognition of this role has spurred numerous studies exploring various aspects such as educators' knowledge (Retnawati et al., 2018), experiences (Jeffrey et al., 2014; Sulistiyo et al., 2017), challenges they face (Retnawati et al., 2016), preparedness (Baya'a & Daher, 2013; Zurqoni et al., 2020), perceptions (Safi'i et al., 2019), and educational policies that affect educators' performance (Hermanu et al., 2022).

In higher education, evaluating educators' teaching performance is an essential aspect of ensuring quality. Different institutions adopt various models and strategies for conducting these assessments. The assessment methods employed also vary, ranging from traditional paper-and-pencil-based assessments (Taufiq, 2015) to online assessments utilizing tools like Google Forms (Batubara, 2016) or custom-built assessment systems integrated with the institution's information infrastructure (Ikram et al., 2018). Each institution chooses the most suitable assessment approach based on its needs and resources.

Researchers have extensively explored the development and utilization of performance assessment instruments for evaluating educators in teaching practices. These studies have been reported for over a decade, with some even dating back to the late 20th Century, highlighting the importance of assessing the quality of teaching performance (Van Tassel-Baska et al., 2006; Henkel, 1997; Martin & Martin, 1989; Poole et al., 1998). However, assessing teaching quality and educator performance remains a prominent issue in education. As the field of education acknowledges the ever-changing landscape of society, it recognizes the need to adapt to new challenges and advancements, such as the digitalization of learning (Djidu & Retnawati, 2022) and the disruptive impact of events like the pandemic (Djidu et al., 2021). Consequently, there is a continuous transformation in the understanding of educator performance to effectively support learners across different educational levels, including primary, secondary, and higher education, in navigating the demands of the modern era (Poole et al., 1998). This is evident in the growing body of research that focuses on the development and validation of observation protocols aimed at assessing the quality of teaching practices (Johnson et al., 2022; Johnson, Crawford et al., 2020; Johnson, Zheng et al., 2020; Noben et al., 2022; Rodgers et al., 2022).

Observation instruments play a crucial role in examining the quality and performance of teaching practices carried out by educators. They serve as valuable tools for educators to navigate the complex instruction landscape. Klette and Blikstad-Balas (2018) highlight the significance of observation instruments as guiding compasses for educators, helping them steer the course of their teaching endeavours. While numerous studies have reported assessment results on teaching quality, only a limited number have focused on the assessment instruments' quality. A literature review by Rodgers et al. (2022) identified 102 studies conducted between 1975 and 2020 that employed observation scales to assess teaching practices. However, it was found that only a small fraction of these studies provided evidence of the validity of the instruments used (p. 419). This implies that users interested in adopting such instruments should first explore the quality and validity of the measurement tools before employing them for their intended purposes.

Institutions often face the challenge of determining the number of raters required to assess the quality of teaching practices. This issue involves considerations of cost and time associated with evaluating educator performance. Evaluating the performance of educators in teaching practices typically involves gathering feedback

from all students, requiring each student to assess all educators. However, it is possible to obtain accurate assessment results without involving every student in the evaluation process. To implement this approach, it is necessary to conduct trials and analyze the quality of the assessment instrument using G-Theory and D-Theory. This analysis helps determine the minimum number of raters needed to achieve reliable results (Bimpeh et al., 2020; Brennan, 2001, 2011).

Based on these considerations, this study aims to validate the performance assessment instrument for educators in teaching practices at a higher education institution in Southeast Sulawesi and estimate its reliability. The instrument's construct validity will be demonstrated through exploratory factor analysis (EFA), while the generalizability theory (G-Theory) will be used to estimate its reliability.

METHOD

This paper is part of ongoing research to develop an instrument for assessing the performance of educators in delivering instruction at the higher education. The development of this assessment instrument is a part of the internal quality assurance system in higher education institutions, which mandates evaluation activities across various aspects of the institution's mission. The development process of the performance assessment instrument followed 15 stages, aligned with the stages of developing non-test instruments (Gable & Wolf, 1993; McCoach et al., 2013). These stages include: (1) developing conceptual definitions; (2) formulating operational definitions; (3) develop blueprint and items; (4) expert judgement; (5) determining response formats; (6) crafting completion instructions; (7) preparing a draft instrument for readability testing; (8) finalizing the draft instrument; (9) conducting an initial pilot test; (10) analyzing the pilot test results; (11) revising the instrument; (12) conducting a large-scale test; (13) preparing the final instrument; (14) establishing validity and reliability; and (15) creating the assessment implementation manual. Figure 1 provides a concise overview of the stages involved in the instrument development process.

In this study, we involved 225 undergraduate students from a state university in Southeast Sulawesi as active participants who utilized the developed instrument to assess educators' performance in delivering instruction. These students, who were in their second and third years at the Faculty of Education, had completed relevant courses covering the fundamentals of instructional theory, educational psychology, teacher competencies, and the teaching profession. To establish the instrument's construct validity, we analyzed the assessment results of the 225 students using EFA. Additionally, we randomly selected 19 educators to be evaluated by three raters each.

Five experts specializing in education and learning actively assessed this instrument's quality. The findings revealed that the instrument demonstrates strong content validity. This paper primarily focus on discussing the evidence of construct validity and score dependability of the instrument. Construct validity evidence was provided through the utilization of Exploratory Factor Analysis (EFA), while the reliability of scores was estimated using G-Theory. The application of G-Theory in developing this instrument involved incorporating the components of individuals, raters, and items as sources of variance that influence assessment scores. Unlike approaches such as internal consistency, parallel tests, or test-retest methods, G-Theory estimates reliability by considering additional factors that affect score variability (Brennan, 2011). Within this context, the variability of the true score, which

reflects educators' performance in instructional delivery, is influenced by both raters and items. In other words, rater subjectivity can impact the scores assigned, making it crucial to consider when estimating reliability. Estimating reliability using internal consistency, test-retest, or parallel forms approaches is not feasible when dealing with scores provided by multiple raters, as each individual will have multiple scores corresponding to each rater (e.g., if there are ten raters, each individual will have ten scores). Conversely, G-Theory allows for considering the entire variability of scores stemming from raters, items, and other factors in the estimation process, without aggregating scores from all raters beforehand.

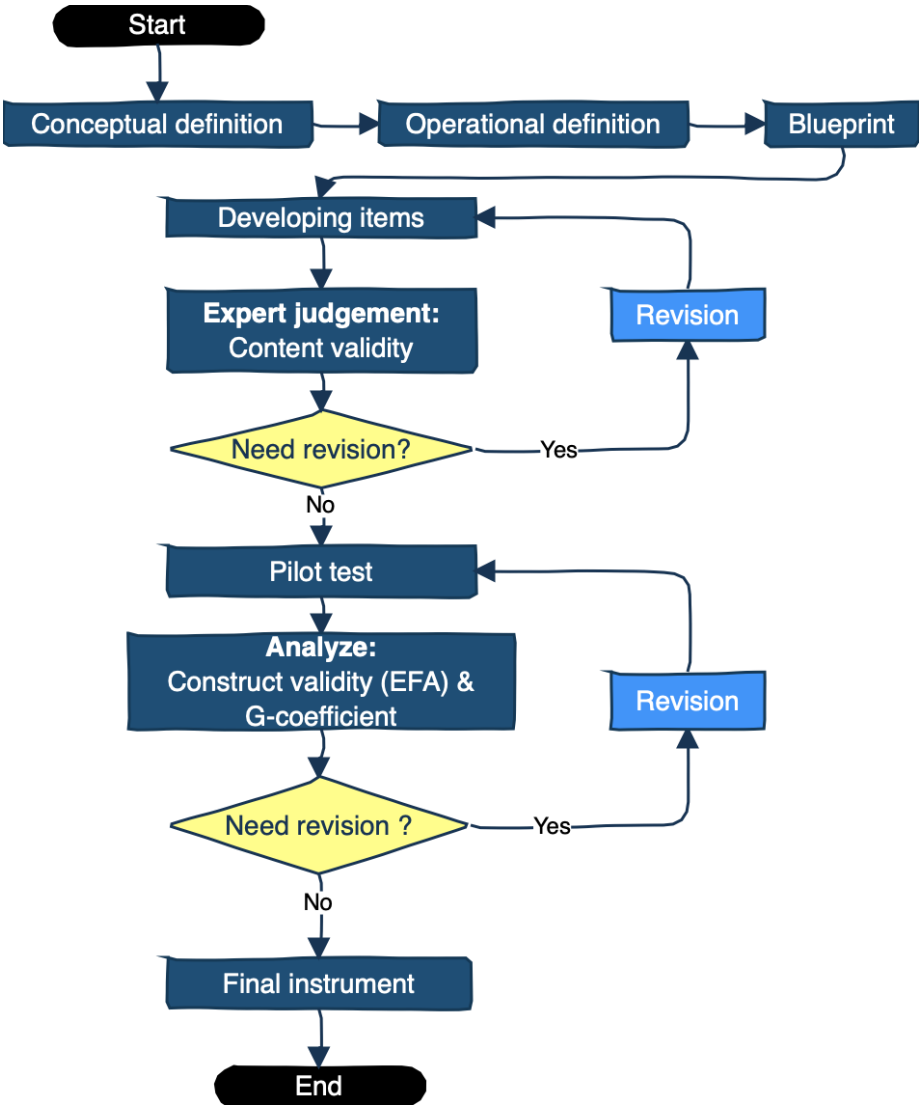

Figure 1. Development Stages

Using a nested design, we applied G-Theory (G-Study) followed by D-Theory (D-Study). Data analysis was performed using the 'gstudy' and 'dstudy' functions in the 'gtheory' package in RStudio (Moore, 2016; R Core Team, 2022). This paper discusses the results of the performance assessment instrument's pilot data analysis using EFA and G-Theory, which are crucial stages before determining the instrument's suitability. We present all the syntax for the analysis processes in text format to facilitate replication or reuse by readers who wish to estimate G-coefficients using RStudio.

Table 1. Multifacets Universes of Admissible

NPO	Design	Main Effect	Interaction Effect
1	P x I x R	P, I, R	PI, PR, IR, PIR
2	P (I x R)	P, I, (I:R)	PR, P(I:R)
3	P(I : R)	P,I:P(I:R)	
4	P: (I:R)	P,I:P, R:P	

The instrument is considered good construct validity if the EFA results indicate that the developed items measure the intended constructs. In this context, the instrument is expected to measure components related to educators' performance in instructional implementation. To evaluate the instrument's reliability, G-Study and D-Study are employed, ensuring that the estimated coefficients meet the minimum threshold of >0.70 (Allen & Yen, 1979; Brookhart & McMillan, 2020; Ebel & Frisbie, 1991; Miller et al., 2009). The G-Theory design utilized is the multi facets universes of admissible observation, where interactions among Person (P), Item (I), and Rater (R) are considered. The specific interaction design is presented in Table 1.

We analyze two components in the G-Study and D-Study. The G-Study estimates a set of component variances, with the number of components determined by the chosen model. In contrast, the D-Study builds upon the results obtained from the previous G-Study. It estimates, utilizes, and interprets the magnitudes of the variances needed for decision-making (Brennan, 2011). It is crucial to pay close attention to the specifications of the generalization universe in the D-Study. The outcomes of the D-Study provide insights into the generalizability of the study's findings within a particular measurement procedure.

RESULTS AND DISCUSSION

The instrument for evaluating the performance of educators in higher education has been developed, focusing on five indicators. Each indicator is assessed using a minimum of six items. These indicators encompass the pedagogical, professional, personality, and social aspects in alignment with the required competencies.

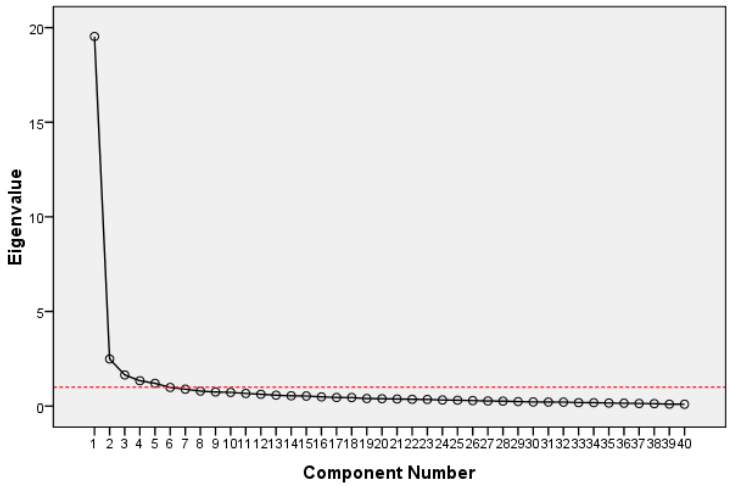

Figure 2. The Scree Plot displays the number of factors

The analysis shown in Figure 2 reveals that the instrument measures five prominent factors. A detailed breakdown of these findings is presented in Table 2. The exploratory factor analysis results provide valuable insights, indicating that the

assessment instrument encompasses four factors contributing to measuring educators' performance. Compared to the competencies outlined in the Indonesian Law Number 14 of 2005 on Teachers and Lecturers, which encompass pedagogical, professional, personality, and social competencies, the EFA results demonstrate the alignment of the instrument's items with these competencies. However, it should be noted that not all competencies mentioned in the Law on Teachers and Lecturers align perfectly with the grouping of items derived from the factor analysis.

Table 2. Rotated Component Matrix (EFA)

Items	Components				
	1	2	3	4	5
B01	0.275	**0.724**	0.218	0.1	0.188
B02	0.292	**0.73**	0.161	0.041	0.168
B03	0.198	**0.651**	0.057	0.165	0.282
B08	0.453	**0.499**	0.142	0.424	-0.064
B18	0.48	**0.565**	0.308	0.255	0.008
B19	0.362	**0.615**	0.331	0.196	0.005
B20	0.391	**0.602**	0.283	0.25	-0.15
B04	**0.502**	0.485	0.082	0.188	0.37
B05	**0.536**	0.392	0.244	0.293	0.148
B07	**0.5**	0.44	0.039	0.415	0.143
B09	**0.581**	0.432	0.144	0.366	-0.07
B10	**0.669**	0.242	0.049	0.191	0.221
B11	**0.704**	0.34	0.174	0.147	0.198
B12	**0.644**	0.328	0.025	0.298	0.13
B13	**0.673**	0.398	-0.001	0.249	0.213
B14	**0.654**	0.43	0.137	0.237	0.073
B15	**0.631**	0.349	0.053	0.153	0.149
B16	**0.536**	0.379	0.035	0.133	0.202
B17	**0.614**	0.447	0.147	0.12	0.213
B21	**0.565**	0.452	0.348	0.251	-0.16
B22	**0.68**	0.314	0.35	0.181	-0.024
B23	**0.7**	0.245	0.289	0.195	0.068
B24	**0.73**	0.124	0.217	0.214	0.104
B25	**0.719**	0.123	0.263	0.143	0.296
B26	**0.695**	0.222	0.143	0.213	0.209
B27	**0.672**	0.059	0.273	-0.041	0.312
B28	**0.465**	0.444	0.195	0.085	0.354
B29	0.056	0.242	**0.768**	0.139	0.14
B30	0.081	0.269	**0.779**	0.074	0.207
B31	0.209	0.145	**0.555**	0.068	0.532
B32	0.28	0.123	**0.695**	0.372	0.108
B33	0.355	0.045	**0.512**	0.277	0.15
B34	0.386	0.069	**0.484**	0.48	0.213
B06	0.348	0.375	0.241	**0.412**	0.157
B35	0.121	0.305	0.259	**0.687**	0.267
B36	0.334	0.144	0.215	**0.697**	0.259
B40	0.282	0.159	0.252	**0.7**	0.236

Items	Components				
	1	2	3	4	5
B37	0.348	0.048	0.239	0.228	**0.648**
B38	0.186	0.15	0.283	0.234	**0.639**
B39	0.162	0.21	0.093	0.462	**0.621**

A cluster of seven items emerged and formed a factor called "readiness and planning." These items demonstrated factor loadings ranging from 0.5 to 0.7. The items within this factor revolve around the preparedness and alignment of goals with the documents prepared and presented by educators to students. Additionally, a set of 20 items formed another factor. Upon analyzing these items, it was observed that they related to pedagogical skills, conceptual mastery, and the ability to facilitate student learning. Hence, this factor was named "pedagogy and professionalism," in accordance with the competencies outlined in the Law on Teachers and Lecturers. The remaining six items measured aspects pertaining to the role model behavior and self-control of educators. Consequently, these six items were labelled "personality," aligning with the competencies mentioned in the Law on Teachers and Lecturers.

In addition, the analysis of the extracted factors revealed a set of items that assessed the educators' ability to establish effective relationships with students within the classroom setting. This factor was named "social relationships within classroom". Finally, three items measured the educators' competence in building social connections outside the classroom, resulting in the factor named " social relationships beyond classroom".

Table 3. Sample items on Instrument

Factor	Sample item
Readiness and planning	The lecturer is always ready to give lectures and/or conduct practical sessions. The lecturer has the necessary course materials and equipment. The lecturer consistently arrives on time. The content delivered by the lecturer aligns with the designated competencies. Examinations and/or assignments correspond to the learning objectives of the course. The lecturer communicates the teaching and assessment methods to the students. The lecturer prepares the materials diligently.
Pedagogy and professionalism	The lecturer tries to stimulate students' interest in the course at the beginning of the lecture. The lecturer effectively communicates the objectives and content and responds to questions. The lecturer provides dedicated time for discussing the course material. The lecturer can guide discussions to achieve the intended goals. The lecturer employs diverse teaching methods. The lecturer explains the connections between the taught field/topics and other areas.

Factor	Sample item
	The lecturer utilizes research findings to enhance the quality of teaching.
Personality	The lecturer possesses a strong belief in their teaching abilities. The lecturer exercises prudence in decision-making. The lecturer serves as a role model in demeanour and conduct. The lecturer demonstrates self-discipline in diverse circumstances.
Social relationships within classroom	The lecturer is skilled at creating an engaging classroom atmosphere. The lecturer demonstrates impartiality and fairness in their interactions with students. The lecturer actively seeks and values feedback, criticism, and input from others. The lecturer exhibits tolerance and respect for the diversity of students.
Social relationships beyond classroom	The lecturer willingly offers their time for consultations outside of regular class hours. The lecturer possesses a thorough knowledge of the students who are enrolled in the course. The lecturer effortlessly establishes positive relationships with all academic community members, including students.

The factor analysis results indicate that the developed instrument measures five factors (Fig. 3) relevant to the four competencies of teachers and lecturers mentioned in the Teachers and Lecturers Law: pedagogical, personality, professional, and social competencies. Pedagogical competency refers to the educator's ability to manage student learning. Personality competency pertains to the educator's ability to possess a strong, noble, wise, and authoritative personality, serving as a role model for students. Professional competency relates to the educator's ability to have comprehensive and in-depth knowledge of the subject matter. Social competency involves the educator's ability to communicate and interact effectively and efficiently with students, fellow educators, parents/guardians, and the surrounding community. These four competencies are the primary modalities required to become a professional educator (Avalos, 2011; Teachers and Lecturers Law of the Republic of Indonesia No. 14 of 2005).

The analysis revealed a factor of pedagogy and professionalism, which represents a combination of items related to both pedagogical and professional competencies. This finding is expected, considering the close connection between these two competencies. Previous studies have also provided insights into the intersection and relationship between these competencies (Bell et al., 2010; Getenet, 2017; Hill et al., 2008; Kang, 2018; Mardiah & Yulhendri, 2020; Morris et al., 2009; van Driel & Berry, 2012; Wu & Cai, 2022). These studies support the strong correlation between an educator's abilities in performing their pedagogical and professional duties.

However, it is important to note that three items required revision as they did not align with the measured factor. These three items pertained to the suitability of the material, appropriateness of assessment outcomes, and mastery of the subject

matter. The factor analysis revealed that these items should be classified under the "readiness and planning" factor. We conducted focus group discussions and interviews with the raters to identify the attributes associated with these items. The findings indicated that although the wording of the items was appropriate, the raters believed these three items were more fitting within the domain of preparation and planning. This is because the assessment process, material suitability, and subject matter mastery are closely linked to an educator's preparedness to deliver instruction. Consequently, the three items were revised in terms of their wording based on the feedback received during the focus group discussions and interviews.

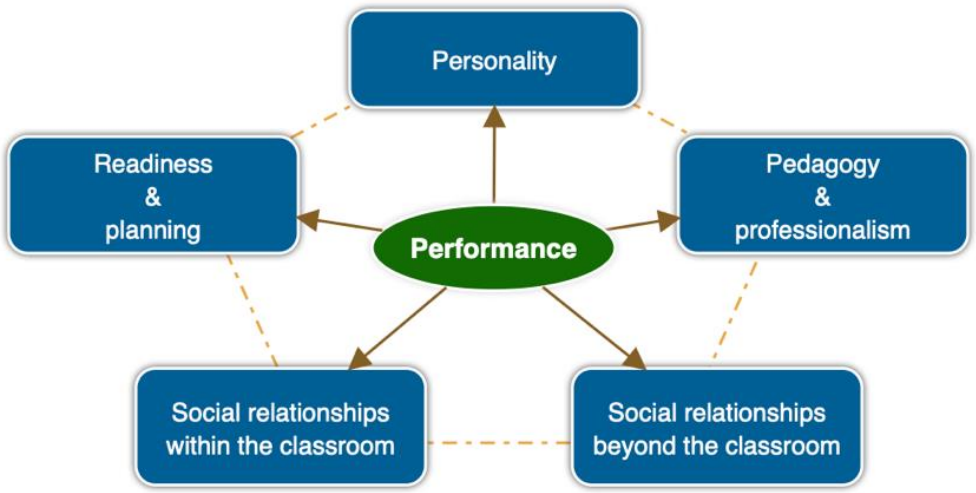

Figure 3. The Five Aspect of Performance Assessment for Educators in Higher Education

The findings reveal a similar pattern in assessing personality and social factors. While these factors are distinct, their close relationship sometimes makes it challenging to differentiate between them. It is worth noting that the raters in this study were students with limited age and experience in evaluating performance. Examining the fourth and fifth factors that emerged, it becomes evident that these items capture social competence as defined in the Teacher and Lecturer Law. However, the analysis results indicate some variations between the two factors. It suggests that the social interactions of educators within the classroom and beyond the classroom can be considered different. Hence, these factors do not merge into a single factor based on the analysis. In other words, classroom management and self-control challenges for educators inside and outside the classroom may possess distinct characteristics and perspectives. Leadership, role modelling, and attentiveness are among the elements assessed in this study, which are crucial for educators to address as they closely relate to classroom management quality (Taylor et al., 2011).

Once the validity evidence was obtained, the researchers conducted a G-Study to examine the magnitude of performance score variance influenced by other factors, namely, raters, items, and their interactions. The G-Study analysis was performed using RStudio with the following codes.

```
> modelG <- "scor ~(1 | person)+(1 | item)+(1 | rater)+(1 | item:rater)+(1 | rater:person)"
> g1 <- gstudy(data1,modelG)
> DStudy <- dstudy(g1, colname.objects = "person", data = data1, colname.scores = "scor")
```

```
> DStudy$components
> DStudy$generalizability
# Output
    source          var             percent     n
1   item:rater      3.970834e-13    0.0         120
2   item            1.985367e-04    0.1         40
3   rater:person    1.748585e-02    11.5        3
4   rater           0.000000e+00    0.0         3
5   person          1.334330e-01    87.4        1
6   Residual        1.541137e-03    1.0         120
[1] 0.88  # G.coef.
```

We applied the command to analyze the trial data, which we organized into four columns: person, rater, item, and the rightmost column representing the scores. The data analysis revealed two primary factors that accounted for the largest sources of variance: person (87.4%) and the interaction between rater and person (11.5%). These findings indicate that raters have minimal influence on the measurement outcomes. The person factor exhibited the highest percentage of variance, indicating that the measured scores effectively and precisely capture the assessed performance.

The analysis yielded a G-Study coefficient of 0.88, indicating a high level of precision. This coefficient represents the relative G-Study coefficient for three raters and 40 items, surpassing the minimum threshold of 0.7. The next step involved utilizing the G-Study results in the D-Study to estimate the required number of raters for achieving high measurement precision. The following syntax was employed to estimate the G coefficient for one rater and 1 item.

```
> r1 <- dstudy(g1, colname.objects = "person")
> r1$components
> r1$generalizability
#Output
    source          var             percent     n
1   item:rater      4.765001e-11    0.0         1
2   item            7.941469e-03    2.1         1
3   rater:person    5.245755e-02    13.8        1
4   rater           0.000000e+00    0.0         1
5   person          1.334330e-01    35.2        1
6   Residual        1.849364e-01    48.8        1
[1] 0.36. #G.coef for 1 rater
```

The D-Study analysis showed that the precision obtained from assessing one rater and 1 item was remarkably low at only 0.36. Subsequently, we utilized Rstudio to perform the D-Study analysis for different numbers of raters, starting from 2 and continuing onwards. In this stage of the D-Study, we calculated the G coefficient for various scenarios (rater = 2, 3, ..., 10) using the following codes.

```
> rater=1
> D_study <- data.frame("n_Rater"=c(1:10),"n_Item"=c(1:1),"G.Coef"='')
> while(rater<=10){
```

```
+D_study[rater,"G.Coef"]=round(r1$var.universe/(r1$var.universe +
(r1$var.error.rel/rater)),3)
+ rater=rater+1}
> D_study
#Output
```

	n_Rater	n_Item	G.Coef
1	1	1	0.36
2	2	1	0.529
3	3	1	0.628
4	4	1	0.692
5	5	1	0.738
6	6	1	0.771
7	7	1	0.797
8	8	1	0.818
9	9	1	0.835
10	10	1	0.849

A G coefficient of 0.738 was obtained from the D-Study for five raters. This D-Study outcome will be applied in evaluating educators' performance during the teaching process. Gathering assessments from 5 raters has demonstrated satisfactory result precision, leading to time savings and improved assessment efficiency.

We conducted this study to validate the construct and establish the instrument's reliability using EFA and Generalizability Theory. Through these analyses, we gained valuable insights into the factors that contribute to the performance of educators in higher education, as well as the specific items used for measurement. The results have practical implications for institutions, as they can consider streamlining the assessment process by involving a manageable number of raters to evaluate performance. Notably, the D-Study findings demonstrate that a reliable assessment can be achieved with just five raters. This suggests the potential for time and resource savings, improving the efficiency of the assessment process. However, institutions may still choose to involve a larger number of raters while carefully considering efficiency and the potential impact of student fatigue and stress on evaluation accuracy. We present the D-Study results of up to 10 raters as a valuable consideration for institutions aiming to use a larger number of raters. With this number, the reliability level reaches an impressive 0.849, indicating higher precision and reliability in the assessment outcomes.

The instrument developed and validated in this study has a broader scope as it aims to evaluate the performance of educators in the implementation of teaching across diverse programs and disciplines within a university. Rather than specifically targeting the measurement of educators' performance in a particular field of knowledge, the instrument focuses on assessing their overall performance in instructional delivery and interactions within the learning environment. Therefore, it only somewhat measures the specific skills associated with domain competencies.

The researchers strongly believe that the orientation change and performance of educators greatly support the development of skills and learning outcomes for students (Chen & Terada, 2021). Therefore, using high-quality assessment instruments for performance evaluation serves as a gateway to understanding the quality and acts as a diagnostic tool for identifying weaknesses and potential future improvement

areas. Furthermore, it is essential to continue developing the assessment instrument for educator performance in instructional delivery, considering various core competencies of the 21st Century. These competencies can be accommodated within the performance assessment, particularly in the pedagogical and professional factors, and the measurement can be expanded to encompass other aspects that are not currently captured by this instrument. The social factor, which encompasses collaboration, communication, and teamwork abilities, can also serve as valuable input for the future development of similar instruments.

CONCLUSION

Based on our study findings, we draw the following conclusions regarding the validity and reliability of the instrument for measuring educators' performance in teaching at the higher education level. Firstly, we have established that the developed instrument demonstrates strong construct validity. Secondly, the measurement instrument encompasses five relevant factors associated with the core competencies of teachers and lecturers outlined in the Teachers and Lecturers Law: readiness and planning, pedagogy and professionalism, personality, social relationships within the classroom, and social relationships beyond the classroom. Thirdly, the performance assessment instrument exhibits high reliability, as evidenced by the G-study and D-study results, which yielded coefficients of G exceeding 0.7. Lastly, it is recommended to involve a minimum of five raters in order to achieve assessment results with a reliability above 0.7 when evaluating educators' performance in teaching at the higher education level.

RECOMMENDATION

Although this study successfully demonstrated the validity and reliability of the developed performance assessment instrument, there are still several limitations that can guide further research. Firstly, performance measurement focuses on fundamental aspects related to teaching implementation, such as conceptual mastery, classroom management, and social relationships. However, the instrument has yet to measure in-depth aspects of specific fields of expertise or disciplines. Secondly, the instrument has not incorporated various 21st Century competencies, such as communication skills, collaboration, and other essential abilities.

In conclusion, we recommend that performance assessments in higher education embrace the generalizability theory as a framework for developing assessment instruments and designing appropriate and efficient assessment methods. Expanding the instrument to encompass core 21st-century competencies, including collaboration, communication, and teamwork skills, would be beneficial in capturing vital social aspects within the learning context. Additionally, involving diverse stakeholders with experience and expertise in assessing educator performance would enhance the use of raters. Lastly, conducting further studies to validate this instrument in different educational environments would broaden the generalizability of the findings.

Author Contributions

Conceptualization, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, and software, HD; methodology, HD, EI, and W; validation, EI, and W; formal analysis, HD, EI, and W; supervision, EI and W; project administration, HD. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Acknowledgment
We would like to express our heartfelt gratitude to the dedicated participation and invaluable support from the esteemed faculty members and academic staff of Universitas Sembilanbelas November Kolaka (USN Kolaka). Their active engagement, scholarly insights, and unwavering encouragement have shaped the research and elevated its academic merit. We acknowledge and appreciate their commitment to fostering a conducive environment for intellectual growth and collaborative research within the university community. Thank you all for your significant contributions, which have enriched the quality and depth of this work.

Conflict of interests
The authors declare no conflict of interest.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Wadstworth.

Avalos, B. (2011). Teacher professional development in Teaching and Teacher Education over ten years. *Teaching and Teacher Education*, *27*(1), 10–20. https://doi.org/10.1016/j.tate.2010.08.007

Batubara, H. H. (2016). Penggunaan Google Form sebagai alat penilaian kinerja dosen di Prodi PGMI UNISKA Muhammad Arsyad Al Banjari. *Jurnal Pendidikan Dasar Islam*, *8*(1), 39–50. http://ejournal.unsub.ac.id/index.php/sendinusa/article/view/661

Baya'a, N., & Daher, W. (2013). Mathematics teachers' readiness to integrate ICT in the classroom. *International Journal of Emerging Technologies in Learning*. https://doi.org/10.3991/ijet.v8i1.2386

Bell, C. A., Wilson, S. M., Higgins, T., & Mccoach, D. B. (2010). *Measuring the Effects of Professional Development on Teacher Knowledge : The Case of Developing Mathematical Ideas*. *41*(5), 479–512.

Bimpeh, Y., Pointer, W., Smith, B. A., & Harrison, L. (2020). Evaluating Human Scoring Using Generalizability Theory. *Applied Measurement in Education*, *33*(3), 198–209. https://doi.org/10.1080/08957347.2020.1750403

Brennan, R. L. (2001). Generalizability theory: Statistics for social science and public policy. In *New York: Springer-Verlag.* (Vol. 30).

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Brookhart, S. M., & McMillan, J. H. (2020). Classroom assessment and educational measurement. In *Classroom Assessment and Educational Measurement*. Routledge. https://doi.org/10.4324/9780429507533-5

Chen, Y. C., & Terada, T. (2021). Development and validation of an observation-based protocol to measure the eight scientific practices of the next generation science standards in K-12 science classrooms. *Journal of Research in Science Teaching*, *58*(10), 1489–1526. https://doi.org/10.1002/tea.21716

Djidu, H., Mashuri, S., Nasruddin, N., Sejati, A. E., Rasmuin, R., Ugi, L. E., & Arua, A. La. (2021). Online learning in the post-Covid-19 pandemic era: Is our higher education ready for it? *Jurnal Penelitian Dan Pengkajian Ilmu Pendidikan: E-Saintika*, *5*(2), 139–151. https://doi.org/10.36312/esaintika.v5i2.479

Djidu, H., & Retnawati, H. (2022). Digitizing mathematics and science learning: What do we need to prepare? *5th International Conference on Current Issues in Education (ICCIE 2021)*, *640*, 296–301. https://www.atlantis-press.com/article/125969632.pdf

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice-Hall International, Inc.

Gable, R. K., & Wolf, M. B. (1993). Instrument Development in the Affective Domain. In *Instrument Development in the Affective Domain*. Springer Netherlands. https://doi.org/10.1007/978-94-011-1400-4

Getenet, S. T. (2017). Adapting technological pedagogical content knowledge framework to teach mathematics. *Education and Information Technologies*, *22*(5), 2629–2644. https://doi.org/10.1007/s10639-016-9566-x

Gil-Flores, J., Rodríguez-Santero, J., & Torres-Gordillo, J.-J. (2017). Factors that explain the use of ICT in secondary-education classrooms: The role of teacher characteristics and school infrastructure. *Computers in Human Behavior*, *68*, 441–449. https://doi.org/10.1016/j.chb.2016.11.057

Henkel, M. (1997). Teaching Quality Assessments. *Evaluation*, *3*(1), 9–23. https://doi.org/10.1177/135638909700300102

Hermanu, A. I., Sari, D., Sondari, M. C., & Dimyati, M. (2022). Is it necessary to evaluate university research performance instrument? Evidence from Indonesia. *Cogent Social Sciences*, *8*(1). https://doi.org/10.1080/23311886.2022.2069210

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). *Content Knowledge : Conceptualizing and Measuring Teachers ' Topic-Specific Knowledge of Students*. *39*(4), 372–400.

Hu, B. Y., Fan, X., Yang, Y., & Neitzel, J. (2017). Chinese preschool teachers' knowledge and practice of teacher-child interactions: The mediating role of teachers' beliefs about children. *Teaching and Teacher Education*, *63*, 137–147. https://doi.org/10.1016/j.tate.2016.12.014

Ikram, F. F. D., Komala, N., & Syaefullah, T. W. (2018). Analisa Sistem EDOM Politeknik Negeri Jakarta Menggunakan Technology Acceptance Model (TAM). *MULTINETICS*, *4*(1), 34. https://doi.org/10.32722/vol4.no1.2018.pp34-38

Jeffrey, L. M., Milne, J., Suddaby, G., & Higgins, A. (2014). Blended Learning : How Teachers Balance the Blend of Online and Classroom Components. *Journal of Information Technology Education*, *13*, 121–140. https://doi.org/10.28945/1968

Johnson, E. S., Crawford, A., Moylan, L. A., & Zheng, Y. (2020). Validity of a Special Education Teacher Observation System. *Educational Assessment*, *25*(1), 31–46. https://doi.org/10.1080/10627197.2019.1702461

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2020). Examining rater accuracy and consistency with a special education observation protocol. *Studies in Educational Evaluation*, *64*, 0–18. https://doi.org/10.1016/j.stueduc.2019.100827

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2022). Evaluating an explicit instruction teacher observation protocol through a validity argument approach. *The Journal of Experimental Education*, *90*(2), 419–434. https://doi.org/10.1080/00220973.2020.1811194

Kang, E. (2018). Exploring Elementary Teachers' Pedagogical Content Knowledge and Confidence in Implementing the NGSS Science and Engineering Practices. *Journal of Science Teacher Education*, *29*(1), 9–29. https://doi.org/10.1080/1046560X.2017.1415616

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal*, *17*(1), 129–146. https://doi.org/10.1177/1474904117703228

Mardiah, M., & Yulhendri, Y. (2020). Pengaruh IPK, micro teaching, dan praktik pengalaman lapangan (PPL) terhadap kompetensi pedagogik mahasiswa calon guru jurusan Pendidikan Ekonomi FE UNP. *Jurnal Ecogen*, *3*(1), 165–175. https://doi.org/10.24036/jmpe.v3i1.8535

Martin, B. A., & Martin, J. H. (1989). Assessing the Lecture Performance of University Faculty: A Behavioral Observation Scale. *Journal of Education for Business*, *64*(4), 157–160. https://doi.org/10.1080/08832323.1989.10117350

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). Instrument development in the affective domain. In *Journal of Chemical Information and Modeling* (3rd ed.). Springer.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). Measurement and assessment in teaching. In *Library of Congress Catalogging in Publication Data*. Pearson Education, Inc.

Moore, C. T. (2016). *gtheory: Apply Generalizability Theory with R*. http://evaluationdashboard.com

Morris, A. K., Hiebert, J., & Spitzer, S. M. (2009). *Mathematical Knowledge for Teaching in Planning and Evaluating Instruction : What Can Preservice Teachers Learn ? 40*(5), 491–529.

Noben, I., Deinum, J. F., & Hofman, W. H. A. (2022). Quality of teaching in higher education: reviewing teaching behaviour through classroom observations. *International Journal for Academic Development*, *27*(1), 31–44. https://doi.org/10.1080/1360144X.2020.1830776

Poole, M., Harman, E., & Deden, A. (1998). Managing the Quality of Teaching in Higher Education Institutions in the 21st Century. *Australian Journal of Education*, *42*(3), 271–284. https://doi.org/10.1177/000494419804200305

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Rahardja, U., Lutfiani, N., Setiani Rafika, A., & Purnama Harahap, E. (2020). Determinants of Lecturer Performance to Enhance Accreditation in Higher Education. *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, 1–7. https://doi.org/10.1109/CITSM50537.2020.9268871

Retnawati, H., Apino, E., Djidu, H., Ningrum, W. P., Anazifa, R. D., & Kartianom, K. (2019). Scaffolding for international students in statistics lecture. *Journal of Physics: Conference Series*, *1320*(1). https://doi.org/10.1088/1742-6596/1320/1/012078

Retnawati, H., Djidu, H., Kartianom, K., Apino, E., & Anazifa, R. D. (2018). Teachers' knowledge about higher-order thinking skills and its learning strategy. *Problems of Education in the 21st Century*, *76*(2), 215–230. http://oaji.net/articles/2017/457-1524597598.pdf

Retnawati, H., Hadi, S., & Nugraha, A. C. (2016). Vocational high school teachers' difficulties in implementing the assessment in curriculum 2013 in Yogyakarta province of Indonesia. *International Journal of Instruction*, *9*(1), 33–48. https://doi.org/10.12973/iji.2016.914a

Rodgers, W. J., Morris-Mathews, H., Romig, J. E., & Bettini, E. (2022). Observation Studies in Special Education: A Synthesis of Validity Evidence for Observation Systems. *Review of Educational Research*, *92*(1), 3–45.

https://doi.org/10.3102/00346543211042419

Safi'i, I., Warni, S., & Yanti, P. G. (2019). Persepsi Guru Bahasa Indonesia tentang Hubungan antara Penerapan Full Day School dengan Penguatan Karakter Siswa. *Jurnal Pendidikan Karakter*, *9*(2). https://doi.org/10.21831/jpk.v9i2.27361

Stylianides, G. J. (2007). Investigating the guidance offered to teachers in curriculum materials: the case of proof in mathematics. *International Journal of Science and Mathematics Education*, *6*(1), 191–215. https://doi.org/10.1007/s10763-007-9074-y

Sulistiyo, U., Mukminin, A., Abdurrahman, K., & Haryanto, E. (2017). Learning to teach: A case study of student teachers' practicum and policy recommendations. *The Qualitative Report*, *22*(3), 712–731. https://nsuworks.nova.edu/tqr/vol22/iss3/3

Taufiq, R. (2015). Penilaian Kinerja Dosen Dalam Bidang Belajar Mengajar Di Fakultas Teknik Universitas Muhammadiyah Tangerang. *Faktor Exacta*, *5*(1), 77–85. https://journal.lppmunindra.ac.id/index.php/Faktor_Exacta/article/view/185

Taylor, M., Yates, A., Meyer, L. H., & Kinsella, P. (2011). Teacher professional leadership in support of teacher professional development. *Teaching and Teacher Education*, *27*(1), 85–94. https://doi.org/10.1016/j.tate.2010.07.005

Undang-Undang Republik Indonesia Nomor 14 Tahun 2005 tentang Guru dan Dosen, (2005).

van Driel, J. H., & Berry, A. (2012). Teacher professional development focusing on pedagogical content knowledge. *Educational Researcher*, *41*(1), 26–28. https://doi.org/10.3102/0013189X11431010

Van Tassel-Baska, J., Quek, C., & Feng, A. X. (2006). The development and use of a structured teacher observation scale to assess differentiated best practice. *Roeper Review*, *29*(2), 84–92. https://doi.org/10.1080/02783190709554391

Wu, Y., & Cai, J. (2022). Does school teaching experience matter in teaching prospective secondary mathematics teachers? Perspectives of university-based mathematics teacher educators. *ZDM – Mathematics Education*, *0123456789*. https://doi.org/10.1007/s11858-022-01344-8

Zurqoni, Z., Retnawati, H., Rahmatullah, S., Djidu, H., & Apino, E. (2020). Has arabic language learning been successfully implemented? *International Journal of Instruction*, *13*(4). https://doi.org/10.29333/iji.2020.13444a