KINDERGARTEN: Journal of Islamic Early Childhood Education
p-ISSN: 2621-0339 |e-ISSN: 2621-0770, hal. 15-23
Vol. 6, No. 1, April 2023
DOI: http://dx.doi.org/10.24014/kjiece.v6i1.22136

Intrarater Reliability of Accreditation Assessment for Early Childhood Education

Petrus Redy Partus Jaya [1], Theresia Alviani Sum [2], dan Felisitas Ndeot[3]

[1,2,3] PG PAUD, FKIP, Universitas Katolik Indonesia Santu Paulus Ruteng
e-mail corresponden: petrusredypartusjaya@gmail.com

ABSTRAK. *Penelitian ini mengukur reliabilitas intrarater antara asesor akreditasi satuan Pendidikan Anak Usia Dini (PAUD) di Provinsi Nusa Tenggara Timur (NTT). Dalam penilaian akreditasi satuan PAUD tahun 2022, 50 lembaga yang tersebar di 15 kabupaten di Provinsi NTT dijadikan sebagai sampel. Teknik pengujian reliabilitas intrarater yang digunakan adalah teknik intraclass correlation coefficient (ICC). Hasil penelitian menunjukkan tingkat kesepahaman yang baik antara Asesor A dan Asesor B, dengan koefisien ICC dalam penilaian akreditasi satuan PAUD tahun 2022 di Provinsi Nusa Tenggara Timur tergolong tinggi dengan besaran estimasi ICC 0,644. Namun, koefisien ICC antara Asesor A dengan Validator maupun antara Asesor B dengan Validator berada pada kategori sedang dengan tingkat kesepakatan yang moderat. Selain itu, dalam penelitian ini diperoleh nilai effect size yang kecil. Hal ini menunjukkan bahwa faktor subyektif asesor seperti lingkungan asal maupun pengalaman melakukan akreditasi berkorelasi sangat kecil terhadap skor akreditasi. Oleh karena itu, perlu dilakukan upaya-upaya untuk meningkatkan kesepakatan antara Asesor A dan Validator serta antara Asesor B dan Validator guna meningkatkan reliabilitas intrarater dalam penilaian akreditasi satuan PAUD di Provinsi Nusa Tenggara Timur.*

Kata Kunci: *Reliabilitas intrarater; Akreditasi; Pendidikan Anak Usia Dini*

ABSTRACT. *This study measured intrarater reliability between assessors of Early Childhood Education (PAUD) unit accreditation in the East Nusa Tenggara Province (NTT), Indonesia. In the 2022 PAUD unit accreditation assessment, 50 institutions across 15 districts in the NTT Province were sampled. The intrarater reliability testing technique used was the intraclass correlation coefficient (ICC). The results showed good agreement between Assessor A and Assessor B, with the ICC coefficient in the 2022 PAUD unit accreditation assessment in the NTT Province being relatively high with an estimated ICC of 0.644. However, the ICC coefficient between Assessor A and Validator, as well as between Assessor B and Validator, was in the moderate category with a moderate level of agreement. In addition, the study found a small effect size, indicating that subjective factors such as assessors' backgrounds and accreditation experience had a very small correlation with accreditation scores. Therefore, efforts should be made to improve agreement between Assessor A and Validator as well as between Assessor B and Validator to improve intrarater reliability in the PAUD unit accreditation assessment in the NTT Province.*

Keyword: *Intrarater reliability; accreditation; Early childhood education*

INTRODUCTION

Early childhood education is one of the important stages in human development. In Indonesia, the government prioritizes the development of quality early childhood education, one of which is through accreditation. Accreditation of early childhood education institutions is an important process to ensure that these institutions provide quality services for young children. As part of the quality assurance process, early childhood education assessors play a crucial role in monitoring and evaluating the effectiveness of management and service programs provided by early childhood education units. However, confidence in the consistency of assessments among assessors can be a challenge. In the context of early childhood education accreditation, intra-rater reliability is defined as the agreement of assessment results among several assessors of the same early childhood education institution/unit. The reliability of assessments conducted by early childhood education assessors needs to be considered because inconsistencies and biases in scoring can affect the validity of the evaluation results.

Research on intra-rater reliability has been conducted in various fields, both academic and industrial. Nevertheless, this research topic is still very important to study because it has crucial implications in decision-making and formulation of recommendations. Errors in assessment can lead to incorrect decisions and inappropriate recommendations. Furthermore, this cycle of errors will continue to occur and have a negative impact. This is emphasized in Grimes and Ford's (2014) study, which found that low intra-rater reliability in the assessment process can lead to decision-making errors and can impact institutional performance (Grimes, P. W., & Ford, 2014). In addition, the topic of consistency in assessment among assessors is not only always relevant to be researched, but also relevant to be used as an evaluation material. Research by Ozcelik (Özçelik, 2021) found that there is an imbalance among assessors in assessing the quality of the same journal. The results of this research are used as an evaluation material to determine efforts to improve the consistency of assessment among assessors. Similarly, research by Zakiyah et al. explains that in assessing the quality of study programs, there is a significant difference in assessment among assessors. This research also has an impact on recommendations to improve agreement among assessors in providing assessments (Zakiyah, Y., Muslimin, I. A., & Siswoyo, 2019).

In the context of accreditation for early childhood education institutions (PAUD), research on intrarater reliability among assessors is important to be conducted. As accreditation is an important process to ensure quality services provided by PAUD units, inter-assessor reliability is also important to ensure consistency in assessment results among assessors. The use of qualitative assessment instruments and approaches can be seen as a challenge in conducting PAUD accreditation. This approach is more dependent on different perceptions and assessments by assessors and can lead to differences in interpretation in assessments. This can affect agreement among assessors and can result in inconsistent assessments. According to Hosseini et al., qualitative approaches can lead to subjective and non-standardized assessments that can affect the validity and reliability of the accreditation process (Hosseini, S. S., Zandieh, M., & Afzali, 2020). However, qualitative approaches in assessment also have positive implications, especially if they are aimed at helping to improve and develop PAUD units according to the characteristics and contexts of the accredited units (Katsikas, S. L., Natsis, A. G., & Tsioumis, 2019). In this context, it is very important to always conduct research and evaluation related to intrarater reliability or inter-assessor consistency in PAUD assessment.

Intrarater reliability is a term used in research to measure how consistent two or more raters are in assessing the same object or subject. Generally, this is done to avoid bias or subjective assessment that can affect assessment results. Cohen (Cohen, 1960) explained inter-rater reliability as "the degree to which two or more observers agree in their ratings of a given set of objects." Consistent with Cohen's opinion, Fleiss asserted that intrarater reliability can be defined as the extent to which different assessors agree in assessing the same target (Shrout, P. E., & Fleiss, 1979). There are several factors that contribute to intrarater reliability coefficients. These factors include: *Assessor Experience***:** Assessor experience is one of the determining factors of intrarater reliability coefficients. Some studies have shown that assessor experience can significantly impact intrarater reliability coefficients in various assessment contexts, including accreditation assessments (Barcikowski, R. S., & Ketrow, 2014). In the context of early childhood education and care (ECEC) accreditation, experienced assessors tend to be more consistent than new assessors who are still lacking in knowledge and skills related to ECEC accreditation assessment.

Assessment Criteria: Clarity and specificity of assessment criteria can also affect intrarater reliability coefficients. Research has shown that unclear criteria with vague limitations can cause inconsistencies in various assessment and evaluation contexts (Hsu, Y. C., Liang, J. C., & Tsai, 2019). In the context of ECEC accreditation, clarity and specific criteria can help ensure that assessors evaluate observations or assessments based on the same perspective consistently. *Training and Support:* Previous studies have shown that the adequacy of training in assessment

DOI: http://dx.doi.org/10.24014/kjiece.v6i1.22136

instruments and techniques for assessors is a crucial factor in ensuring reliable and consistent evaluations. In ECEC accreditation, preparing assessors with sufficient training can help ensure that assessors have the necessary knowledge and skills to perform consistent assessments (Sibbald, B., Shen, J., McBride, A., & Cumming, 2013). *Subjective/Personal Bias:* Assessor subjective biases can impact the intrarater reliability of accreditation assessments. Research has shown that assessor bias can affect inconsistencies and unreliable evaluations. In the context of ECEC accreditation, subjective biases must be addressed to ensure that assessors are aware of these biases and determine strategies to mitigate their impact on the accreditation assessment process (Heilbronner, R. L., & Strosser, 2013). *Communication:* Effective communication between assessors during the evaluation (accreditation) process can improve intrarater reliability coefficients between them. Chou et al emphasized this factor in their research. According to them, effective communication between assessors significantly increases their level of agreement in giving assessments. Efforts to build effective communication between assessors can ensure that they share an understanding of accreditation/assessment criteria, thus enabling evaluations to be done consistently (Chou, Y. C., Lin, H. W., & Chang, 2015).

Environmental Factor: The environment in which the assessor is located also has the potential to disrupt the reliability of accreditation assessments. Some studies have proven this factor. One of them, Harris et al (2018), found that environmental factors can interfere with and divert the level of accuracy and consistency of evaluations. In the context of early childhood education accreditation, the belief that assessments are carried out in a disturbance-free environment can increase the coefficient of reliability between assessors (Harris, R. B., He, J., & Chen, 2018). In addition to the factors that have been described, Eviati and Indrawati (Eviati, E., & Indrawati, 2018), added several potential factors that can significantly affect the coefficient of reliability between assessors. These factors include: Assessor characteristics, such as educational background, work experience, and technical skills; assessment instrument characteristics, such as feasibility, validity, and instrument reliability; and, assessment context characteristics, such as school characteristics, child age groups, and environmental conditions.

Some common intrarater reliability techniques include Pearson's correlation coefficient, intraclass correlation coefficient (ICC), and Kappa coefficient. When determining intrarater reliability testing techniques, data characteristics, measurement scales, and the number of evaluators need to be considered. The Kappa coefficient is more suitable for testing assessment consistency between assessors if the collected data is nominal and ordinal. If the measurement scale is interval or ratio, the intraclass correlation coefficient (ICC) technique can be used. The number of evaluators also determines the selection of techniques. For example, if there are only two evaluators, the Kappa coefficient or Pearson correlation can be used. However, if there are more than two evaluators, ICC is more suitable (Fleiss, 1981).

METHOD

This research was conducted using a quantitative approach. The purpose of the research was to measure the level of intrarater reliability coefficient among PAUD accreditation assessors in the NTT Province. The assessment results can be used as an evaluation material to review the accreditation process and improve the competence of PAUD assessors in the NTT Province. The number of institutions sampled to test inter-rater assessment consistency was 50 institutions spread across 15 districts in the NTT Province. The intrarater reliability tested was the assessment results given by Assessor A, Assessor B, and Validator Assessor. The data collected were scores of the visitation assessment given by Assessor A, Assessor B, and Validator Assessor in the 2022 PAUD accreditation. These assessment scores were interval data type. Therefore, the intrarater reliability testing technique used was the intraclass Correlation Coefficient (ICC) technique. The ICC coefficient categories refer to Fleiss' opinion. According to Fleiss (Fleiss, 1986), the ICC value can be categorized into the following five groups:

Table 1.
Criteria for ICC Coefficient by Fleiss

ICC Coefficient	Category	Meaning
< 0,40	Low	Significant inconsistency among assessors
0,40 s.d. 0,59	Moderate	Moderate agreement level among assessors
0,60 s.d. 0,74	High	High Good agreement level among assessors
0,75 s.d. 0,89	Very High	Very good agreement level among assessors
0,90 s.d. 1,00	Very High	Almost perfect agreement level among assessors

Source: (Fleiss, 1986: 1 – 32)

RESULT AND DISCUSSION

Based on the accreditation score data of 50 PAUD institutions sampled from 15 districts in NTT Province, the average assessment scores were obtained for assessor A, B, group score (assessor A-B), and validator assessor score as shown in Figure 1 below.

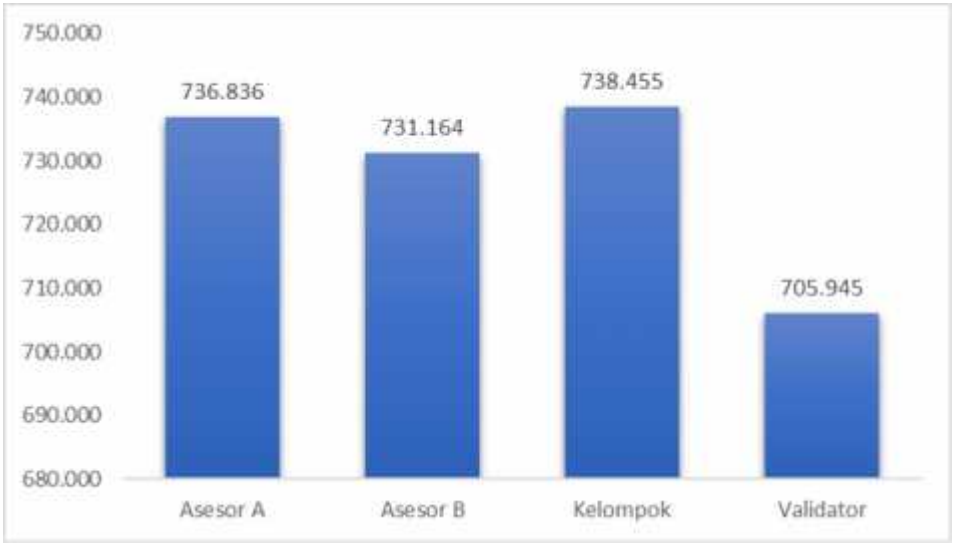

Figure 1. Average Scores of Assessment by Assessor A, B, Group, and Validator.

In Figure 1, it can be seen that the average score of assessor A tends to be similar to the group assessment score, and higher than the score of assessor B. Upon closer examination, the group assessment score is slightly higher than the assessment scores of assessor A and assessor B. This can happen because in the group assessment process, assessor A plays an important role. Assessor A assesses the group while coordinating or communicating with assessor B. Based on this mechanism, it can be assumed that the assessor A factor greatly determines the results of the group assessment. Higher scores can occur due to the accumulation of several accreditation items taken from the assessment results of assessor B.

The situation is different when comparing the average group assessment score with the validator assessment score. The validator assessment is much lower than the group and assessor A and B assessments. The validator assessment tends to refer to the evidence attached by assessor A and assessor B. The validity of the attached evidence will impact the similarity of scores with the assessment of assessor A, B or group. If the score is lower, it means that some evidence agreed upon by assessor A or B is declared invalid by the validator.

Although small, these score differences can potentially lead to different accreditation ranking conclusions, especially if the difference is on the threshold of the score range with

DOI: http://dx.doi.org/10.24014/kjiece.v6i1.22136

different ranks. This potential difference has been explained by Manshur and Haryanto (2015): "During accreditation assessment, there is a possibility that the score given by assessors approaches the threshold between categories, so there is a risk of errors in determining the appropriate accreditation category (Mansoer, R. M., & Haryanto, 2015)." Therefore, the average assessment scores of several assessors need to be tested for differences. Are there significant differences or just differences in the high and low averages? Testing these differences is done using a one-way ANOVA technique. This testing requires testing the prerequisite of variance homogeneity. The following are the results of the variance homogeneity testing for the assessment scores of assessor A, B, group, and validator using Levene's test.

Table 2.
Test for Equality of Variances (Levene's)

F	df1	df2	p
2.329	3.000	216.000	0.075

Table 2 shows that the p-value from the Levene's test is 0.075. This p-value is greater than 0.05. Thus, the assumption of homogeneity of variance for the scores of assessor A, B, Group, and Validator is accepted. This assumption test is further supported by the Q-Q Plot display below.

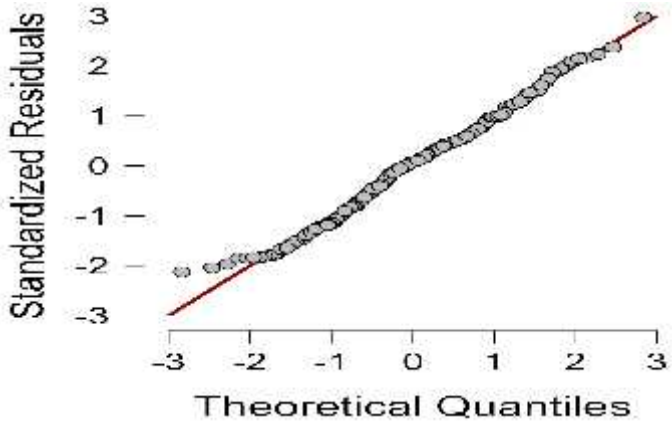

Figure 1. Q-Q Plot of Assessment Scores

Assumption of variance homogeneity is fulfilled when the score distribution of each assessor group is around the linear line. The Q-Q Plot in the figure shows that the score distribution of each assessor group is not far from the linear line.

After the assumption of variance homogeneity is fulfilled, testing for the differences in the mean scores among assessors can be performed. The results of this test are shown in the following Table 3.

Table 3.
Test for Differences in Mean Scores among Assessor A, Assessor B, and Validator.

Cases	Sum of Squares	df	Mean Square	F	P	η^2
Assessors	37606.273	3	12535.424	1.758	0.156	0.024
Residuals	1.541e+6	216	7132.035			

Note. Type III Sum of Squares

Based on Table 3, it can be seen that the p-value is > 0.05. Based on this value, it is concluded that there is no significant difference between the assessments of Assessor A, Assessor B, Group, and Validator. In addition to the p-value or significance of the difference, the effect size coefficient (η^2) should also be considered because in 2022, accreditation assessors for early childhood education are determined by a policy where one assessor comes from the accredited area and the other comes from a different area. The environmental/origin factor of the assessor is assumed to potentially cause subjective assessments.

Effect size (η^2) in one-way ANOVA testing is a measure of how much the independent variable affects the dependent variable in a population. Effect size (η^2) is used to interpret the magnitude of the effect or influence of the independent variable on the dependent variable. In the context of accreditation, assessors are viewed as the independent variable and assessment scores are viewed as the dependent variable. According to Cohen, the value of effect size (η^2) ranges from 0 to 1, with the following categories: $\eta^2 = 0.01$: small effect size; $\eta^2 = 0.06$: medium effect size; $\eta^2 = 0.14$: large effect size (Cohen, 1988). Based on this effect size (η^2) value criteria, it is concluded that the effect size (η^2) of the assessment between assessors (0.024) is in the small category. This means that the influence of the assessor's subjective factor is classified as small. Although there were no significant differences in the mean scores between the assessors and the effect size (η^2) was small, consistency testing among assessors still needs to be carried out to determine the consistency coefficient of assessment between assessors. A high consistency coefficient of assessment among assessors indicates agreement and consensus on accreditation assessment items and agreement on the evidence presented.

In general, compared to the assessment results of Assessor A, Assessor B, and Validator, the level of agreement is in the high category, or there is a good level of agreement among assessors with an ICC coefficient of 0.644. The ICC coefficient can be seen in Table 4 below:

Table 4.

Intrarater Reliability of Assessors A, B, and Validator.

Type	Point Estimate	Lower 95% CI	Upper 95% CI
ICC1,1	0.644	0.534	0.741

Note. 56 subjects and 3 raters/measurements. ICC type as referenced by Shrout & Fleiss (1979).

The ICC coefficient category among all assessor groups (A-B-Validator) is classified as high when assessed together. However, different information is obtained when the ICC coefficient is calculated based on the assessor group. The comparison of coefficients based on the assessor group is presented in the following figure:

DOI: http://dx.doi.org/10.24014/kjiece.v6i1.22136

Figure 2. Comparison of ICC Coefficients among Assessor Groups.

Assessor A and B tend to have higher ICC coefficients compared to the validator assessor because assessor A and B are involved in the initial or formative assessment, while the validator assessor is involved in the final or summative assessment. Formative assessment is done repeatedly to help improve the quality of the assessment. On the other hand, summative assessment is done at the end of the assessment process to determine whether an object of assessment meets the set standards or not. In formative assessment, assessors A and B have more opportunities to discuss and compare their assessments, thus helping to improve the consistency of assessment between assessors. This is related to the factor of communication between assessors; the more often assessors communicate to align their perceptions, the assessment will tend to be equivalent, which impacts the high ICC coefficient. Mokkink et al. (Mokkink, L. B., van der Vleuten, C. P. M., Bouter, L. M., Sollie, A. W., & Schellevis, 2010) have proven this condition in their research entitled Inter-rater agreement and reliability of the COSMIN. The study showed that assessors involved in the initial process or formative assessment (assessment visitation) tend to have higher ICC coefficients compared to assessors who assess at the final stage of accreditation.

The ICC coefficient obtained from this study describes that the potential factor contributing to the level of agreement/consistency of accreditation assessors in early childhood education in the NTT Province is communication between assessors. This inter-assessor communication can also deconstruct other factors such as accreditation assessment criteria and assessor experience. Malini and Ramya (Malini, P. S., & Ramya, 2019) through factor analysis techniques have proven the correlation between communication factors, consistency of interpretation of assessment criteria, and assessor experience. Intense communication can improve assessors' understanding of accreditation assessment criteria. The danger is if this agreement is based on wrong perceptions and interpretations of accreditation assessment items.

CONCLUSION

The intrarater reliability coefficient (ICC) in the assessment of PAUD unit accreditation in the Nusa Tenggara Timur province in 2022 is considered high with an estimated ICC of 0.644. This high category means that there is a good level of agreement among assessors. This good level of agreement tends to occur between assessor A and assessor B. Different conclusions are reached when the ICC coefficient is tested separately. The ICC coefficient between Assessor A and Validator as well as between Assessor B and Validator are in the moderate category with a moderate level of agreement. This condition is assumed to be the effect of intensive communication between Assessor A and Assessor B. Both assessors are tasked with assessing during the visitation and have more time to build communication in agreeing on the interpretation of accreditation assessment items. However, this communication will not be meaningful if it is based on incorrect interpretations of accreditation assessment criteria. In addition, from the results of this test, a small effect size value is obtained. This indicates that subjective assessor factors such as their environment of origin and experience in conducting accreditation are highly correlated with accreditation scores.

Based on the results of this study, several recommendations can be formulated for the next accreditation process, including: Communication between assessors needs to be continued to form the same understanding and interpretation of various PAUD accreditation assessment criteria; training and preparation of assessors remain an important stage in the accreditation process; mapping assessors based on the same origin region as the assessment does not have a subjective impact on accreditation assessment. Therefore, this mapping model can still be carried out but in collaboration with assessors from different regions; and BAN PAUD and PNF need to consider the development of accreditation instruments with a combined approach. This means that the proportion of qualitative accreditation criteria needs to be balanced with the proportion of quantitative criteria. This is aimed at mitigating and reducing different perceptions among assessors and improving assessment consistency among assessors.

REFERENSI

Barcikowski, R. S., & Ketrow, S. M. (2014). Assessing the inter-rater reliability of the Wilson Reading System® (WRS) coding system. *Journal of Psychoeducational Assessment*, *32*(1), 78–86.

Chou, Y. C., Lin, H. W., & Chang, Y. T. (2015). Evaluating the inter-rater reliability of an oral presentation assessment instrument using generalizability theory. *BMC Medical Education*, *15*(1), 56.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Eviati, E., & Indrawati, D. (2018). Evaluasi Hasil Belajar Anak Usia Dini dengan Menggunakan Pendekatan CIPP pada Kelompok B di TK Negeri Tegalrejo Yogyakarta. *Jurnal Pendidikan Anak Usia Dini*, *7*(2), 105–118.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions (2nd ed.)*. Wiley.

Fleiss, J. L. (1986). *Reliability Measurement. In the Design and Analysis of Clinical Experiments*. John Wiley & Sons.

Grimes, P. W., & Ford, J. K. (2014). Reliability testing: What do we know? *Human Resource Management Review*, *24*(4), 271–282.

Harris, R. B., He, J., & Chen, Q. (2018). Noise and response bias in classroom observations: A randomized controlled trial. *Journal of Educational and Behavioral Statistics*, *43*(5), 569–592.

DOI: http://dx.doi.org/10.24014/kjiece.v6i1.22136

Heilbronner, R. L., & Strosser, G. L. (2013). (2013). The effect of examiner bias on the forensic assessment of juvenile sex offenders. *Journal of Forensic Psychology Practice*, *13*(5), 362–378.

Hosseini, S. S., Zandieh, M., & Afzali, A. (2020). Accreditation of higher education institutions and quality assurance: A review of the literature. *Journal of Education and Practice*, *11*(19), 84–91.

Hsu, Y. C., Liang, J. C., & Tsai, C. C. (2019). Investigating the inter-rater reliability of the science teacher assessment framework in Taiwan. *International Journal of Science Education*, *41*(4), 481–499.

Katsikas, S. L., Natsis, A. G., & Tsioumis, K. A. (2019). Accreditation of higher education institutions: A review of the literature. *Journal of Educational and Social Research*, *9*(2), 11–18.

Malini, P. S., & Ramya, P. (2019). Factors influencing inter-rater reliability in accreditation assessment of educational institutions. *Journal of Engineering Education Transformations*, *32*(4), 245–250.

Mansoer, R. M., & Haryanto, S. (2015). *Sistem Akreditasi Pendidikan Indonesia*. Prenada Media.

Mokkink, L. B., van der Vleuten, C. P. M., Bouter, L. M., Sollie, A. W., & Schellevis, F. G. (2010). Inter-rater agreement and reliability of the COSMIN (Consensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Medical Research Methodology*, *10*(1), 82.

Özçelik, N. S. (2021). Intrarater reliability in assessing the quality of scientific journals. *Journal of Academic Librarianship*, *47*(4).

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. In *Psychological bulletin* (pp. 420–429).

Sibbald, B., Shen, J., McBride, A., & Cumming, C. (2013). (2013). Factors affecting agreement between physicians' and nurses' clinical decisions for newly admitted nursing home residents. *Journal of Interprofessional Care*, *27*(1), 76–83.

Zakiyah, Y., Muslimin, I. A., & Siswoyo, S. (2019). Intrarater reliability of the accreditation assessment process for study programs in Indonesian higher education. *Quality in Higher Education*, *25*(2), 155–169. https://doi.org/10.1080/13538322.2019.1587816