Revista Electrónica de LEEME, (2023) (52) (pág.18-pág.38)

revista electrónica DE LEEME
Electronic Journal of Music in Education
Revista arbitrada de investigación y aplicaciones en Educación Musical
Peer-reviewed journal of research and applications in Music Education
ISSN:1575-9563

ARTICLES

Psychometric properties and standard-setting study of the Piano Performance Test for prospective teachers

Propiedades psicométricas y estudio de establecimiento de normas del Test de Interpretación Pianística para futuros docentes

Salim Sever[1]
Department of Elementary Education, Ankara University, Ankara (Turquía)
C. Deha Dogan[2]
Department of Educational Evaluation and Measurement, Ankara University, Ankara (Turquía)
Omer Kamis[3]
Department of Educational Evaluation and Measurement, Çankırı Karatekin University, Çankırı (Turquía)
Gulsah Sever[4]
Department of Arts Education
Division of Music Education, Gazi University, Ankara (Turquía)

doi:10.7203/LEEME/52.27217
Reception: 01-08-2023 Revision: 08-08-2023 Acceptance: 25-09-2023

Abstract

The purpose of this study was to investigate the psychometric properties of the Piano Performance test for Elementary School Teachers and to undertake a standard-setting study for this scale. This study included three groups of participants: students (n=100), raters (n=2) used to test the psychometric features of the musical instrument performance test and experts (n=6) for the standard-setting study of the test. In this study, the researchers developed a music performance test and analytical rubric. The results showed that the one-factor structure was appropriate for the musical instrument performance test, which explained 66% of the total variance. The Cronbach Alpha coefficient showed that the internal consistency of the scale was acceptable (.83). Moreover, generalizability studies and the intra-class correlation coefficient indicated excellent rater reliability for the scale. The results of the item discrimination analysis show that the musical instrument performance test is capable of discriminating participants who had high and low levels of ability to play the piano.

Key words: Music-education; piano performance; pre-service teachers; inter-rater reliability; rubric; standard-setting.

Resumen

El objetivo de este estudio era investigar las propiedades psicométricas del test de Interpretación de Piano para Profesores de Enseñanza Primaria y realizar un estudio de fijación de estándares para esta escala. Este estudio incluyó tres grupos de participantes: estudiantes (n=100), calificadores (n=2) utilizados para comprobar las características psicométricas de la prueba de rendimiento en instrumentos musicales y expertos (n=6) para el estudio de fijación de estándares de la prueba. En este estudio, los investigadores desarrollaron una prueba de interpretación musical y una rúbrica analítica. Los resultados mostraron que la estructura de un factor era adecuada para la prueba de ejecución de instrumentos musicales, que explicaba el 66% de la varianza total. El coeficiente Alfa de Cronbach mostró que la consistencia interna de la escala era aceptable (.83). Además, los estudios de generalizabilidad y el coeficiente de correlación intraclase indicaron una excelente fiabilidad de la escala por parte de los evaluadores. Los resultados del análisis de discriminación de ítems muestran que la prueba de interpretación de instrumentos musicales es capaz de discriminar entre participantes con niveles altos y bajos de habilidad para tocar el piano.

Palabras claves: Educación musical; interpretación pianística; profesores en formación; fiabilidad entre evaluadores; rúbrica; establecimiento de normas.

[1] Professor of Musicology, Faculty of Educational Sciences, https://orcid.org/0000-0003-4028-4514
* Contact and correspondence: Salim Sever, Department of Elementary Education, Ankara University, salimsvr@gmail.com, Cemal Gürsel Cad. Ankara Üniversitesi Cebeci Yerleşkesi Çankaya, 06590 Ankara. Turquía.
[2] Associate Professor of Educational Evaluation and Measurement, Faculty of Educational Sciences, https://orcid.org/0000-0003-0683-1334
[3] Lecturer, Educational Evaluation and Measurement, Faculty of Education., https://orcid.org/0000-0003-0605-087X
[4] Associate Professor of Music Education, Faculty of Educational Sciences, https://orcid.org/0000-0003-0559-6993

1. Introduction

Music education is an important aspect of childhood development, and including music lessons in elementary and preschool teachers' curricula can have a significant impact on children's learning outcomes (Swanwick, 2016). Research showed that music education can enhance children's language development, cognitive skills, and social-emotional well-being (Schellenberg, 2018). Furthermore, incorporating music into the preschool curriculum can help to promote creativity and foster a love of music in young children (Standley, 2016). Thus, it is essential for elementary and preschool teachers to receive training in music education and to incorporate music lessons into their daily routines in order to support the holistic development of their students. Therefore, according to Scripp and Kaufman (2019), it is essential that elementary and preschool teachers have a strong foundation in music, including piano skills. The piano is one of the fundamental instruments that provides a solid foundation for music education to provide effective music instruction for young children. Piano training can improve teachers' music literacy, provide opportunities for creative expression, and enhance their ability to integrate music into the classroom curriculum (Robinson, 2016; Guven, 2020). Scripp and Kaufman (2019) stated that piano instruction can help preschool teacher candidates gain proficiency in music theory, ear training, and performance skills, which can be applied to teaching young children. According to the research of Price and Burnsed (1989), the most needed skills in preschool and primary education are singing and playing. Piano (keyboard), is capable of creating different sounds, allows accompaniment, and lets teachers use their voice as singing or leading while playing simultaneously. By learning to play the piano, preschool and elementary school teachers gain a solid foundation in music theory, harmony, and technique, which enables them to effectively teach and guide young children in their musical development (Lee, 2009). In this regard, it is an effective musical instrument to be used in elementary school and preschool music courses.

Assessment is an integral part of the teaching process. Without a valid and reliable assessment, we cannot be sure about the quality and effectiveness of the teaching process. Piano performance assessment plays a critical role in evaluating piano performers' technical abilities, interpretative skills, and overall musicality. According to Boud (1995) performance assessment enables players to receive feedback on their musical strengths and weaknesses, as well as to identify areas for improvement. Performance assessments also provide valuable information for music teachers to adjust their instructional strategies to meet the needs of their students (Elliott, & Silverman, 2014).

Piano teachers have different scoring habits when evaluating their students' piano performance. Some teachers may use a more lenient scoring system to encourage their students, while others may use a stricter scoring system to motivate their students to improve. Research has shown that structured assessment methods, such as exams and competitions, can have a positive impact on students' motivation and self-esteem (Mak, & Fancourt, 2019). These evaluations also offer a way to assess and compare performance levels across different students and institutions, providing a common standard for excellence in piano playing (Kim *et al.*, 2021).

Different measurement tools and models have been developed to assess music performance (Abeles, 1973; Boyle, & Radocy, 1987; Watkins, & Farnum, 1954; Nichols, 1991; Palmer, 1996; Stanley *et al.*, 2002; Zdzinski, & Barnes, 2002; Juslin, 2003; Wapnick *et al.*, 2004; Wapnick *et al.*, 2005; Russell, 2010; Alessandri *et al.*, 2016). A primary difficulty with music performance assessment is managing its subjective nature. Rubrics can be used as a set of guidelines to objectively assess a student's performance (Wesolowski, 2012). A rubric is a coherent set of criteria that includes descriptions of levels of performance quality across a range of dimensions. An analytical rubric specifies expectations for a given task by dividing it into its parts and providing a detailed description (Stevens, & Levi, 2005). Rubrics for piano performance might include criteria such as rhythmic accuracy, pitch accuracy, tempo, accent, dynamics, tone quality, interpretation, and overall quality (Duerksen, 1972), technical (tone, intonation, rhythmic accuracy, articulation, and technique) and musical (tempo, dynamics, timbre, interpretation and musical expression) (Russell, 2010), right-hand expression, phrasing, dynamics, rubato, form/structure, tone balance, pedaling, attention to rhythm and meter, articulation, technical competence, tempo, expression of several parts (Thompson *et al.*, 1998). Although the rubrics created for the performance contain different evaluation items, they can be grouped under three main headings: technical competence, musical interpretation, and stage presence. By using a rubric, teachers can provide more detailed feedback to students and help them identify areas for improvement (McMillan, & Hearn, 2008).

When assessing music performance, consistency is crucial, and using analytical rubrics increases inter-rater reliability. In order to test the reliability of an assessment tool, various statistical techniques can be applied based on various test theories. One of the primary assessment models is classical test theory (CTT). In CTT, reliability is defined in terms of internal consistency and stability (Urbina, 2004). The Cronbach's Alpha coefficient is typically used as the measure of internal consistency (Crocker, & Algina, 1986). The Cohen Kappa and interclass correlation techniques also provide information about inter-rater reliability (Gliner *et al.*, 2009; Howell, 2013).

The other testing construct is generalizability theory (GT). If more than one rater is used in assessing a student's musical performance, the variance of rater facet (variable) provides information about rater reliability. G and Phi coefficients also provide information about the reliability of the measurement process (Shavelson, & Webb, 1991). Thus, assessing piano performance using both classical test theory and generalizability theory might provide strong evidence for inter-rater reliability.

The standard grading measure is concerned with establishing a cut-off score that separates the competent from the incompetent student on a test or performance-based assessment (Ben-David, 2000). A cut-off score is simply the score that serves to delineate one level from another. If the cut-off scores are not appropriately set, the results of the assessment could come into question. For that reason, standard setting is a critical component of the test development process (Bejar, 2008). It is vital to establish cut-off scores for music performance assessments.

There are various approaches to setting standards for both written and performance-based assessment (Southgate *et al.*, 2001). A large number of methods have been developed and are used to set standards for both written and performance-based examinations (Cizek, & Bunch, 2007). Playing the piano is also a musical performance that can be measured; therefore, it is treated as a standard setting subject.

Among these methodological frameworks, Angoff is a commonly used method for standard-setting which requires expert raters. This approach is based on the concept of the borderline or minimally competent student. In other words, it is based on the determination of the student who performs the minimum acceptable playing performance according to the scoring made by two or more experts. The extended Angoff method is a parallel approach for polytomously-scored items (Cizek, & Bunch, 2007). Extended Angoff is one of the most appropriate methods to define a cut-off score for piano performance scale for elementary and preschool teachers.

In literature there are some studies focusing on inter-rater reliability of musical performance tests (Boyle, & Radocy, 1987; Thompson *et al.*, 1998; Thompson, & Williamon, 2003). Also, Bergee (2003, 2007) used G Theory test inter-rater reliability of musical performance. Wrigley and Emmerson (2011) used the structural equation model in the reliability study of the performance rubric they developed for all musical instruments.

In evaluating the student's performance, jury points were taken into consideration. Musical performance is multi-layered and complex. For that reason, evaluation is often forced into a personal and emotional nature. It has been recognized that examiners who rated the same performance may hold different constructs, can weigh the same constructs differently, or may not hold any conscious constructs but rely on intuition or a gut feeling and, as a result, have difficulty articulating how they arrive at a performance rating (Boyle, & Radocy, 1987; Thompson *et al.*, 1998). Being an expert to evaluate is also an important issue as poor inter-rater consistency has resulted when raters who are novices in scoring the assessment define a given construct differently from those more familiar with the tool (Thompson, & Williamon, 2003). Moreover, there is some research investigating the severity of raters and rater presicion using Many-Facet Rasch Partial Credit Measurement Model (Wesolowski, 2012, Wesolowski *et al.*, 2015, 2016).

The passing grade may vary from institution to institution, but in most cases, it is between 60 and 70 out of 100. Can a teacher, playing at the 60-point level, convey the musical piece to the students? Which point level is good enough to convey the music to the students? In other words, what should be the rubric score equivalent of the minimum listenability level of a piano performance? To examine this, data were collected through a rubric that measures the dimensions of rhythm, melody, harmony, technique and tempo. The main purpose of this study is to determine the passing grade in the assessment made with a rubric with 5 criteria, each of which is scored between 1 and 5.

In the relevant literature, no studies were found that focus on developing a piano performance scale for elementary and pre-service teachers – notwithstanding that this would be a quite useful evaluative measure for use in elementary and preschool music courses. Creating such an instrument and testing its psychometric properties (validity and reliability) will thus contribute to the literature. This scale can be used to assess the piano performance of pre-service teachers on educational faculties.

Moreover, it has been seen that many factors impact rubric scores, having largely to do with issues of validity and reliability. The use and comparison of various statistical tools is considered important in measuring inter-rater reliability. Using both classical test theory and generalizability theory together will help researchers to assess validity and reliability problems more effectively. However, in the literature, there are very few studies using a combination of these approaches to test inter-rater reliability. This is why different methods based on CTT and G Theory were used together in this research. Besides, in the literature, there is no research aimed at defining cut-off scores for piano performance for preschool and elementary school prospective teachers.

The present study adds a statistical perspective on previous research in the field by examining the psychometric features and also includes the standard-setting study for the Piano

Performance Test for Pre-Service Teachers (PPT-PT). In this manner, it presents a valid and reliable way of assessing piano performance using rubrics.

The current study is aimed at investigating the psychometric properties and defining the cut-off score of the PPT-PT based on different measurement theories. The pertinent research questions are:

1. What is the construct validity of the PPT-PT? (Does PPT-PT effectively measure its intended construct?)
2. What is the reliability of the PPT-PT based on the CTT and GT? (What degree of error is involved in PPT-PT scores in the context of CTT and GT?)
3. What is the inter-rater reliability based on the CTT and GT? (What is the level of agreement among the independent raters scoring the PPT-PT?)
4. What is the cut-off score of the PPT-PT based on the extended Angoff method? (What is the minimum score that students should get to be successful in PPT-PT? According to the Extended Angoff method?)

2. Method

2.1. Design

This descriptive research aims to investigate the psychometric properties of an evaluative instrument, the PPT-PT. The current study also aims to define a cut-off score for PPT-PT. In educational research, researchers summarize the characteristics of individuals, groups, measurement tools, etc. with no intervention in the process (Fraenkel *et al.*, 2015); thus, in this study, the psychometric features and standards of the PPT-PT were described without any intervention.

2.2. Participants

The study participants consisted of three groups: students, raters used to test the psychometric features of the PPT-PT, and experts for the standard-setting study of the test.

Table 1. Demographic properties of the participant subjects

Participants	N	Grade
Elementary education program	50	Second
Pre-school education program	50	Third
Total	100	

Study Group 1 was comprised of 100 undergraduate students enrolled in the Preschool and Elementary Education Departments of a state university in Ankara. The participants were selected using convenience sampling. For practical reasons, it was not possible to use random sampling.

Study Group 2 consisted of two independent raters instructing music-education courses employed in different state universities in Ankara. One of the raters is a professor in the music-education field and has been instructing various undergraduate and graduate courses, including music theory, music-education, and music sociology, for 19 years. The other rater is an associate professor in music-education and has been instructing individual and group violin undergraduate and graduate courses for 12 years. Both raters are also experienced in scoring musical instrument performances using various evaluative rubrics.

Group 3 comprised six experts from the field of music-education and performance, who were asked to estimate borderline examinee performance for each of the criteria in the PPT-PT (Cizek, & Bunch, 2007).

2.3. Instruments

For this study, the researchers developed a piano performance test and analytical rubric. The performance test required the students to play tunes on the piano. The piano piece included in this research was taken from the popular tunes taught in music courses in elementary schools.

The piano piece was in 4/4 meter and 32 measures. It was composed in the scale of D minor using six consecutive notes between C and A. The left-hand part included tonic and dominant chords in the main and first conversion positions. To assess students' performance, the raters used a rubric based on five criteria (tempo, rhythm, melody, harmony, and technique), with each criterion scored out of 5 points total. The piano performance analytical rubric (Appendix A) was developed considering the steps displayed in Table 2 (Andrade 2000; Haladyna, 1997; Kutlu *et al.*, 2014; Moskal, 2000).

Table 2. Rubric development process

Steps	Procedure
1	Defining the features/sub-dimensions to be assessed
2	Defining the range of scores for each sub-dimension
3	Defining the performance indicators
4	Forming the draft form of an analytical rubric
5	Taking expert opionion
6	Forming the final draft of the analytical rubric

The standard-setting expert form, developed by the researchers, consists of two parts and is administered, correspondingly, in two rounds. In the first round, the experts rate the borderline examinee performance independently for each of the five criteria. Then the experts revealed their estimates of the borderline examinee performance and discussed any inter-rater disagreement. In the second round, the experts rate the borderline examinee performance regarding the previous discussion and fill out the second part of the form (Appendix B).

2.4. Collection of data

For the data collection, the students performed the piano piece and each rater scored the students' performances independently. Since all students performed the same musical piece and each rater scored each student's performance, a fully crossed design was achieved (PxIxR). GT also allows for nested designs, in which different groups of raters graded different groups of students (Brennan, 2001). But in this study, a fully crossed design was used, with every student in the study evaluated by the same set of raters. The data structure example of the fully crossed design is presented in Table 3.

Table 3. Example of the data structure for a fully crossed design

	Item 1		**Item 2**		. . .	**Item 5**	
	R1	**R2**	**R1**	**R2**		**R1**	**R2**
P1	X	X	X	X	. . .	X	X
P2	X	X	X	X	. . .	X	X
.	.	.	.	.	. . .	.	.
.	.	.	.	.	. . .	.	.
P100	X	X	X	X	. . .	X	X

R: Rater, P: Person.

Each rater scored the students' performance individually and independently. This process took approximately ten minutes for each student and was completed in ten days.

The data collection of the standard-setting study was completed in the following five stages. In Stage 1, the experts were instructed as to what the standard-setting comprised and explained the extended Angoff method used in the study. In Stage 2, the experts discussed the competencies of borderline students concerning their musical performance. In Stage 3, the experts defined the borderline examinee performance for each criterion in the PPT-PT individually using the 'standard-setting' expert form. In Stage 4, the experts revealed their estimates of the borderline examinee performance and discussed any inter-rater disagreement. In the final stage, the experts redefined the borderline examinee performance for each of the criteria (considering what they had discussed in the previous section). Data collection for the standard-setting study session was completed in approximately 2 hours. Approval was obtained from the Ethics Committee of the Ankara University (ID: 85434274-050.04.03/2022).

2.5. Data analysis

Analysis for Validity. To define the construct validity of the Scale, exploratory and confirmatory factor analyses were calculated. In order to define item discrimination, the Mann-Whitney U test was conducted to compare the upper and lower 27% of participants.

Analysis for Reliability. To test inter-reliability, the Cronbach's Alpha coefficient was calculated based on CTT. For testing inter-rater reliability, the Cohen Kappa and interclass correlation coefficients were utilized based on CTT.

The GT framework was also utilized in determining rater reliability; the rater variance components were examined. For the reliability analysis, the G and Phi coefficients were calculated based on the G theory. For data analysis, EduG 6.1, SPSS 21.0, and LISREL 7.0 software packages were used.

Data Analysis for Standard-setting. Following the extended Angoff method, the PPT-PT and descriptive statistics were computed for the data collected from the experts.

3. Results

3.1. Findings on construct validity

The construct validity of the scale was assessed applying exploratory and confirmatory factor analyses. For practical reasons, it was not possible to apply the analyses to two different groups of participants because the sample size was insufficient. This is why exploratory and confirmatory factor analyses were computed based on the same group; this may be considered a limitation of the study. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) comprise multivariable statistics and require some assumptions and sample sizes; however, for the data based on the students' performance (such as musical instrument performance), it was not practical to obtain large sample sizes. In the present study, the sample size was 100; considering the number of items in the performance test (five items), this seemed sufficient for EFA and CFA (Barret, & Kline, 1981).

Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) was utilized; the KMO value was calculated as .83. According to Leech et al. (2005), a KMO value of .80–.90 and higher can be interpreted as very good. Bartlett's test verifies whether our correlation matrix is different from the unit matrix. Therefore, a statistically significant Bartlett's test (rejection of the H_0 hypothesis) means that the correlations between the variables differ from zero. In this study, the Bartlett's test was found to be statistically significant (*p* < .05). Both KMO and Bartlet's sphericity test results show that the data are suitable for factor analysis calculation. Thus, the data set is suitable for factorization. In other words, it can be stated that the criteria in the rubric (rhythm, tempo, melody, harmony, technique) have a sufficient level of correlation with each other to form a factor. In the current study, there was no missing value in the research data.

The results of EFA indicated that the scale had a one-factor structure. In total, five items explained 63% of the total variance. To decide on the number of factors, the scree plot and Horn’s parallel analysis were examined. Both showed that a one-factor structure was appropriate for the PPT-PT (Figure 1).

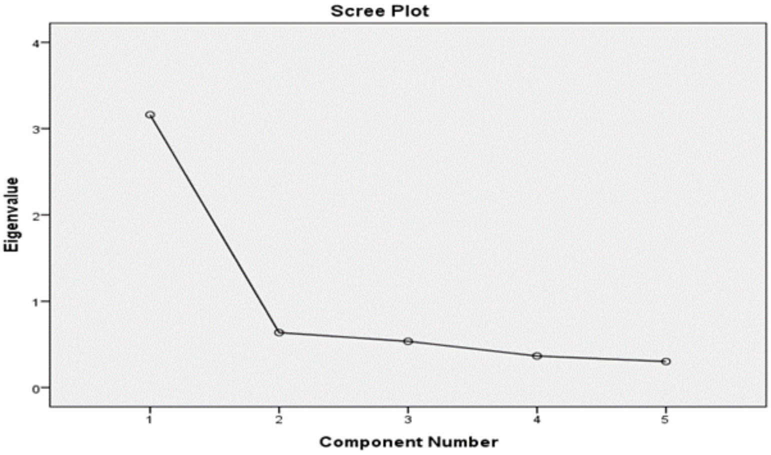

Figure 1. *Scree Plot for the PPT-PT*

The fact that there is a sharp decline in the graph after the first eigenvalue and that it flattens out and loses its slope indicates that a one-factor structure is dominant. In other words, the decline and flattening of the graph indicates that the criteria (tempo, rhythm, melody, harmony, technique) in the rubric are components of a single construct. The factor loadings for the items of the PPT-PT were presented in Table 4.

Table 4. Factor loadings for the PPT-PT

Criteria	***Loadings***
	Factor 1
Tempo	.845
Melodic accuracy	.820
Rhythmic accuracy	.788
Harmonic accuracy	.779
Fingering	.739

As shown in Table 4, all factor loadings were between .74 and .85. This shows that all items had a strong relationship with the measured construct. Therefore, there was no need to omit any of the items since, according to Tabachnick and Fidell (2007), factor loadings above .40 are good enough to keep the items in the scale.

The CFA was utilized to confirm the one-factor structure of PPT-PT. Figure 2 shows the standardized coefficients from the latent variable to the observed variables and t values.

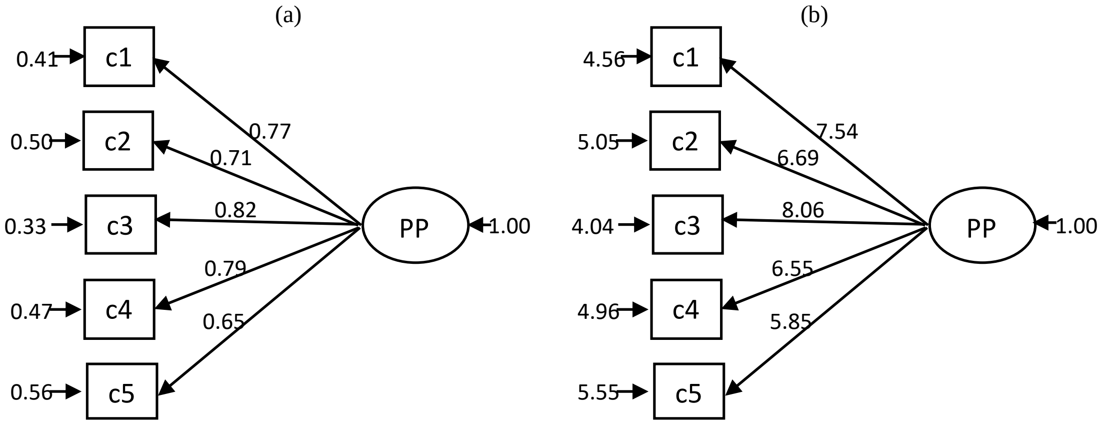

Figure 2. Standardized coefficients and t values for the relationship between the latent and observed variables

As shown in Figure 2(a), the standardized coefficients for the relationship between the latent and observed variables were between .65 and .82, respectively. Figure 2(b) shows that the t values for the relationship between latent and observed variables were above the critical ratio (2.58) and were statistically significant at the .01 level (the cut-off points for the fit indices are presented in Table 5).

Table 5. Cut-Off points for the Fit Indices

Fit Indices	Cut-off Points
Chi cuadrado / df	Chi cuadrado /df ≤2.5 excellent fit; ≤5 mediocre fit
GFI-AGFI- CFI- NNFI	GFI-AGFI- CFI- NNFI ≥ 0.95 excellent fit; ≥0.90 good fit
SRMR-RMSEA	SRMR-RMSEA ≤0.05 excellent fit; ≤0.08 good fit; ≤0.010 poor fit

Nota. df: degrees of freedom, GFI: Goodness of fit index, AGFI: Adjusted goodness of fit index, CFI: Comparative fit index, NNFI: Non-normed fit index, SRMR: Standardized root mean square residual, RMSEA: Root mean square error of approximation (Tabachnick y Fidell, 2007).

The ratio of the degrees of freedom was calculated as 3.937/5 = 0.89. This result can be considered as an indicator of an excellent fit (Tabachnick, & Fidell, 2007). The results of the other fit indices were as follows: non-normed fit index (NNFI): .96; comparative fit index (CFI): .98; goodness of fit index (GFI): .95; root mean square error of approximation (RMSEA): .11; and standardized RMR: .04.

The NNFI, CFI, and GFI values, being above .95, were the indicators of an excellent fit (Hooper *et al.,* 2008; Sümer 2000). The RMSEA and standardized RMR values below .05 indicated an excellent fit. The GFI and AGFI values, on the other hand, suggested the presence of a mediocre fit (Brown, 2006; Tabachnick, & Fidell, 2007). In this study, all fit indices showed an excellent fit, except for the RMSEA value.

When the results were examined overall, it was seen that, although the RMSEA value was relatively high, the ratio of chi-square to the degrees of freedom and the NNFI, CFI, and standardized RMR values were at the expected level. Therefore, the one-factor structure that was determined as a result of EFA was validated by CFA and the construct validity of the scale was at the expected level.

Thus, EFA and CFA results are statistical evidence that the criteria in the rubric (tempo, rhythm, melody, harmony, technique) come together to form piano playing skills. As a result, the sum of the scores obtained from the criteria shows the piano playing skill level of the performer.

3.2. Findings on item discrimination

The discrimination levels of the scale items were calculated to investigate whether there was a significant difference between the scale scores and the factor scores of the upper and lower 27% of the participants. The upper 27% students refer to high performance group, and the lower 27% of the students refer to low performance group. Item discrimination was measured for the PPT-PT using the Mann-Whitney U test. The results of this test indicate that the upper group had higher mean rank scores for each item on the scale; this difference was significant at the .01 level. Moreover, item-total correlations were computed; it was found that all correlations suggested that the PPT-PT was capable of discriminating high and low performance groups.

3.3. Findings on internal consistency and generalizability studies

Cronbach’s Alpha was utilized to determine the coefficient of the internal consistency of the PPT-PT and calculated as .83. The results indicated that the internal consistency of the scale was high.

Table 6. Variance component values for a fully crossed design

Source	SS	df	MS	Corrected	%
S	142.43288	72	1.97823	0.15049	29.5
I	49.46575	4	12.36644	0.06753	13.2
R	2.08356	1	2.08356	-0.00111	0.0
SI	91.33425	288	0.31713	0.08658	17.0
SR	21.61644	72	0.30023	0.03125	6.1
IR	9.33425	4	2.33356	0.02999	5.9
SIR	41.46575	288	0.14398	0.14398	28.2
Total	357.73288	729			100%

As shown in Table 6, the variance component for the student main effect (S object of measurement) was 0.15, which explained approximately 30% of the total variance. This showed that the PPT-PT was capable of defining individual differences among the students. Considering the main effect of the item, it was seen that the variance component was 0.07, explaining 13% of the total variance. This means that the difficulties of the items in the PPT-PT were diversified. Most importantly, the rater main effect was examined and revealed that the variance component was 0.0011, which explained almost none of the total variance. In other words, the amount of variance from the rater was very low in the total variance. This means that the raters were consistent with each other.

Concerning the interaction effects of the variance components, student-item (SI), student-rater (SR), and item-rater (IR) were 0.08 (17%), 0.03 (6.1%), and 0.02 (5.9%), respectively. The variance component of the interaction effect of student-item-rater (SIR) was 0.14, which was the second highest component. The results show that 28% of the total variance consisted of random errors. Moreover, the G coefficient, which referred to the relative error variance was .76. This value was sufficient, considering PPT-PT had a relatively small number of items (Diederich, 1973). Furthermore, decision studies showed that, if the number of raters increased to 3, 4, and 5, the G coefficients would become .80, .82 and .83, respectively. Therefore, it would be better to use at least three raters for the musical instrument performance scale to obtain higher G coefficients.

3.4. Findings on the inter-class correlation coefficient

This section shows an inter-class correlation coefficient (ICC), which indicates rater reliability for Raters 1 and 2. The ICC between the raters was found to be .593 ($F_{364\text{-}365} = 3.81$, $p < .01$), suggesting moderate inter-rater reliability. Based on the 95% confidence interval of the ICC estimate, the values less than .50, between .50 and .75, between .75 and .90, and greater than .90 are indicative of poor, moderate, good, and excellent reliability, respectively.

3.5. Findings on the standard-setting study

Table 7. Results of the standard setting study

Expert No:		C1	C2	C3	C4	C5	Mean of row	SD of the row
1	Round 1	5	5	5	5	5	5.00	0.00
	Round 2	5	4	4	3	5	4.20	0.84
2	Round 1	4	5	4	4	4	4.20	0.45
	Round 2	4	5	4	4	4	4.20	0.45
3	Round 1	4	4	4	4	4	4.00	0.00
	Round 2	4	4	4	3	4	3.80	0.45
4	Round 1	5	4	4	3	3	3.80	0.84
	Round 2	4	4	4	3	3	3.60	0.55
5	Round 1	4	4	5	3	3	3.80	0.84
	Round 2	4	4	5	3	3	3.80	0.84
6	Round 1	4	5	5	5	3	4.40	0.89
	Round 2	4	5	5	5	3	4.40	0.89
7	Round 1	5	5	5	4	3	4.40	0.89
	Round 2	4	5	5	3	3	4.00	1.00
8	Round 1	4	4	5	4	3	4.00	0.71
	Round 2	4	4	5	4	3	4.00	0.71
9	Round 1	4	4	4	3	2	3.40	0.89
	Round 2	4	3	4	3	2	3.20	0.84
	Mean of column 1	4.33	4.44	4.55	3.89	3.33	4.11	
	Mean of column 2	4.33	4.22	4.44	3.44	3.33	3.95	
	SD of column 1	0.5	0.53	0.53	0.78	0.87	0.64	
	SD of column 2	0.33	0.66	0.53	0.73	0.87	0.62	

C: Criteria
SD: Standard Deviation

As shown in Table 7, when the cut-off scores for each item at the end of the second round were examined, it was seen that the experts allocated the lowest cut-off point to Item 5 (3.33) and the highest cut-off point to Item 3 (4.44). Also, the experts stated that the students at the borderline should have an average of 3.95 points from each item. The standard deviation values were examined to determine the variability between the experts' scores; it was therefore concluded that the variability between the expert opinions was less in Round 2 (0.62) than in Round 1 (0.64).

To compute the cut-off score, the mean of the total points of each expert for each item was taken. Then, the means were summed to obtain the cut-off score: 4.33+4.22+4.44+3.44+3.33 = 19.75

According to the findings obtained from the experts (using the extended Angoff method), for a student to be considered sufficient in their performance, they must obtain at least 19.75 out of 25 from the PPT-PT. Since it is not possible to obtain decimal scores from this scale, the cut-off score was rounded to 20.

4. Discussion and conclusions

This study aimed to investigate the psychometric properties and define the cut-off score of PPT-PT using different methods. The results of the EFA and CFA analyses showed that the PPT-PT had a one-factor structure, which means that the scale measured one latent trait, i.e., musical instrument performance (indicating sufficient construct validity). Therefore, we can state that the Scale purely measured piano performance without involving other traits. Thus, with this scale, students' musical abilities can be evaluated fairly.

Cronbach's Alpha coefficient result indicates that the internal consistency of the scale was high. In other words, the items/criteria were inter-correlated, and it was appropriate to sum the scores from different items/criteria. This study employs an analytical rubric to assess students' musical performances, which involves evaluating specific sub-behaviors, including tempo, melodic accuracy, rhythmic accuracy, harmonic precision, and fingering. When both EFA CFA results and Cronbach's Alpha value are evaluated together, it can be interpreted that these theoretically predicted sub-behaviors come together in a consistent manner and constitute musical performance. In other words, this assumption in the literature has been confirmed in a sense.

The results of the item discrimination analysis revealed that the PPT-PT was good at discriminating students who had high and low levels of ability to play a musical instrument. This demonstrates a strong correlation between the expressions delineated for the sub-components of musical performance and their direct relevance to the art of musical performance itself. Consequently, employing this rubric can facilitate the assessment of students' piano performances using repertoire akin to the examples provided herein.

The findings of the generalizability studies showed that the test was capable of defining individual differences among the students. Furthermore, the amount of variance from the raters was very low in total, meaning that the raters were consistent. In other words, there was an agreement among the raters when they independently scored the students' performances. However, the random error variance, which showed the amount of error that interfered with the measurement process, was relatively high. This means that some other variables (e.g., time, type of rubric, etc.) should be considered when undertaking generalizability analyses in further studies. Furthermore, the G coefficient indicated an acceptable reliability level for the investigated test, which is also consistent with the Cronbach's Alpha coefficient. However, the decision studies utilized indicated that, in order to increase reliability, it was better to use three raters. In this study, we had two raters. It is therefore recommended for future instructors or researchers applying the PPT-PT to utilize at least three raters to obtain more reliable results.

The findings related to the ICC indicated moderate inter-rater reliability. The level of inter-rater reliability was relatively acceptable. On the other hand, similar to the results of the decision study, it was considered to be better to use three raters than two. When assessing both generalizability and intraclass correlation coefficient analyses together, it becomes evident that the explanations provided for the sub-behaviors in piano performance were consistently understood by various raters. Consequently, this suggests a high level of accuracy in introducing and defining the criteria (sub-behaviors) during the development of the scoring key.

The results of the standard-setting study suggested that the cut-off score of the PPT-PT was relatively high. This shows the importance of the piano performance being assessed as a whole. For example, if a student plays the rhythm, notes, chords, tempo, etc., partly accurately, they can achieve a score of 13 of 25 (52%) from this scale. Thus, a student who receives a score of 13 out of 25 (52%) can be considered successful. However, in this case, the performed piece would nonetheless be incomprehensible to the audience. Moreover, in this study, the participants were teacher candidates, who are expected to teach these pieces to their students. Therefore, a score of at least 20 of 25 points (80%) on the PPT-PT was required in order to pass the examination.

This study shows that the PPT-PT is valid and reliable and can be used to assess undergraduate students' piano performances with a cut-off score of 20 of 25 (80%). Therefore, it is advised that instructors use this Scale to assess students' performance. However, it is highly recommended that a minimum of 20 be used as a cut-off score. Moreover, instructors will obtain more reliable results if they use at least three raters. Nevertheless, it should not be forgotten that the raters must be experienced and well-qualified. Otherwise, the reliability of the scale may decrease (Kamış, & Doğan, 2018).

Future researchers can repeat similar studies with more samples and raters. Additionally, they can use other techniques for standard-setting analyses, such as Bookmark, Nedelsky, etc. (Cizek, & Bunch, 2007). Moreover, if the researchers have a larger sample (more than 250 participants), it is advised using the many-faceted Rasch model to test the validity and reliability of the PPT-PT or other developed scales.

References

Abeles, H.F. (1973). Development and validation of a clarinet performance adjudication rating scale. *Journal of Research in Music Education*, *21*, 246-255. https://doi.org/10.2307/3345094

Alessandri, E., Williamson, V.J., Eiholzer, H., & Williamon, A. (2016). A critical ear: analysis of value judgments in reviews of Beethoven’s piano sonata recordings. *Frontiers in Psychology*, *7*, 391. https://doi.org/10.3389/fpsyg.2016.00391

Andrade, H.G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, *57*(5),13-18. https://www.researchgate.net/publication/285750862_Using_rubrics_to_promote_thinking_and_learning

Barret, P., & Kline, P. (1981). The observation to variable ratio in factor analyses. *Journal of Personality and Group Behavior*, *2*, 94-98. https://www.researchgate.net/publication/232561774_The_Observation_to_Variable_Ratio_in_Factor_Analysis

Bejar, I.I. (2008). *Standard setting: What is it? Why is it important?* (R&D Connections No. 7.). https://www.ets.org/Media/Research/pdf/RD_Connections7.pdf

Ben-David, M.F. (2000). AMEE guide no.18: Standard setting in student assessment. *Medical Teacher*, 22(2), 120-130. https://doi.org/10.1080/01421590078526

Bergee, M.J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education*, *51*(2), 137-150. https://doi.org/10.2307/3345847

Bergee, M.J. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education*, *55*(4), 344-358. https://doi.org/10.1177/0022429408317515

Boud, D. (1995). *Enhancing Learning Through Self-Assessment*. Routledge.

Boyle D.J., & Radocy, R.E. (1987). *Measurement: Evaluation of musical experiences*. Schirmer Books.

Brennan, R.L. (2001). *Statistics for social science and public policy generalizability theory*. Springer.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. Guilford Publications, Inc.

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Cengage Learning.

Diederich, P.B. (1973). *Short-cut statistics for teacher-made tests*. Educational Testing Service.

Duerksen, G.L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education., 20*, 268-272. https://doi.org/10.2307/3344093

Elliott, D., & Silverman, M. (2015). *Music matters: A Philosophy of music education*. Oxford University Press.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2015). *How to design and evaluate research in education* (9th ed.). McGraw Hill Education.

Guven, E.D. (2020). Piano-accompanied solfège reading experiences of preservice music teachers. *Research Studies in Music Education*, *43*(3), 417-433. https://doi.org/10.1177/1321103x19871078

Gliner, J.A., Morgan, G.A., & Leech, N.L. (2009). *Research methods in applied settings: An integrated approach to design and analysis*. Routledge.

Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking*. Allyn and Bacon.

Hooper, D., Coughan, J., & Mullen, M. (2008). Structural equation modeling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, *6*(1), 53-60. http://mural.maynoothuniversity.ie/6596/

Howell, D.C. (2013). *Statistical methods for psychology*. Cengage Learning.

Juslin, P.N. (2003). Five facets of musical expression: a psychologist's perspective on music performance. *Psychology of Music*, *31*, 273-302. https://doi.org/10.1177/03057356030313003

Kamış, Ö., & Doğan, C.D. (2018). An investigation of reliability coefficients estimated for decision studies in generalizability theory. *Journal of Education and Learning*, *7*(4), 103-113. https://doi.org/10.5539/jel.v7n4p103

Kim, S., Park, J.M., Rhyu, S., Nam, J., & Lee, K. (2021). Quantitative analysis of piano performance proficiency focusing on difference between hands. *PLoS ONE*, *16*(5), 1-28. https://doi.org/10.1371/journal.pone.0250299

Kutlu, Ö., Doğan, C.D., & Karakaya (2014). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme [Measurement and evaluation: An assessment based on performance and portfolio]*. Pegem Academy.

Lee, Y. (2009). Music practices and teachers' needs for teaching music in public preschools of south korea. *International Journal of Music Education*, *27*(4), 356-371. https://doi.org/10.1177/0255761409344663

Leech, N.L., Barrett, K.C. y George, A.M. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed.). Lawrence Erlbaum Associates, Publishers.

Mak, H.W., & Fancourt, D. (2019). Arts engagement and self-esteem in children: Results from a propensity score matching analysis. *Annals of the New York Academy of Sciences*, *1449*(1), 36-45. https://doi.org/10.1111/nyas.14056

McMillan, J.H., & Hearn, J. (2008). Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons*, *87*(1), 40-49. https://www.jstor.org/stable/42923742

Moskal, M.B. (2000). Scoring rubric: what, when and how? *Practical Assessment, Research & Evaluation*, *7*, 1-5. https://doi.org/10.7275/a5vq-7q66

Nichols, J.P. (1991). A factor analysis approach to the development of a rating scale for snare drum performance. *Dialogue Instrumental Music Education*, *15*, 11-31. Michigan University Press.

Palmer, C. (1996). On the assignment of structure in music performance. *Music Perception*, *14*, 23-56. https://doi.org/10.2307/40285708

Price, H.E., & Burnsed, V. (1989). Classroom teachers' assessments of elementary education music methods. *Update: Applications of Research in Music-education*, *8*(1), 28-32. https://doi.org/10.1177/875512338900800107

Robinson, N.R. (2016). Developing a critical consciousness for diversity and equity among preservice music teachers. *Journal of Music Teacher Education*, *26*(3), 11-26. https://doi.org/10.1177/1057083716643349

Russell, B.E. (2010). *The Empirical Testing of a Musical Performance Assessment Paradigm. PhD*. University of Miami.

Schellenberg, E.G. (2018). Music and cognitive abilities. En S. Hallam, I. Cross y M. Thaut (Eds.), *The Oxford handbook of music psychology* (2nd ed.) (pp.261-275). Oxford University Press.

Scripp, L., & Kaufman, B. (2019). *Music Learning as Youth Development*. https://doi.org/10.4324/9780429436956-11

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory a primer*. Sage Publications.

Southgate, L., Hays, R.B., Norcini, J., Mulholland, H., Ayers, B., Woolliscroft, H., Cusimano, M., McAvoy, P., Ainsworth, M., Haist, S., & Campbell, M. (2001). Setting performance standards for medical practice: A theoretical framework. *Medical Education*, *35*(5), 474-481. https://doi.org/10.1046/j.1365-2923.2001.00897.x

Standley, J.M. (2016). Music and early childhood development. En A.C. Lehmann, J.A. Sloboda y R. Woody (Eds.), *Psychology for musicians: Understanding and acquiring the skills* (pp.103-118). Oxford University Press.

Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, *18*(1), 46-56. https://doi.org/10.1177/1321103x020180010601

Stevens, D.D., & Levi, A. (2005). *Introduction to rubrics: An assessment tool to save time, convey effective feedback, and promote student learning*. Styus.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar [Structural equation modeling: Basic concepts and applications]. *Türk Psikoloji Yazıları*, *3*(6), 49-74. https://www.researchgate.net/publication/281981476_Yapidotlesssal_esitlik_modelleri_Temel_kavramlar_ve_ornek_uygulamalar

Swanwick, K. (2016). Music education and human development. En R. Colwell y C. Richardson (Eds.), *The new handbook of research on music teaching and learning* (pp.3-23). Oxford University Press.

Tabachnick, B.G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson Education.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, *21*(1), 21-41. https://doi.org/10.1525/mp.2003.21.1.21

Thompson, W.F., Diamond, C.T.P., & Balkwill, L.L. (1998). The adjudication of six performances of a Chopin etude: A study of expert knowledge. *Psychology of Music*, *26*(2), 154-174. https://doi.org/10.1177/0305735698262004

Urbina, S. (2004). *Essentials of psychological testing*. John Wiley & Sons.

Wapnick, J., Ryan, C., Lacaille, N., & Darrow, A.-A. (2004). Effects of selected variables on musicians' ratings of high-level piano performances. *International Journal of Music Education*, *22*(1), 7-20. https://doi.org/10.1177/0255761404042371

Wapnick, J., Ryan, C., Campbell, L., Deek, R., Lemire, R., & Darrow, A.A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performance. *Journal of Research in Music Education*, *53*, 162-176. https://doi.org/10.1177/002242940505300206

Watkins, J.G., & Farnum, S.E. (1954). *The Watkins-Farnum performance scale, form A; A standardized achievement test for all band instruments*. Hall Leonard Music Press.

Wesolowski, B.C. (2012). Understanding and developing rubrics for music performance assessment. *Music Educators Journal*, *98*(3), 36-42. https://doi.org/10.1177/0027432111432524

Wesolowski, B.C., Wind, S.A., & Engelhard, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the Multifaceted Rasch Partial Credit Model. *Music Perception: An Interdisciplinary Journal*, *33*(5), 662-678. https://doi.org/10.1525/mp.2016.33.5.662

Wesolowski, B.C., Wind, S.A., & Engelhard, G. (2015). Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae*, *19*(2), 147-170. https://doi.org/10.1177/1029864915589014

Wrigley, W.J., & Emmerson, S.B. (2011). Ecological development and validation of a music performance rating scale for five instrument families. *Psychology of Music*, *41*(1), 97-118. https://doi.org/10.1177/0305735611418552

Zdzinski, S.F., & Barnes, G.V. (2002). Development and validation of a string performance rating scale. *Journal of Research in Music Education*, *50*, 245-255. https://doi.org/10.2307/3345801

APPENDIX A - Piano Performance Test for Pre-Service Teachers

Analytical rubric

Student's name: **Department:** **..../..../ 20**

	5 Points	4 Points	3 Points	2 Points	1 Point
Tempo	**Tempo is considered and even** *Played at the right tempo from beginning to end.*	**Tempo is even** *Played at the right pace, with a few hesitations.*	**Tempo is mostly even** *Paused while playing or hesitated a lot*	**Tempo is mostly erratic and uneven** *Paused many times while playing or Tempo was unstable due to hesitation.*	**Couldn't complete the piece.**
Rhythmic Accuracy	**The rhythm is coherent and makes musical sense** *Played the rhythms completely right*	**The rhythm is mostly coherent and makes musical sense** *Made minor rhythmic mistakes in one or two measures (less than 10%) during the performance*	**The rhythm is often coherent and makes little musical sense** *Made three to six minor mistakes (11% to 30%) or stopped once during the performance*	**The rhythm is incoherent and does not make musical sense.** *Made seven to ten minor rhythmic mistakes* (31% to 50%) *or stopped several times (more than twice) during the performance.*	**The rhythm was almost imperceptible**
Melodic Accuracy	**The melody feels complete and coherent; the piece makes musical sense** *Played the melody completely right*	**The melody feels less complete and coherent; the piece makes less musical sense** *Played one or two incorrect notes (less than 10%) during the performance.*	**The melody feels musically incomplete** *Played three-five incorrect notes (11% to 30%) during the performance.*	**The melody is not complete or coherent** Played six to ten *incorrect notes* (31% to 50%) *during the performance.*	**The melody was almost imperceptible**
Harmonic Accuracy	**Chords are evident and musically played** *Played all the chords completely right*	**Chords are used in a manner that supports the piece musically** *Played one or two incorrect chords (less than 10%) or incorrect voices in chords.*	**The chords do not always align with the melody** *Played three to five incorrect chord (11% to 30%) or incorrect voices in chords.*	**Chords are used, but do not align appropriately with melody** *Played six to ten incorrect chords (31% to 50) or incorrect voices in chords.*	**The chords were almost imperceptible**
Fingering	**All fingering was played correctly** *Played the whole piece with the right fingering.*	**Almost all fingering was played correctly** *Played with one or two incorrect fingering (less than 10%) during the performance*	**Some of the fingering was correct, but the overall impression was inconsistent** *Played with three to five incorrect fingering (11% to 30%) during the performance*	**Fingering was not followed** Played with six to ten incorrect fingering *(31% to 50%)* during the performance	**Demonstrated a lack of understanding of appropriate fingering**

If necessary, the rater can fill out this part.

Extraordinary performance that cannot be assessed via the PPT-PT:

Weaknessess of the student that cannot be assessed via the PPT-PT:

APPENDIX B- Extended Angoff Method Expert Opinion Form

Expert ID : ____________

ROUND 1

Rate the borderline examinee performance from 1 to 5 points for each criteria given below

CRITERIA	Tempo (Criteria 1)	Ryhtmic Accuracy (Criteria 2)	Melodic Accuracy (Criteria 3)	Harmonic Accuracy (Criteria 4)	Fingering (Criteria 5)
SCORES					

Reveal your estimates of the borderline examinee performance and discuss any inter-rater disagreement.

ROUND 2

Rate the borderline examinee performance from 1 to 5 points regarding the previous discussion for each criteria given below

CRITERIA	Tempo (Criteria 1)	Ryhtmic Accuracy (Criteria 2)	Melodic Accuracy (Criteria 3)	Harmonic Accuracy (Criteria 4)	Fingering (Criteria 5)
SCORES					

1st and 2nd Round Descriptive Statistics (This part will be filled out by the researcher)

	Minimum Score		Maximum Score		Mod		Median		Mean	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Criteria 1										
Criteria 2										
Criteria 3										
Criteria 4										
Criteria 5										

R1: Round 1 R2: Round 2