EDUCATIONAL SCIENCES: THEORY & PRACTICE eISSN: 2148-7561, ISSN: 2630-5984

Received: 27 June 2021

Revision received: 15 October 2021

Accepted: 28 November 2021

www.jestp.com

DOI 10.12738/jestp.2021.3.0010 ♦ **December** 2021 ♦ 21(4) ♦ 155-164

Article

Interatter Reliability E-Assessment-Based Dance Practice Assessment

Dinny Devi Triana
State University of Jakarta, Indonesia
Email: dinnydevi@unj.ac.id
https://orcid.org/0000-0002-3588-7890

Rivo Panji Yudha
17 Agustus 1945 University, Cirebon
Email: rivoyudha@yahoo.co.id
https://orcid.org/0000-0001-8833-5304

Ibnu Salman
National Research and Innovation Institution,
Email: ibnusalman81@gmail.com
https://orcid.org/0000-0001-8791-8372

Abstract

The purpose of this article was to analyze the inter-rater reliability in assessing Javanese dance practice *Golek Clunthang* learned by the students of Dance Education study program in State Univresity of Jakarta through e-assessment methods. The problem faced was how the inter-rater reliability works in assessing dance practice through e-assessment technology, since most dance assessments get critics that would determine the scores. Another problem was the element of subjectivity and the lack of consistency of the raters in the scoring process. Therefore, this researcher aimed at presenting the results of the inter-rater reliability of the e-assessment-based dance practice assessment instrument. The instrument was tested on 2 raters with 18 students in two groups of dance class. Inter-rater reliability was analyzed by Intraclass Correlation Coefficient. However, it was necessary to equalize the perception of the ratters in understanding the criteria and rubrics of the assessment, because the consistency of the reassessment would affect the reliability between raters, and this can be solved through the use of e-assessment in conducting the assessment. The advantage of assessing dance using e-assessment is that the assessor can re-assess, with the same results. E-assessment technically makes it easier to conduct an assessment, anywhere and anytime. This learning and assessment system will be interactive, because it can provide transparent feedback and scores. However, it is highly recommended to use more than 2 appraisers in order to minimize the element of subjectivity.

Keywords

Reliability, Inter-rater, Practice Assessment, Dance, E-assessment

Correspondence to Dinny Devi Triana, Dance Education Study Program, Faculty of Language and Arts, State Universitas Negeri Jakarta, email dinnydevi@unj.ac.id

Citation: Triana D. D., Yudha R. P, Salman I. (2021). Interatter Reliability E-Assessment-Based Dance Practice Assessment. *Educational Sciences: Theory and Practice, 21*(4), 155 - 164. http://dx.doi.org/10.12738/jestp.2021.3.0010

Dance is classified as a complicated motor skill that necessitates a lot of practice (Enghauser, 2012). Dance learning is a broad idea because it refers to a skill domain that is tied to behavior and knowledge. In other words, learning goals must result in a change in behavior or activity as a result of stimulus and response (Thorndike, 1978 – 1949). When children learn one of the traditional dances that must be expressed based on the culture, behavioral transformation occurs. Dance learning demands a hierarchical learning model that requires high-thinking, emotional, and physical operation, in addition to memorizing and absorbing facts and skills (Warburton, 2011).

In dance learning, not only is the analysis that exhibits skills required to determine whether a learning process is successful, but also it is the assessment that focuses on performance, such as movement order, the application of pace based on rhythm, strength, and sentiments (Steven & Hesketh, 2013). This difficulty explains the dance evaluation employing high-precision motion capture using a motion introduction sensor system via Markov's "hiding" model method, as well as some movements acquired in real time (Laraba & Tilmanne, 2016). The motion capture approach creates a virtual simulator prototype in which a dance performance is collected and compared to a template of previously taught dance motions (Aristidou et al., 2014).

A tool known as Performance Competence Evaluation Measure (PCEM) is used to evaluate aspects from dance performances (Krasnow & Chatfield, 2009) by using 2 steps, which involves first the literature reviews to study the measuring instrument and to describe a dance performance on the qualitative measuring instrument development, and second, to test the validity and reliability of PCEM with 3 judges and 20 samples by showing the reliability of intra-rater and inter-rater (Krasnow & Chatfield, 2009). Dance assessments which mostly explore metacognitive skills, creativity, communication skill and skill to work productively (Ridgway et al., 2004), which is hard to do objectively. PCEM was developed by Krasnow and Chatfield (2009) and it was used as a foundation research tool in this article, to analyze the inter-rater reliability in assessing dance practice with the use of technology.

Assessment is a controversial topic in dance education, and some experts argue that arts cannot be assessed objectively (Hernandez, 2012), while others acknowledge the importance of assessment in dance classes where students direct their own learning about composition, technique, and practice (Englebright & Mahoney, 2012; Harding, 2012). The argument depends on the definition of summative and formative assessment. Summative assessment or learning assessment is created to evaluate learning for the purpose of setting the score, grades, or ranks. On the other hand, formative assessment is considered a learning assessment (Stiggins, 2006) and is meant to give feedbacks for the students and teachers to improve their growth in the learning process (Schildkamp et al., 2020). More specifically, formative assessment is a process to clarify a learning goal and performance of work, subjects, or units; it gives continuous feedback about students' improvements to reaching the goals; and to revise instructions and students' works based on the feedback (Shute, 2008; Wisniewski et al., 2020).

Assessments that measure students' skills comprehensively, as in dance assessments, demand students to do the real life's works and to show the essence in applying their knowledge and skills, and therefore the authentic assessment is used (Mueller, 2005). A more complete data can be collected through authentic assessment, along with the documentation of the students' skills based on the learning order that they experience (Donnelly, 2006). On the other hand, topics about assessment has a different characteristic (Clements et al., 2003; Ginsburg et al., 2016), so not all topics can be assessed the same way. Therefore, there needs to be different assessing techniques (Kane, 2001). One of the techniques that can be authentically given to students are practice assignments.

The research (Jonsson & Svingby, 2007; Smit et al., 2017; Wulan, 2008) and some of the experts' arguments (e.g., Kan & Bulut, 2014) show the importance of practice assessment, along with a good rubric. Therefore, the rater inconsistency problem in understanding the rubrics affects the major difference on the score result given in the assessment paper (Andrade & Du, 2005). Some researches for instance, (Smit et al., 2017; Wulan, 2008) shows that most of the time, the rater is inconsistent in using the rubric. This is due to the lack of teachers' experience in operating the rubric and also the rubric's quality itself (Adnan & Bulut, 2014). Other than that, the inconsistencies also happen due to the lack of understanding of constructions or rubrics' aspects. This condition becomes a serious problem and pushes the empirical test of practice assessment reliability that is used in assessing students' dance practice assignments.

The rater tries to use a wide range when assessing the practice result. However, with the same training and teaching experience, rater may evaluate students' work differently (Lumley, 1998; Shafer et al., 2001). The difference is caused by how the raters understand and apply the assessment rubric, as well as subjectivity level in giving the assessment (Eckes, 2008). As a result, the students receive bias assessment results. This becomes a serious problem that encourage the empirical testing of practice assessment reliability that is used to assess students' practice assignments.

In these few decades, dance practice assessments have undergone several innovations from experts, but entering the 21st century and the development of information and technology (IT) it continues to bring major changes to society, especially since the use of internet in teaching was applied in some countries (Pang, 2020). So as in assessments, it is seen as an important part of a system with the purpose to strengthen the education quality and to facilitate mobility amongst students. Students have to be assessed using criteria, rules and procedures that are applied consistently (Ali et al., 2018; Zhang et al., 2019), because it seems that assessment is something that is always there in universities, even though there are changes and different goals. There are many ways of doing dance assessment using the help of technology, such as motion capture technology that is captured through the Kinect-based tracking of human skeleton (Alexiadis & Daras, 2014), motion analysis (Aristidou et al., 2014) as well as prototype by combining the real and virtual world (Hachimura et al., 2004).

Assessment using ICT is later known as e-assessment, which encompass the whole assessment process, starting from designing the work to keeping the results with the help of ICT (Joint Information Systems Committee (JISC), 2005). So as the use of e-assessment where the learning system and assessment is using computers, so the assessment becomes interactive because it can give transparent scores and feedback (Mackenzie, 2003). In all academic fields, especially the ones having a major development, there is always the need to analyze in order to get the overview and to learn where the gaps need to be filled (Creswell & Creswell, 2017; Webster & Watson, 2002). This can help support the assessment process in dance practice learning now by using Electronic Data Exchange to accelerate the communication between schools and assessment authorities; the process of learning marking and recording scores can be improved. A system where students' works are scanned and distributed has more advantage than the conventional system in terms of logistics (e.g., posting and tracking papers in a huge amounts of numbers), and a continuous monitoring can guarantee a reliable high marking. The works these days push the limit in some fields such as understanding the text and automatically analyze students' analysis and strategy (Ridgway et al., 2004).

Based on this study, the problem in this article stated to find out how the raters' reliability worked in assessing dance practice through the use of e-assessment technology. The *Golek Clunthang* dance is learned in formal and informal institutions, so a valid instrument was needed to do assessment as a result of learning. Therefore, in order to meet the need of the dance assessment, an instrument to assess the *Golek Clunthang* dance based on the movement variety was used. This helped assessors in doing a more valid and reliable assessment through the online system.

The purpose of this article was to analyze the reliability between raters in assessing Java dance practice *Golek Clunthang* learned by students of Dance Education study program in State University of Jakarta through e-assessment. E-assessment can be a solution to the assessment today and future (Mackenzie, 2003), so that the bias of assessing dance practice and the subjectivity can be eliminated. This article tested the reliability between assessors in assessing the dance practice using e-assessment. To test the reliability between two assessors, students were given scores by the raters on the same day but in a different time. The scores produced by each rater were compared to one another for each size through the e-assessment.

Literature Review

Inter-Rater Reliability

Most of the studies about assessment behavior report about the reliability of the assessors. This type of reliability can be considered as retests reliability from a single test. This is most commonly reported in medical literature, but it is rarely reported in the context of psychology or education assessment (Jonsson & Svingby, 2007). The reliability of intra-assessor can be reported as a single index to the whole assessment project, or for each assessor separately. In the last case, they are usually reported using the kappa Cohen statistic, or as a correlation coefficient between two texts from the same sets of essays (Shohamy et al., 1992) for the example of individual-sized study of intra-assessor reliability]. In the assessment program's description, the intra-assessor's reliability is indexed by the average of individual-assessor's reliability, with the intra-class correlation (ICC) or with the generalization index from the retest aspect that refers to the whole group but not to the individual-assessor.

Inter-rater reliability or reliability between observers is about how far do two or more assessors (or observers, code designers, examiners) agree. This discusses about the problem of applying the ranking system consistently. The reliability between assessors can be evaluated using different statistics. Some of the more common statistics are the percentage agreement, kappa, moment-product correlation, and intra-class correlation coefficient. A high reliability score between assessors is based on the strong agreement between the two assessors. A low reliability score between assessors is based on the weak agreement between two assessors (Lange, 2018).

The equivalence decision of a measuring tool is done by the inter-rater reliability method that can be done by using 3 methods, such as: Percent Agreement, Cohen's Kappa and Pearson's Product Moment Correlation. Percent Agreement and Cohen's Kappa were used to assess the reliability of an instrument that produces a nominal data (e.g., sick or not sick, whether an event happens or not) from an observation result. Meanwhile the interval or ratio scale data is done by the Pearson's Product Moment Correlation test, which is to correlate the measuring results between observers (Gwet, 2014; Hallgren, 2012; McHugh, 2012). When the Intra-class correlation coefficient (ICC) gets higher when considering a reliability between assessors, as well as the statistic size to determine a reliability level, the reliability between assessors gets better (Gross & Battié, 2002). The best reliability score between assessors is 0.90 to 1.00, while the good ICC score is 0.75 to 0.90. The medium score is 0.50 to 0.75, while the low score is less than 0.50.

Javanese Dance Practice "Golek Clunthang"

The Javanese traditional culture is one of the Indonesian cultures that is constantly affected by globalization. However, the Javanese traditional culture that is still entrenched in the society, which is now still supported and conserved by people in cities or villages, amongst the noblemen and common people, is the traditional ceremony related to life cycle, as well as an art form which is dance of beksa (Wibowo, 1981). Javanese culture and the people now are experiencing changes and shifts in so many aspects of life, affected by: 1) Social, politics, and culture, 2) The spirit of nationalism, 3) Globalization flow (SUMARYONO, 2003).

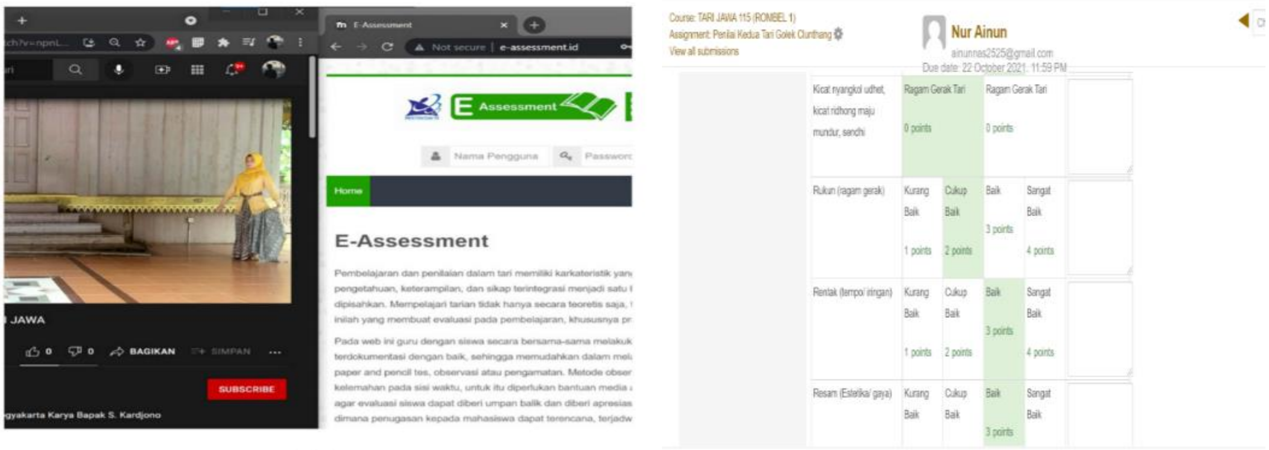
The Javanese dance *Golek Clunthang* is a subject studied in the Dance Education study program at State University of Jakarta. During this pandemic, the learning and assessment is done through e-assessment, so as not to reduce the quality of learning. To maintain the objectivity of the assessment of the learning outcomes of Javanese dance *Golek Clunthang*, an external assessor who has good Javanese dance competence is used.

E-assessment

Technology offers new steps to assess learning that will produce a rich source of data and can broaden the ways where the educators understand learning mastery and teaching effectiveness (Vendlinksi & Stevens, 2002). E-assessment (Joint Information Systems Committee (JISC), 2005) including the use of computers as a part of every activity that are related to assessment, such as summative, formative, or diagnostic, as well as online assignments submission in the form of e-portfolio or reflective blog, feedback delivered through an audio file that is recorded in computers. Technology-based assessment or technology enhanced is a new assessment helped by computers, often called by e-assessment.

E-assessment as a partner for e-learning (Mackenzie, 2003) offers a synchronization of teaching and assessment methods (Ashton & Thomas, 2006; Gipps, 2005), provides various improvements in task designs by doing e-portfolios, simulations and interactive games, and skill assessment that is not easily to assess conventionally (JISC, 2010). This provides a solution to students who were learning remotely. E-assessment allows students to understand their weakness (Miller, 2008) and feedback that is considered impersonal (Earley, 1988) and non-judgmental (Beevers et al., 2010). Therefore, the use of digital can be a 'window' (Nicholas et al., 2009) to students' thinking that provides information, both for students and teachers.

The following is an e-assessment that has been developed based on the needs in learning and dance assessment (Triana & Juniasih, 2019) that is used to assess the Javanese dance practice *Golek Clunthang* along with its rubric. Students uploaded videos to the e-assessment, then the assessors gave scores to the assessment menu in the e-assessment, as shown in Figure 1:



E-Assessment

Pembelajaran dan penilaian dalam tari memiliki karkateristik yan
pengetahuan, keterampilan, dan sikap terintegrasi menjadi satu
dipisahkan. Mempelajari tarian tidak hanya secara teoretis saja,
inilah yang membuat evaluasi pada pembelajaran, khususnya pr

Pada web ini guru dengan siswa secara bersama-sama melakuk
terdokumentasi dengan baik, sehingga memudahkan dalam mel
paper and pencil tes, observasi atau pengamatan. Metode obser
kelemahan pada sisi waktu, untuk itu diperlukan bantuan media
agar evaluasi siswa dapat diberi umpan balik dan diberi apresias
dimana penugasan kepada mahasiswa dapat terencana, terjadw

Course: TARI JAWA 115 (ROMBEL 1)
Assignment: Penilai Kedua Tari Golek Clunthang
View all submissions

Nur Ainun
ainunnas2525@gmail.com
Due date: 22 October 2021, 11:59 PM

Kicat nyangkol udhet, kicat ridhong maju mundur, sendhi	Ragam Gerak Tari 0 points		Ragam Gerak Tari 0 points	
Rukun (ragam gerak)	Kurang Baik 1 points	Cukup Baik 2 points	Baik 3 points	Sangat Baik 4 points
Rentak (tempo/ iringan)	Kurang Baik 1 points	Cukup Baik 2 points	Baik 3 points	Sangat Baik 4 points
Resam (Estetika/ gaya)	Kurang Baik	Cukup Baik	Baik 3 points	Sangat Baik

Figure 1. *Evaluation of E-Assessment Dance Practice*

Methods

Participants

Participants involved in this study were 18 of 52 students of the dance education study program who taught the Javanese dance course *Golek Clunthang*, who were asked to do assignments by demonstrating the dance, and the lecturers assessed the practice results done by the students through the e-assessment. Samples were selected dandomly by uploading videos that matched the requirements, which are clarity, video clarity, and shooting distance. The assessors as raters consisted of 2 people who have competence as Javanese dance teachers in non-formal institutions. Assessors were selected from external party so that the assessment could be done objectively, and they had no psychological relationship with the students tested.

Procedure

Data in this study was collected through practice assessment sheets through e-assessment. The sheets contain the criteria or aspects that were assessed in practice assignments and graded by 2 raters. The sheets were then used to assess practice assignment given to the students. Potential raters were filtered by researchers to see whether they fulfilled the criteria to participate. To guide the lecturers in assessing, a rubric in the e-assessment was developed as a guide for raters to do the assessment. Rubrics are used to guarantee the objectivity of assessment. Raters did the assessment by giving a score through observations with a scale of 1 (incorrect), 2 (inappropriate), 3 (precise) and 4 (very precise) based on 1) pillars of *wiraga*, which are the variety of movements that must be done correctly, 2) Rentak or *wirama*, which is the accuracy of movement with tempo / accompaniment, 3) variety of *wirupa*, which is the accuracy of style according to aesthetics, 4) taste of *wirasa*, which is appreciaton or expression that is performed according to the character of the dance.

Data Analysis

The data analysis technique used is descriptive quantitative. The data analysis technique in this study was a reliability technique by looking at the relevance of expert judgment and internal consistency. Other than that, inter-rater approach was used in estimating the reliability of the instrument. Data were checked by the Levene's test for variance equality. Data were exported from e-assessment in a .csv format to Excel to count the score order and the total score for the four combined sequences. The related Intraclass correlation coefficient (ICC) and 95% of Confidence Interval (CI) were calculated in SPSS. Inter-rater reliability was assessed using a two-way mixed effects model for single raters, single measurement scores for each dancer or demonstrator, definition of consistency and checking of all raters. Inter-rater reliability was assessed using a two-way random effects model for multiple raters, a single measurement score for each dancer-demonstrator, definition of consistency and checking of all raters. ICC values are considered low for 0.49, moderate for 0.50 – 0.69, high for 0.70 – 0.89, and very high for 0.90 – 1.00 (Munro, 2005).

Results

The evaluation of dance practice assessment was done with the help of 2 external raters who mastered the Javanese dance *Golek Clunthang and* did not have psychological affinity with the students being assessed. The steps of the assessment were as follows: (1) the perception equation was carried out a day before the assessment was done, the researchers gave the instrument to the ratters and explained the means consisted in the indicator items; (2) Raters were given training and simulations on how to assess through e-assessment; each rater observed the video uploaded by students, then assessed based on the rubric provided in the system, then the rater filled in the assessment items by giving the score of 1, 2, 3, 4 as an evaluation result. This was done so that when the rater conducted an assessment, it could avoid misinterpretation of the assessment items (3) the rater conducted an assessment of students through e-assessment, (4) the researcher downloaded all the results of the assessment that had been carried out by the rater through e-assessment, then calculated the level of agreement of the two raters, (5) the researcher held a discussion with the raters and asked for input on the assessment instrument used.

After the researchers had downloaded all the research result through the e-assessment and counted the level of agreement (reliability) between the two raters, it was obtained by calculating the inter-rater reliability coefficient using the Intra-class Correlation Coefficient. The results of the calculations using the SPSS version 22 program are presented in Table 1.

Table 1. *Intraclass Correlation Coefficient*

	Intraclass Correlation[b]	*95% Confidence Interval*		*F Test with True Value 0*			
		Lower Bound	*Upper Bound*	*Value*	*df1*	*df2*	*Sig*
Single Measures	.543[a]	.116	.800	3.376	17	17	.008
Average Measures	.704[c]	.208	.889	3.376	17	17	.008

Two-way mixed effects model where people effects are random and measures effects are fixed

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition. The between-measure variance is excluded from the denominator variance

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise

Based on the data in table 1, it can be seen that the results of the ICC calculation using the SPSS v.22 analysis results were obtained, which showed that the average agreement between raters was 0.543, while the consistency for raters was 0.704, which means it has high stability (Munro, 2005). Based on the calculation results, it can be explained that raters had a fairly high consistency in evaluating the Javanese dance *Golek* Clunthang.

Table 2 shows the outputs of the ANOVA analysis, and whether there were statistically significant differences between the groups' means. The results seem to indicate that the pairwise comparisons of the mean score raters assigned to the dance practice assessment differed significantly from each other. It can be seen clearly that the significance level is 0.000, placed below 0.001 ($p < 0.001$). Therefore, there is a significant difference statistically in the average score given by different assessors. This may mean that there is a contrast difference between scores given by the assessors for the same essay product and the reliability between assessors is very low. The result uses the help from Version 22 of SPSS Program, presented in Table 2.

Table 2. *ANOVA analysis*

		Sum of Squares	*df*	*Mean Square*	*F*	*Sig*
Between People		3029.324	17	178.196		
Within People	Between Items	10.606	1	10.606	8.201	.001
	Residual	897.356	17	52.786		
	Total	907.962	18	50.442		
Total		3937.285	35	112.494		

Discussion

In the dance practice assessment, the total of inter-rater scores were highly correlated on the retests, but the individual order correlation was lower, ranging from 0.543 – 0.704. It is possible that the results between were increased, if the previous raters observed the practice result before doing the assessment for the second time. High results between raters shows that the observation of dance assessment was kept fresh in the raters' mind during the first assessment. It was in line with the explanation of inter-rater reliability (Scheel et al., 2018), that inter-rater reliability refers to the consistency of data recorded by two or more raters, measuring the same subject during one experiment. Inter-rater reliability helps in determining whether the measurement tool provides support and confidence in assessing dance practice.

The presence of unskilled raters is often inconsistent in reassessment and will affect the reliability of inter-raters. Therefore, using e-assessment will help, because the advantages of assessing dance using e-assessment is that the assessor can re-assess with the same results. E-assessment technically makes it easier to conduct an assessment. However, it is highly recommended to use more than 2 raters.

This difference can also be observed by examining the inter-rater consensus between raters. It seems that different weighting for each sub-criterion could result in more consistent assessments as stated by raters, because the results of different correlation coefficients obtained using the Fischerz transform also support the idea that inter-rater scores are similar but not equal. Pairwise comparison with ANOVA tells us that inter-rater scores are never significantly the same. This means that different scores were given to the practice assessment at different times. If a lower score is given for the same practice assessment in two different sessions around the boundary score, it obviously means that success and failure depend on the source of variation. At this point, raters and the time elapsed between assessments may appear to be variation sources.

The coefficient G also indicates that scoring is more precise and effective when a scale is used, as it increases reliability between raters. Considering some limitations, further research on the effectiveness and usefulness of the scale was carried out because it was difficult to conclude what process the raters went through when they were assessing dance practice through e-assessment. The more information there is, the more reliable it is, because more data will be retrieved (Zlatkin-Troitschanskaia et al., 2019).

The description above shows that the use of e-assessment in *Golek Clunthang* dance can be done with ICT-based assessment, and this will certainly help make it easier for assessors or teachers when assessing conventional dance practices towards digital / paperless-based assessments.

Conclusion

This study proves that practice assessment through e-assessment can minimize subjectivity and seems to have to pay attention to the internal responses of assessors with different functions or ways in using dance practice assessment criteria. Statistical analysis clearly shows that the assessment of dance practice through e-assessment is more reliable and consistent. The correlation coefficient is higher and is supported by the rater himself, as seen in the SPSS data.

However, when the total score and rater assessment results are examined, it can be seen clearly that the scores are different from each other even though the correlation coefficient is high and significant. It may be more accurate if Kline's (1986) cut-off coefficient (0.70) for meaningful correlation could be increased to 0.90, at least to provide more reliable evidence or ranking. This means that the scores given are more similar or closer to each other. The training provided and the agreement of the assessors before the assessment process for each group of students also seems to need to equalize perceptions of the criteria and indicators of assessment. In order to obtain verbal descriptions as concrete information, it is best to understand this process, and to define the decision-making process of the rater, systematic steps need to be considered in conducting a follow-up assessment.

In terms of inter-rater reliability in assessing dance practice through e-assessment, it is considered reliable in the high category. Some of the weaknesses that showed up during the research process, such as the criteria that became the object of assessment and were stated in the dance assessment sheet, were not made based on in-depth theoretical studies. Other than that, there is no further validity test from the new dance practice assessment sheet that has been done. In response to these weaknesses, the following suggestions are: (a) in developing the criteria or aspects of the assessment, it should be derived from an in-depth theoretical study. With a deep and strong theory, valid operational definitions will be produced on each criterion or aspect; (b) when the dance assessment sheet is obtained, the results of the rotation of several aspects must be further tested for validity. This way, the validity of the new instrument will be known; (c) lecturers or education practitioners must formulate a brief and clear descriptor on the assessment rubric, (d) to minimize the element of subjectivity in the assessment of dance practice, it is necessary to have inter-rater consistency so that the scores can be accounted for. With a descriptor that is too long, it often confuses raters, and in the end the rater will give the wrong assessment.

Acknowledgement

The researcher would like to thank the Dean of Faculty of Language and Arts, Univeristas Negeri Jakarta, who has given full support to complete this research.

References

Adnan, K., & Bulut, O. (2014). Crossed random-effect modeling: examining the effects of teacher experience and rubric use in performance assessments. *Eurasian Journal of Educational Research*(57), 1-28. https://doi.org/10.14689/ejer.2014.57.4

Alexiadis, D. S., & Daras, P. (2014). Quaternionic Signal Processing Techniques for <newline/>Automatic Evaluation of Dance Performances <newline/>From MoCap Data. *IEEE Transactions on Multimedia*, *16*(5), 1391-1406. https://doi.org/10.1109/TMM.2014.2317311

Ali, S. A. B., Ahmad, M. N., Zakaria, N. H., Arbab, A. M., & Badr, K. B. A. (2018). Assessing Quality of Academic Programmes: Comparing Different Sets of Standards. *Quality Assurance in Education: An International Perspective*, *26*(3), 318-332. https://doi.org/10.1108/QAE-09-2016-0051

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research, and Evaluation*, *10*, 3. https://doi.org/10.7275/g367-ye94

Aristidou, A., Stavrakis, E., & Chrysanthou, Y. (2014). Motion Analysis for Folk Dance Evaluation. GCH, Darmstadt, Germany. 55-64. https://dl.acm.org/doi/abs/10.5555/2854922.2854931

Ashton, H., & Thomas, R. (2006). Bridging the gap between assessment, learning and teaching. Loughborough University, Leicestershire, Leicestershire. https://caaconference.co.uk/pastConferences/2006/proceedings/ashton.pdf

Beevers, C. G., Clasen, P., Stice, E., & Schnyer, D. (2010). Depression symptoms and cognitive control of emotion cues: a functional magnetic resonance imaging study. *Neuroscience*, *167*(1), 97-103. https://doi.org/10.1016/j.neuroscience.2010.01.047

Clements, D. H., Sarama, J., & DiBiase, A.-M. (2003). *Engaging young children in mathematics: Standards for early childhood mathematics education*. Routledge. https://doi.org/https://doi.org/10.4324/9781410609236

Creswell, J. W., & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Inc. https://us.sagepub.com/en-us/nam/research-design/book255675

Donnelly, R. (2006). Teaching Portfolios. In *Case Studies of Case Studies of Good Practice in the Assessment of Student Learning in Higher Education*. HEA. https://arrow.tudublin.ie/ltcbk/10/

Earley, P. C. (1988). Computer-generated performance feedback in the magazine-subscription industry. *Organizational Behavior and Human Decision Processes*, *41*(1), 50-64. https://doi.org/10.1016/0749-5978(88)90046-5

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185. https://doi.org/10.1177%2F0265532207086780

Enghauser, R. G. (2012). Tracing lines of meaning: A course redesign for dance pedagogy. *Journal of Dance Education*, *12*(2), 54-61. https://doi.org/10.1080/15290824.2012.621401

Englebright, K., & Mahoney, M. R. (2012). Assessment in elementary dance education. *Journal of dance education*, *12*(3), 87-92. https://doi.org/10.1080/15290824.2012.701176

Ginsburg, H. P., Lee, Y.-S., & Pappas, S. (2016). A research-inspired and computer-guided clinical interview for mathematics assessment: Introduction, reliability and validity. *ZDM*, *48*(7), 1003-1018. https://doi.org/10.1007/s11858-016-0794-8

Gipps, C. V. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education*, *30*(2), 171-180. https://doi.org/10.1080/03075070500043176

Gross, D. P., & Battié, M. C. (2002). Reliability of safe maximum lifting determinations of a functional capacity evaluation. *Physical Therapy*, *82*(4), 364-371. https://doi.org/10.1093/ptj/82.4.364

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC. https://books.google.com.pk/books?id=fac9BQAAQBAJ

Hachimura, K., Kato, H., & Tamura, H. (2004). A prototype dance training support system with motion capture and mixed reality technologies. RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759), Kurashiki, Okayama, Japan. 217-222. https://doi.org/10.1109/ROMAN.2004.1374759

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23-34. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Harding, M. (2012). Assessment in the high school technique class: Creating thinking dancers. *Journal of Dance Education*, *12*(3), 93-98. https://doi.org/10.1080/15290824.2012.701172

Hernandez, B. (2012). The case for multiple, authentic, evidence-based dance assessments. *Journal of Physical Education, Recreation & Dance*, *83*(1), 5-56. https://doi.org/10.1080/07303084.2012.10598700

JISC, E. (2010). Effective assessment in a digital age-a guide to technology-enhanced assessment and feedback. In *Technology enhanced Assessment* (pp. 26-28).

Joint Information Systems Committee (JISC). (2005). *Effective Practice with e-Assessment: An overview of technologies, policies and practice in further and higher education*. Joint Information Systems Committee.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement*, *38*(4), 319-342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Krasnow, D., & Chatfield, S. J. (2009). Development of the "performance competence evaluation measure": assessing qualitative aspects of dance performance. *Journal of Dance Medicine & Science*, *13*(4), 101-107. https://europepmc.org/article/med/19930811

Lange, R. T. (2018). Inter-rater Reliability. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of Clinical Neuropsychology* (pp. 1844-1844). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-319-57111-9

Laraba, S., & Tilmanne, J. (2016). Dance performance evaluation using hidden Markov models. *Computer Animation and Virtual Worlds*, *27*(3-4), 321-329. https://doi.org/10.1002/cav.1715

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language profficiency. *English for Specific Purposes*, *17*(4), 347-367. https://doi.org/10.1016/S0889-4906(97)00016-1

Mackenzie, D. (2003). Assessment for E-Learning: What are the features of an ideal e-assessment system? http://library.oum.edu.my/oumlib/node/211545

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282. https://hrcak.srce.hr/89395

Miller, D. (2008). *The Comfort of Things*. Polity Press. https://books.google.com.pk/books?id=65NMV1ZgsNgC

Mueller, J. (2005). The authentic assessment toolbox: enhancing student learning through online faculty development. *Journal of Online Learning and Teaching*, *1*(1), 1-7. https://jolt.merlot.org/documents/VOL1No1mueller.pdf

Munro, B. H. (2005). *Statistical Methods for Health Care Research*. Lippincott Williams & Wilkins. https://books.google.com.pk/books?id=a34z_Ah2-LgC

Nicholas, D., Huntington, P., Jamali, H. R., Rowlands, I., & Fieldhouse, M. (2009). Student digital information-seeking behaviour in context. *Journal of Documentation*, *65*(1), 106-132. https://doi.org/10.1108/00220410910926149

Pang, X. (2020). Research on the Application of Internet Technology in Dance Teaching. 2020 2nd International Conference on Applied Machine Learning (ICAML), Changsha, China. 204-207. https://doi.org/10.1109/ICAML51583.2020.00049

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment*. Futurelab. http://www.futurelab.org.uk/resources/publications-reports-articles/literature-reviews/Literature-Review204

Scheel, C., Mecham, J., Zuccarello, V., & Mattes, R. (2018). An evaluation of the inter-rater and intra-rater reliability of OccuPro's functional capacity evaluation. *Work*, *60*(3), 465-473. https://doi.org/10.3233/WOR-182754

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Shafer, W. D., Swanson, G., Bene, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, *14*(2), 151-170. https://doi.org/10.1207/S15324818AME1402_3

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33. https://doi.org/10.2307/329895

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189. https://doi.org/10.3102%2F0034654307313795

Smit, R., Bachmann, P., Blum, V., Birri, T., & Hess, K. (2017). Effects of a rubric for mathematical reasoning on teaching and learning in primary school. *Instructional Science*, *45*(5), 603-622. https://doi.org/10.1007/s11251-017-9416-2

Steven, C., & Hesketh, I. (2013). Increasing learner responsibility and support with the aid of adaptive formative assessment using QM designer software. In S. Brown, Bull, Joanna, Race, Phil, (Ed.), *Computer-assisted Assessment of Students*. Routledge. https://doi.org/https://doi.org/10.4324/9780203062340

Stiggins, R. (2006). Assessment for learning: A key to motivation and achievement. *EDge: the latest information for the education practitioner*, *2*(2), 1-19.

SUMARYONO, E. (2003). *ETIKA PROFESI HUKUM : Norma-Norma Bagi Penegak Hukum*. Kanisius. http://perpustakaan.uin-antasari.ac.id/ucs/index.php?p=show_detail&id=27452

Triana, D. D., & Juniasih, I. (2019). IT-Based Movement Evaluation System in Dance Studios. International Conference on Arts and Design Education (ICADE 2018), Bandung, West Java, Indonesia. 224-228. https://dx.doi.org/10.2991/icade-18.2019.52

Vendlinksi, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *The Journal of Technology, Learning and Assessment*, *1*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1669

Warburton, E. C. (2011). Of meanings and movements: Re-languaging embodiment in dance phenomenology and cognition. *Dance Research Journal*, *43*(2), 65-84. https://doi.org/10.1017/S0149767711000064

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, *26*(2), xiii-xxiii. https://www.jstor.org/stable/4132319

Wibowo, F. (1981). *Mengenal tari klasik gaya Yogyakarta*. Dewan Kesenian Prop-DIY, Proyek Pengembangan Kesenian DIY, Department P. & K.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wulan, A. R. (2008). Skenario baru bagi implementasi asesmen kinerja pada pembelajaran sains di Indonesia. *Jurnal Pendidikan*, *32*(3), 1-10. http://file.upi.edu/Direktori/FPMIPA/JUR._PEND._BIOLOGI/ANA_RATNAWULAN/Skenario_baru_asesmen_kinerja.pdf

Zhang, L.-Y., Liu, S., Yuan, X., & Li, L. (2019). Standards and Guidelines for Quality Assurance in the European Higher Education Area: Development and Inspiration. Proceedings of the International Conference on Education Science and Development (ICESD 2019), Shenzhen, China. 19-20. https://doi.org/10.12783/dtssehs/icesd2019/28072

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology*, *89*(3), 468-484. https://doi.org/10.1111/bjep.12286