Revista Electrónica de LEEME, (2023) (52) (pág.18-pág.37)

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

ARTÍCULOS

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

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

Salim Sever[1]
Departamento de Educación Primaria, Ankara University, Ankara (Turquía)
C. Deha Dogan[2]
Departamento de Evaluación y Medición Educativa, Ankara University, Ankara (Turquía)
Omer Kamis[3]
Departamento de Evaluación y Medición Educativa, Çankırı Karatekin University, Çankırı (Turquía)
Gulsah Sever[4]
Departamento de Educación Artística
División de Educación Musical, Gazi University, Ankara (Turquía)

doi:10.7203/LEEME/52.27217
Recepción: 01-08-2023 Revisión: 08-08-2023 Aceptado: 25-09-2023

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.

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.

[1] Profesor de Musicología, Facultad de Ciencias de la Educación, https://orcid.org/0000-0003-4028-4514
* Contact and correspondence: Salim Sever, Departamento de Educación Primaria, Ankara University, salimsvr@gmail.com, Cemal Gürsel Cad. Ankara Üniversitesi Cebeci Yerleşkesi Çankaya, 06590 Ankara. Turquía.
[2] Profesor Adjunto de Evaluación y Medición Educativa, Facultad de Ciencias de la Educación, https://orcid.org/0000-0003-0683-1334
[3] Candidata al doctorado de Evaluación y Medición Educativa, Facultad de Educación, https://orcid.org/0000-0003-0605-087X
[4] Profesor Adjunto de Educación Musical, Facultad de la Educación, https://orcid.org/0000-0003-0559-6993

1. Introducción

La educación musical es un aspecto importante del desarrollo infantil y la inclusión de clases de música en los planes de estudio del profesorado de Primaria y Infantil puede tener un impacto significativo en los resultados de aprendizaje de los niños y las niñas (Swanwick, 2016). La investigación demostró que la educación musical puede mejorar el desarrollo del lenguaje, las habilidades cognitivas y el bienestar socioemocional de niños y niñas (Schellenberg, 2018). Además, la incorporación de la música en el currículo preescolar puede ayudar a promover la creatividad y fomentar el amor por la música en el alumnado de esa etapa (Standley, 2016). Por lo tanto, es esencial que el profesorado de Primaria y Infantil reciba formación en educación musical e incorporen clases de música en sus rutinas diarias para apoyar el desarrollo holístico de sus alumnos. Por lo tanto, según Scripp y Kaufman (2019), es esencial que el profesorado de Primaria y Infantil tenga una base sólida en música, incluidas las habilidades con el piano. El piano es uno de los instrumentos fundamentales que proporciona una base sólida para que la educación musical proporcione una instrucción musical efectiva para niños pequeños y niñas pequeñas. La formación en piano puede mejorar la alfabetización musical del profesorado, proporcionar oportunidades para la expresión creativa y mejorar su capacidad para integrar la música en el plan de estudios del aula (Robinson, 2016; Guven, 2020). Scripp y Kaufman (2019) afirmaron que la enseñanza del piano puede ayudar a que el profesorado candidato a la etapa de Infantil a adquirir competencia en teoría musical, entrenamiento auditivo y habilidades de interpretación, que pueden aplicarse a la enseñanza del alumnado de dicha etapa. Según la investigación de Price y Burnsed (1989), las habilidades más necesarias en la Educación Infantil y Primaria son cantar y tocar. El piano (teclado), es capaz de crear diferentes sonidos, permite el acompañamiento y permite al profesorado utilizar su voz como canto o dirección mientras tocan simultáneamente. Al aprender a tocar el piano, el profesorado de Infantil y Primaria adquiere una base sólida en teoría musical, armonía y técnica, lo que le permite enseñar y guiar eficazmente a los niños y las niñas en su desarrollo musical (Lee, 2009). En este sentido, es un instrumento musical eficaz para utilizar en los cursos de música de Primaria y Infantil.

La evaluación es parte integrante del proceso de enseñanza. Sin una evaluación válida y fiable, no podemos estar seguros de la calidad y eficacia del proceso de enseñanza. La evaluación de la interpretación pianística desempeña un papel fundamental en la valoración de las capacidades técnicas, las habilidades interpretativas y la musicalidad general de los intérpretes de piano. Según Boud (1995), la evaluación de la interpretación permite a quienes interpretan recibir información sobre sus puntos fuertes y débiles, así como identificar áreas de mejora. Las evaluaciones de rendimiento también proporcionan información valiosa para que el profesorado de Música ajuste sus estrategias de instrucción para satisfacer las necesidades del alumnado (Elliott y Silverman, 2014).

El profesorado de piano tiene diferentes hábitos de puntuación a la hora de evaluar la interpretación pianística de sus estudiantes. Hay docentes que pueden utilizar un sistema de puntuación más indulgente para animar a sus discentes, mientras que otros pueden utilizar un sistema de puntuación más estricto para motivar a sus alumnos a mejorar. La investigación ha demostrado que los métodos de evaluación estructurados, como los exámenes y las competiciones, pueden tener un impacto positivo en la motivación y la autoestima del alumnado (Mak y Fancourt, 2019). Estas evaluaciones también ofrecen una manera de evaluar y comparar los niveles de rendimiento entre diferentes estudiantes e instituciones, proporcionando un estándar común para la excelencia en la interpretación del piano (Kim *et al.*, 2021).

Se han desarrollado diferentes herramientas y modelos de medición para evaluar el rendimiento musical (Abeles, 1973; Boyle y Radocy, 1987; Watkins y Farnum, 1954; Nichols, 1991; Palmer, 1996; Stanley *et al.*, 2002; Zdzinski y Barnes, 2002; Juslin, 2003; Wapnick *et al.*, 2004; Wapnick *et al.*, 2005; Russell, 2010; Alessandri *et al.*, 2016). Una de las principales dificultades de la evaluación de la interpretación musical es la gestión de su naturaleza subjetiva. Las rúbricas pueden utilizarse como un conjunto de directrices para evaluar objetivamente el rendimiento del alumnado (Wesolowski, 2012). Una rúbrica es un conjunto coherente de criterios que incluye descripciones de los niveles de calidad de la interpretación en una serie de dimensiones. Una rúbrica analítica especifica las expectativas para una tarea determinada dividiéndola en sus partes y proporcionando una descripción detallada (Stevens y Levi, 2005). Las rúbricas para la interpretación pianística podrían incluir criterios como la precisión rítmica, la precisión del tono, el tempo, el acento, la dinámica, la calidad del tono, la interpretación y la calidad general (Duerksen, 1972), técnicos (tono, entonación, precisión rítmica, articulación y técnica) y musicales (tempo, dinámica, timbre, interpretación y expresión musical) (Russell, 2010), expresión de la mano derecha, fraseo, dinámica, rubato, forma/estructura, equilibrio tonal, pedaleo, atención al ritmo y al compás, articulación, competencia técnica, tempo, expresión de varias partes (Thompson *et al.*, 1998). Aunque las rúbricas creadas para la interpretación contienen diferentes ítems de evaluación, pueden agruparse bajo tres epígrafes principales: competencia técnica, interpretación musical y presencia escénica. Mediante el uso de una rúbrica, los profesores pueden proporcionar un feedback más detallado a los alumnos y ayudarles a identificar áreas de mejora (McMillan y Hearn, 2008).

A la hora de evaluar la interpretación musical, la coherencia es crucial, y el uso de rúbricas analíticas aumenta la fiabilidad entre evaluadores. Para comprobar la fiabilidad de una herramienta de evaluación, pueden aplicarse varias técnicas estadísticas basadas en diversas teorías de test. Uno de los principales modelos de evaluación es la teoría clásica de los tests (CTT). En la CTT, la fiabilidad se define en términos de consistencia interna y estabilidad (Urbina, 2004). El coeficiente Alfa de Cronbach suele utilizarse como medida de la consistencia interna (Crocker y Algina, 1986). Las técnicas Kappa de Cohen y de correlación entre clases también proporcionan información sobre la fiabilidad entre evaluadores (Gliner *et al.*, 2009; Howell, 2013).

El otro constructo de evaluación es la teoría de la generalizabilidad (TG). Si se utiliza más de un evaluador para evaluar el rendimiento musical de un alumno, la varianza de la faceta (variable) del evaluador proporciona información sobre la fiabilidad del evaluador. Los coeficientes *G* y *Phi* también proporcionan información sobre la fiabilidad del proceso de medición (Shavelson y Webb, 1991). Por lo tanto, evaluar la interpretación pianística utilizando tanto la teoría clásica de los tests como la teoría de la generalizabilidad podría proporcionar pruebas sólidas de la fiabilidad entre evaluadores.

La medida de calificación estándar se ocupa de establecer una puntuación de corte que separe al alumno competente del incompetente en un examen o en una evaluación basada en el rendimiento (Ben-David, 2000). Una puntuación de corte es simplemente la puntuación que sirve para delimitar un nivel de otro. Si las puntuaciones de corte no se fijan adecuadamente, los resultados de la evaluación podrían quedar en entredicho. Por ese motivo, el establecimiento de estándares es un componente crítico del proceso de desarrollo de pruebas (Bejar, 2008). Es fundamental establecer puntuaciones de corte para las evaluaciones de interpretación musical.

Existen varios enfoques para establecer estándares tanto para la evaluación escrita como para la basada en el rendimiento (Southgate *et al.*, 2001). Se ha desarrollado un gran número de

métodos que se utilizan para establecer estándares tanto para los exámenes escritos como para los basados en el rendimiento (Cizek y Bunch, 2007). Tocar el piano también es una actuación musical que se puede medir; por lo tanto, se trata como una asignatura que establece estándares.

Entre estos marcos metodológicos, Angoff es un método comúnmente utilizado para el establecimiento de normas que requiere calificadores expertos. Este método se basa en el concepto de estudiante límite o mínimamente competente. En otras palabras, se basa en la determinación del alumno o de la alumna que realiza la interpretación mínima aceptable según la puntuación realizada por dos o más expertos. El método de Angoff ampliado es un enfoque paralelo para ítems con puntuación politómica (Cizek y Bunch, 2007). Angoff ampliado es uno de los métodos más apropiados para definir una puntuación de corte para la escala de interpretación pianística para docentes de Primaria y Infantil.

En la literatura, hay algunos estudios centrados en la fiabilidad entre evaluadores de pruebas de interpretación musical (Boyle y Radocy, 1987; Thompson *et al.*, 1998; Thompson y Williamon, 2003). Asimismo, Bergee (2003, 2007) utilizó la Teoría G para comprobar la fiabilidad entre evaluadores de la interpretación musical. Wrigley y Emmerson (2011) utilizaron el modelo de ecuaciones estructurales en el estudio de fiabilidad de la rúbrica de rendimiento que desarrollaron para todos los instrumentos musicales.

Al evaluar la actuación del alumnado, se tuvieron en cuenta los puntos del jurado. La interpretación musical tiene múltiples capas y es compleja. Por esa razón, la evaluación se ve a menudo forzada hacia una naturaleza personal y emocional. Se ha reconocido que los examinadores que valoran la misma interpretación pueden tener diferentes constructos, pueden sopesar los mismos constructos de forma diferente o pueden no tener ningún constructo consciente, sino basarse en la intuición o en una corazonada y, como resultado, tener dificultades para articular cómo llegan a una valoración de la interpretación (Boyle y Radocy, 1987; Thompson *et al.*, 1998). Ser un experto para evaluar también es una cuestión importante, ya que se ha observado una escasa coherencia entre evaluadores cuando estos, que son novatos en la calificación de la evaluación, definen un constructo determinado de forma diferente a los que están más familiarizados con la herramienta (Thompson y Williamon, 2003). Por otra parte, hay algunas investigaciones que investigan la severidad de los calificadores y la presición de los calificadores utilizando el Modelo de Medición de Crédito Parcial Rasch de Muchas Facetas (Wesolowski, 2012, Wesolowski *et al.*, 2015, 2016).

La nota de aprobado puede variar de una institución a otra, pero en la mayoría de los casos se sitúa entre 60 y 70 puntos sobre 100. ¿Puede el profesorado, tocando en el nivel de 60 puntos, transmitir la pieza musical al alumnado? ¿Qué nivel de puntuación es suficiente para transmitir la música al alumnado? En otras palabras, ¿cuál debería ser la puntuación equivalente al nivel mínimo de audición de una interpretación pianística? Para examinar esto, se recogieron datos a través de una rúbrica que mide las dimensiones de ritmo, melodía, armonía, técnica y tempo. El objetivo principal de este estudio es determinar la nota de aprobado en la evaluación realizada con una rúbrica con 5 criterios, cada uno de los cuales se puntúa entre 1 y 5.

En la bibliografía pertinente, no se han encontrado estudios centrados en el desarrollo de una escala de interpretación pianística para docentes de Educación Primaria e Infantil, a pesar de que se trataría de una medida de evaluación bastante útil para su uso en cursos de música elemental y preescolar. La creación de un instrumento de este tipo y la comprobación de sus propiedades psicométricas (validez y fiabilidad) contribuirán así a la literatura. Esta escala puede

utilizarse para evaluar el rendimiento pianístico del profesorado en formación de las facultades de pedagogía.

Además, se ha visto que muchos factores influyen en las puntuaciones de las rúbricas, lo que tiene que ver en gran medida con cuestiones de validez y fiabilidad. El uso y la comparación de diversas herramientas estadísticas se consideran importantes para medir la fiabilidad entre evaluadores. El uso conjunto de la teoría clásica de los tests y de la teoría de la generalizabilidad ayudará a los investigadores a evaluar los problemas de validez y fiabilidad con mayor eficacia. Sin embargo, en la bibliografía hay muy pocos estudios que utilicen una combinación de estos enfoques para comprobar la fiabilidad entre evaluadores. Por ello, en esta investigación se utilizaron conjuntamente diferentes métodos basados en la CTT y la Teoría G. Por otra parte, en la bibliografía no hay estudios que tengan como objetivo definir puntuaciones de corte para la interpretación pianística de futuros docentes de Infantil y Primaria.

El presente estudio añade una perspectiva estadística a la investigación previa en este campo mediante el examen de las características psicométricas y también incluye el estudio de establecimiento de normas para el *Piano Performance Test for Pre-Service Teachers* (PPT-PT). De este modo, presenta una forma válida y fiable de evaluar la interpretación pianística mediante rúbricas.

El presente estudio tiene como objetivo investigar las propiedades psicométricas y definir la puntuación de corte del PPT-PT basándose en diferentes teorías de medición. Las preguntas de investigación pertinentes son:

1. ¿Cuál es la validez de constructo del PPT-PT? (¿Mide el PPT-PT efectivamente su constructo previsto?).
2. ¿Cuál es el grado de error de las puntuaciones del PPT-PT en el contexto del CTT y el GT?
3. ¿Cuál es el nivel de acuerdo entre los evaluadores y las evaluadoras independientes que puntúan el PPT-PT?
4. ¿Cuál es la puntuación de corte del PPT-PT según el método de Angoff ampliado? (¿Cuál es la puntuación mínima que debe obtener el alumnado para superar el PPT-PT? según el método de Angoff ampliado)?

2. Método

2.1. Diseño

Esta investigación descriptiva pretende investigar las propiedades psicométricas de un instrumento evaluativo, el PPT-PT. El presente estudio también pretende definir una puntuación de corte para el PPT-PT. En la investigación educativa, quienes investigan resumen las características de individuos, grupos, instrumentos de medición, etc. sin intervención en el proceso (Fraenkel *et al.*, 2015); así, en este estudio, las características psicométricas y los estándares del PPT-PT se describieron sin intervención alguna.

2.2. Participantes

Las personas participantes en el estudio quedaron divididas en tres grupos: estudiantes, calificadores utilizados para probar las características psicométricas del PPT-PT y expertos para el estudio de normalización del test.

Tabla 1. Características demográficas de las personas participantes

Participantes	N	Grado
Programa de Educación Primaria	50	Segundo
Programa de Educación Infantil	50	Tercera
Total	100	

El grupo de estudio 1 estaba formado por 100 estudiantes universitarios matriculados en los Departamentos de Educación Infantil y Primaria de una universidad pública de Ankara. El alumnado participante fue seleccionado mediante un muestreo de conveniencia. Por razones prácticas, no fue posible utilizar un muestreo aleatorio.

El grupo de estudio 2 estaba formado por dos evaluadores independientes que impartían cursos de educación musical en distintas universidades públicas de Ankara. Uno de los evaluadores es Catedrático de Pedagogía Musical y lleva 19 años impartiendo diversos cursos de grado y posgrado, como teoría musical, pedagogía musical y sociología de la música. El otro evaluador es Profesor Asociado de Pedagogía Musical y lleva 12 años impartiendo cursos de violín individuales y en grupo, tanto de grado como de postgrado. Ambos calificadores también tienen experiencia en calificar interpretaciones de instrumentos musicales utilizando diversas rúbricas de evaluación.

El Grupo 3 estaba formado por seis expertos del campo de la educación musical y la interpretación, a los que se pidió que estimaran el rendimiento de los examinandos límite para cada uno de los criterios del PPT-PT (Cizek y Bunch, 2007).

2.3. Instrumentos

Para este estudio, los investigadores desarrollaron una prueba de interpretación al piano y una rúbrica analítica. La prueba de interpretación requería que el alumnado tocara melodías al piano. La pieza para piano incluida en esta investigación procedía de las melodías populares que se enseñan en los cursos de Música de las Escuelas Primarias.

La pieza para piano tenía un compás de 4/4 y 32 compases. Estaba compuesta en la escala de re menor utilizando seis notas consecutivas entre Do y La. La parte de la mano izquierda incluía acordes de Tónica y Dominante en las posiciones principal y primera de inversión. Para evaluar la interpretación del alumnado, los calificadores utilizaron una rúbrica basada en cinco criterios (tempo, ritmo, melodía, armonía y técnica), puntuando cada criterio con un total de 5 puntos. La rúbrica analítica de la interpretación pianística (Anexo A) se desarrolló teniendo en cuenta los pasos que se muestran en la Tabla 2 (Andrade 2000; Haladyna, 1997; Kutlu *et al.*, 2014; Moskal, 2000).

Tabla 2. Proceso de Elaboración de Rúbricas

Pasos	Procedimiento
1	Definir las características/subdimensiones que deben evaluarse
2	Definir el intervalo de puntuaciones para cada subdimensión
3	Definición de los indicadores de resultados
4	Elaborar el borrador de una rúbrica analítica
5	Recabar la opinión de expertos

6	Elaboración del borrador final de la rúbrica analítica

El formulario de expertos para la fijación de normas, desarrollado por los investigadores, consta de dos partes y se administra, correspondientemente, en dos rondas. En la primera ronda, los expertos valoran el rendimiento del examinando límite de forma independiente para cada uno de los cinco criterios. A continuación, los expertos revelan sus estimaciones del rendimiento del examinando dudoso y discuten cualquier desacuerdo entre los evaluadores. En la segunda ronda, los expertos valoran el rendimiento del examinando dudoso en relación con la discusión anterior y rellenan la segunda parte del formulario (Anexo B).

2.4. Recogida de datos

Para la recogida de datos, el alumnado interpretó la pieza para piano y cada evaluador puntuó estas interpretaciones de forma independiente. Puesto que todo el alumnado interpretó la misma pieza musical y cada evaluador puntuó la interpretación de cada estudiante, se consiguió un diseño totalmente cruzado (PxIxR). La TG también permite diseños anidados en los que diferentes grupos de calificadores puntúan a diferentes grupos de estudiantes (Brennan, 2001). Pero, en este estudio, se utilizó un diseño totalmente cruzado, en el que cada estudiante del estudio fue evaluado por el mismo conjunto de calificadores. El ejemplo de estructura de datos del diseño totalmente cruzado se presenta en la Tabla 3.

Tabla 3. Ejemplo de estructura de datos para un diseño completamente cruzado

	Ítem 1		**Ítem 2**		**. . .**	**Ítem 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: Evaluador, P: Persona.

Cada evaluador puntuó el rendimiento del alumnado de forma individual e independiente. Este proceso duró aproximadamente diez minutos para cada estudiante y se completó en diez días.

La recogida de datos del estudio de fijación de normas se llevó a cabo en las cinco etapas siguientes. En la etapa 1, se explicó a los expertos en qué consistía la fijación de normas y se les explicó el método Angoff ampliado utilizado en el estudio. En la Etapa 2, los expertos debatieron sobre las competencias del alumnado límite en relación con su rendimiento musical. En la fase 3, los expertos definieron individualmente el rendimiento de los alumnos límite en cada uno de los criterios del PPT-PT utilizando el formulario de expertos "fijación de normas". En la fase 4, los expertos revelaron sus estimaciones del rendimiento de los alumnos límite y debatieron cualquier desacuerdo entre los evaluadores. En la etapa final, los expertos redefinieron el rendimiento límite de los examinados para cada uno de los criterios (teniendo en cuenta lo que habían debatido en la sección anterior). La recogida de datos para la sesión de estudio de fijación de normas se completó

en aproximadamente 2 horas. La aprobación se obtuvo del Comité de Ética de la Universidad de Ankara (ID: 85434274-050.04.03/2022).

2.5. Análisis de Datos

En cuanto al análisis de validez, para definir la validez de constructo de la Escala, se calcularon análisis factoriales exploratorios y confirmatorios. Para definir la discriminación de los ítems, se realizó la prueba U de Mann-Whitney para comparar el 27% superior e inferior de las personas participantes.

Respecto al análisis de fiabilidad, para probar la fiabilidad interpares, se calculó el coeficiente Alfa de Cronbach basado en el CTT. Para comprobar la fiabilidad entre evaluadores, se utilizaron los coeficientes Kappa de Cohen y de correlación entre clases basados en el CTT. También, se utilizó el marco GT para determinar la fiabilidad de los evaluadores; se examinaron los componentes de varianza de los evaluadores. Para el análisis de la fiabilidad, se calcularon los coeficientes *G* y *Phi* basándose en la teoría *G*. Para el análisis de los datos, se utilizaron los paquetes informáticos EduG 6.1, SPSS 21.0 y LISREL 7.0.

Finalmente, para el análisis de datos para el establecimiento de normas, siguiendo el método ampliado de Angoff, se calcularon la PPT-PT y los estadísticos descriptivos de los datos recogidos de los expertos.

3. Resultados

3.1. Conclusiones Sobre la Validez del Constructo

La validez de constructo de la escala se evaluó aplicando análisis factoriales exploratorios y confirmatorios. Por razones prácticas, no fue posible aplicar los análisis a dos grupos diferentes de participantes porque el tamaño de la muestra era insuficiente. Por ello, los análisis factoriales exploratorio y confirmatorio se computaron sobre la base del mismo grupo; esto puede considerarse una limitación del estudio. El análisis factorial exploratorio (AFE) y el análisis factorial confirmatorio (AFC) comprenden estadísticas multivariables y requieren algunos supuestos y tamaños de muestra; sin embargo, para los datos basados en el rendimiento de los alumnos (como el rendimiento en instrumentos musicales), no resultaba práctico obtener tamaños de muestra grandes. En el presente estudio, el tamaño de la muestra fue de 100; teniendo en cuenta el número de ítems de la prueba de rendimiento (cinco ítems), parecía suficiente para el AFE y el AFC (Barret y Kline, 1981).

Se utilizó la medida Kaiser-Meyer-Olkin de adecuación del muestreo (KMO); el valor KMO se calculó en 0,83. Según Leech *et al.* (2005), un valor KMO de .80-.90 y superior puede interpretarse como muy bueno. La prueba de Bartlett comprueba si nuestra matriz de correlaciones es diferente de la matriz unitaria. Por lo tanto, una prueba de Bartlett estadísticamente significativa (rechazo de la hipótesis H0) significa que las correlaciones entre las variables difieren de cero. En este estudio, la prueba de Bartlett resultó ser estadísticamente significativa ($p < 0,05$). Tanto los resultados de la prueba de KMO como los de la prueba de esfericidad de Bartlet muestran que los datos son adecuados para el cálculo del análisis factorial. Así pues, el conjunto de datos es adecuado para la factorización. En otras palabras, se puede afirmar que los criterios de la rúbrica (ritmo, tempo, melodía, armonía, técnica) tienen un nivel

suficiente de correlación entre sí para formar un factor. En el presente estudio, no faltaba ningún valor en los datos de la investigación.

Los resultados del AFE indicaron que la escala tenía una estructura de un factor. En total, cinco ítems explicaban el 63% de la varianza total. Para decidir el número de factores, se examinaron el diagrama de scree y el análisis paralelo de Horn. Ambos mostraron que una estructura de un factor era adecuada para la PPT-PT (Figura 1).

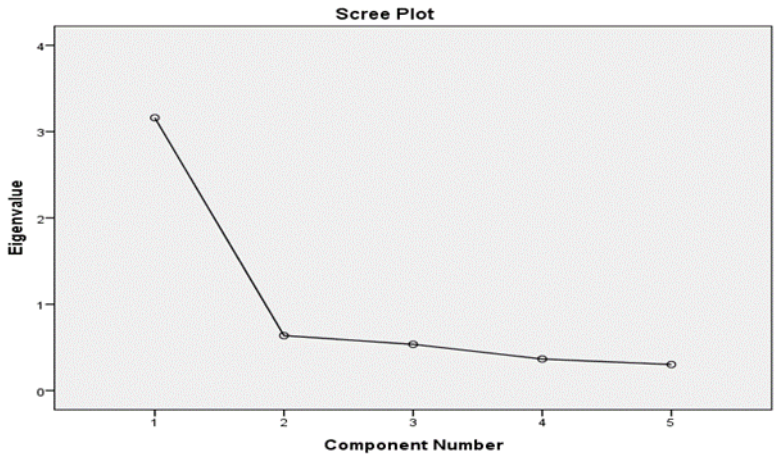

Figura 1. *Scree Plot para el PPT-PT*

El hecho de que se produzca un brusco descenso en el gráfico después del primer valor propio y de que éste se aplane y pierda su pendiente indica que predomina una estructura de un solo factor. En otras palabras, el descenso y el aplanamiento del gráfico indican que los criterios (tempo, ritmo, melodía, armonía, técnica) de la rúbrica son componentes de un único constructo. En la Tabla 4, se presentan las cargas factoriales de los ítems del PPT-PT.

Tabla 4. Cargas Factoriales para el PPT-PT

Criterios	***Cargas***
	Factorial 1
Tempo	.845
Precisión melódica	.820
Precisión rítmica	.788
Precisión armónica	.779
Digitación	.739

Como se muestra en la Tabla 4, todas las cargas factoriales se situaban entre 0,74 y 0,85. Esto demuestra que todos los ítems tenían una fuerte relación con el constructo medido. Esto demuestra que todos los ítems tenían una fuerte relación con el constructo medido. Por lo tanto, no hubo necesidad de omitir ninguno de los ítems ya que, según Tabachnick y Fidell (2007), las cargas factoriales por encima de .40 son lo suficientemente buenas como para mantener los ítems en la escala.

El AFC se utilizó para confirmar la estructura unifactorial de la PPT-PT. La Figura 2 muestra los coeficientes estandarizados de la variable latente a las variables observadas y los valores de t.

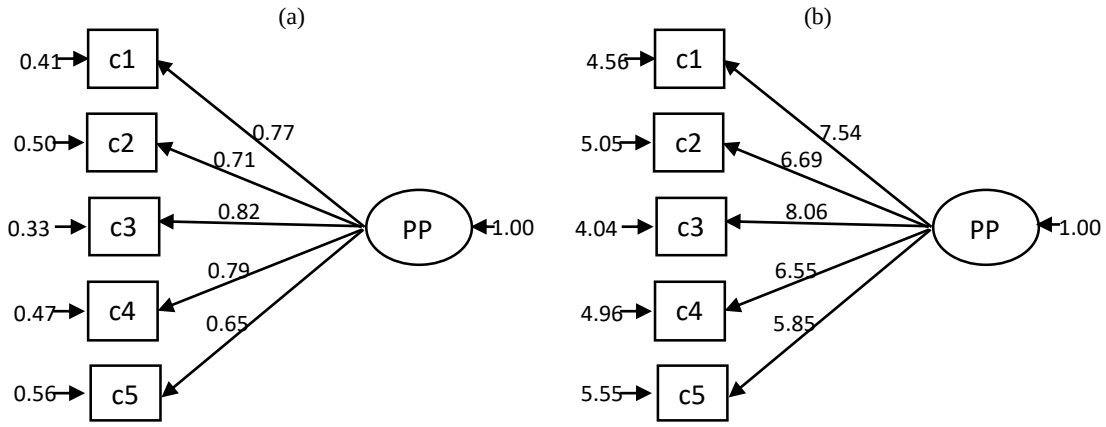

Figura 2. Coeficientes Normalizados y Valores t de la Relación Entre las Variables Latentes y Observadas

Como muestra la Figura 2(a), los coeficientes estandarizados para la relación entre las variables latentes y observadas se situaron entre .65 y .82, respectivamente. La Figura 2(b) muestra que los valores de t para la relación entre las variables latentes y observadas estaban por encima del coeficiente crítico (2,58) y eran estadísticamente significativos al nivel de .01 (los puntos de corte para los índices de ajuste se presentan en la Tabla 5).

Tabla 5. Puntos de Corte de los Indices de Ajuste

Índices de ajuste	Puntos de corte
Chi cuadrado / df	Chi cuadrado /df ≤2.5 excelente ajuste; ≤5 ajuste mediocre
GFI-AGFI- CFI- NNFI	GFI-AGFI- CFI- NNFI ≥ 0.95 excelente ajuste; ≥0.90 buen ajuste
SRMR-RMSEA	SRMR-RMSEA ≤0.05 excelente ajuste; ≤0.08 buen ajuste; ≤0.010 mal ajuste

Nota. df: grados de libertad, GFI: Índice de bondad de ajuste, AGFI: Índice de bondad de ajuste, CFI: Índice de ajuste comparativo, NNFI: Índice de ajuste no normalizado, SRMR: Residuo cuadrático medio normalizado, RMSEA: Error cuadrático medio de aproximación (Tabachnick y Fidell, 2007).

La relación de los grados de libertad se calculó como 3,937/5 = 0,89. Este resultado puede considerarse un indicador de un ajuste excelente (Tabachnick y Fidell, 2007). Los resultados de los demás índices de ajuste fueron los siguientes: índice de ajuste no normalizado (NNFI): .96; índice de ajuste comparativo (CFI): .98; índice de bondad de ajuste (GFI): .95; error cuadrático medio de aproximación (RMSEA): .11; y RMR estandarizado: .04.

Los valores de NNFI, CFI y GFI, al ser superiores a .95, fueron los indicadores de un ajuste excelente (Hooper *et al.*, 2008; Sümer 2000). Los valores de RMSEA y RMR estandarizados inferiores a 0,05 indicaron un ajuste excelente. Los valores de GFI y AGFI, por otro lado, sugerían la presencia de un ajuste mediocre (Brown, 2006; Tabachnick y Fidell, 2007). En este estudio, todos los índices de ajuste mostraron un ajuste excelente, excepto el valor RMSEA.

Cuando se examinaron los resultados en conjunto, se observó que, aunque el valor de RMSEA era relativamente alto, la relación de chi-cuadrado con los grados de libertad y los valores de NNFI, CFI y RMR estandarizado se encontraban en el nivel esperado. Por lo tanto, la estructura de un factor que se determinó como resultado del AFE fue validada por el AFC y la validez de constructo de la escala se situó en el nivel esperado.

Así pues, los resultados de la AFE y la AFC son pruebas estadísticas de que los criterios de la rúbrica (tempo, ritmo, melodía, armonía, técnica) se unen para formar las habilidades para tocar el piano. Como resultado, la suma de las puntuaciones obtenidas a partir de los criterios muestra el nivel de destreza pianística del intérprete.

3.2. Conclusiones sobre la discriminación de ítems

Se calcularon los niveles de discriminación de los ítems de la escala para investigar si existía una diferencia significativa entre las puntuaciones de la escala y las puntuaciones factoriales del 27% superior e inferior del alumnado participante. El 27% superior del alumnado se refiere al grupo de alto rendimiento y el 27% inferior se refiere al grupo de bajo rendimiento. La discriminación de ítems se midió para el PPT-PT mediante la prueba U de Mann-Whitney. Los resultados de esta prueba indican que el grupo superior obtuvo puntuaciones medias más altas para cada ítem de la escala; esta diferencia fue significativa al nivel de 0,01. Además, las correlaciones ítem-total fueron superiores a las del grupo superior. Además, se calcularon las correlaciones ítem-total; se observó que todas las correlaciones sugerían que el PPT-PT era capaz de discriminar los grupos de alto y bajo rendimiento.

3.3. Conclusiones de los Estudios Sobre Coherencia Interna y Generalizabilidad

Se utilizó el alfa de Cronbach para determinar el coeficiente de la consistencia interna del PPT-PT y se calculó como .83. Los resultados indicaron que la consistencia interna de la escala era alta.

Tabla 6. Valores de los Componentes de la Varianza para un Diseño Completamente Cruzado

Fuente	SS	df	MS	Corregido	%
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%

Como se muestra en la Tabla 6, el componente de varianza para el efecto principal estudiante (S objeto de medida) fue de 0,15, lo que explicaba aproximadamente el 30% de la varianza total. Esto demostró que el PPT-PT era capaz de definir las diferencias individuales entre el alumnado. Considerando el efecto principal del ítem, se observó que el componente de varianza fue de 0,07, explicando el 13% de la varianza total. Esto significa que las dificultades de los ítems del PPT-PT estaban diversificadas. Y lo que es más importante, se examinó el efecto principal del evaluador y se observó que el componente de varianza era de 0,0011, lo que no explicaba casi nada de la varianza total. En otras palabras, la cantidad de varianza del evaluador fue muy baja en la varianza total. Esto significa que los evaluadores eran coherentes entre sí.

En cuanto a los efectos de interacción de los compo, en lo que respecta a los efectos de interacción de los componentes de la varianza, estudiante-elemento (SI), estudiante-evaluador (SR) y evaluador-elemento (IR) fueron 0,08 (17%), 0,03 (6,1%) y 0,02 (5,9%), respectivamente. El componente de varianza del efecto de interacción alumno-elemento-evaluador (SIR) fue de 0,14, el segundo componente más alto. Los resultados muestran que el 28% de la varianza total consistió en errores aleatorios. Además, el coeficiente G, que se refería a la varianza relativa del error fue de .76. Este valor era suficiente, teniendo en cuenta que el PPT-PT tenía un número relativamente pequeño de ítems (Diederich, 1973). Además, los estudios de decisión mostraron que, si el número de calificadores aumentaba a 3, 4 y 5, los coeficientes G pasarían a ser de .80, .82 y .83, respectivamente. Por lo tanto, sería mejor utilizar al menos tres calificadores para la escala de ejecución en instrumentos musicales a fin de obtener coeficientes G más elevados.

3.4. Conclusiones sobre el coeficiente de correlación entre clases

Esta sección muestra un coeficiente de correlación interclase (CCI), que indica la fiabilidad de los evaluadores 1 y 2. El CCI entre los evaluadores fue de 0,593 (F364-365 = 3,81), lo que sugiere una fiabilidad interclase moderada. El CCI entre los evaluadores fue de 0,593 (F364-365 = 3,81, $p < 0,01$), lo que sugiere una fiabilidad interevaluadores moderada. Según el intervalo de confianza del 95% de la estimación del CCI, los valores inferiores a 0,50, entre 0,50 y 0,75, entre 0,75 y 0,90 y superiores a 0,90 indican una fiabilidad deficiente, moderada, buena y excelente, respectivamente.

3.5. Conclusiones sobre el estudio normativo

Tabla 7. Resultados del Estudio de Fijación de Normas

Experto Núm.		C1	C2	C3	C4	C5	Media de la Fila	SD de la Fila
1	Ronda 1	5	5	5	5	5	5.00	0.00
	Ronda 2	5	4	4	3	5	4.20	0.84
2	Ronda 1	4	5	4	4	4	4.20	0.45
	Ronda 2	4	5	4	4	4	4.20	0.45
3	Ronda 1	4	4	4	4	4	4.00	0.00
	Ronda 2	4	4	4	3	4	3.80	0.45
4	Ronda 1	5	4	4	3	3	3.80	0.84
	Ronda 2	4	4	4	3	3	3.60	0.55
5	Ronda 1	4	4	5	3	3	3.80	0.84
	Ronda 2	4	4	5	3	3	3.80	0.84
6	Ronda 1	4	5	5	5	3	4.40	0.89
	Ronda 2	4	5	5	5	3	4.40	0.89
7	Ronda 1	5	5	5	4	3	4.40	0.89
	Ronda 2	4	5	5	3	3	4.00	1.00
8	Ronda 1	4	4	5	4	3	4.00	0.71
	Ronda 2	4	4	5	4	3	4.00	0.71
9	Ronda 1	4	4	4	3	2	3.40	0.89
	Ronda 2	4	3	4	3	2	3.20	0.84
	Media de la Columna 1	4.33	4.44	4.55	3.89	3.33	4.11	
	Media de la Columna 2	4.33	4.22	4.44	3.44	3.33	3.95	
	SD de la Columna 1	0.5	0.53	0.53	0.78	0.87	0.64	
	SD de la Columna 2	0.33	0.66	0.53	0.73	0.87	0.62	

C: Criterios
SD: Desviación Estándar

Como se muestra en la Tabla 7, cuando se examinaron las puntuaciones de corte para cada ítem al final de la segunda ronda, se vio que los expertos asignaron el punto de corte más

bajo al ítem 5 (3,33) y el más alto al ítem 3 (4,44). Además, los expertos afirmaron que los alumnos situados en el límite debían tener una media de 3,95 puntos en cada ítem. Se examinaron los valores de la desviación típica para determinar la variabilidad entre las puntuaciones de los expertos; así, se llegó a la conclusión de que la variabilidad entre las opiniones de los expertos era menor en la segunda ronda (0,62) que en la primera (0,64).

Para calcular la puntuación de corte, se tomó la media del total de puntos de cada experto para cada ítem. A continuación, se sumaron las medias para obtener la puntuación de corte: 4,33+4,22+4,44+3,44+3,33 = 19,75.

Según las conclusiones obtenidas de los expertos (utilizando el método de Angoff ampliado), para que un alumno sea considerado suficiente en su rendimiento, debe obtener al menos 19,75 sobre 25 en el PPT-PT. Dado que no es posible obtener puntuaciones decimales a partir de esta escala, la puntuación de corte se redondeó a 20.

4. Discusión y conclusiones

Este estudio tenía como objetivo investigar las propiedades psicométricas y definir la puntuación de corte del PPT-PT utilizando diferentes métodos. Los resultados de los análisis AFE y AFC mostraron que el PPT-PT tenía una estructura de un factor, lo que significa que la escala medía un rasgo latente, es decir, el rendimiento con el instrumento musical (lo que indica suficiente validez de constructo). Por lo tanto, podemos afirmar que la Escala medía puramente el rendimiento con el piano sin implicar otros rasgos. Así pues, con esta escala se pueden evaluar de forma justa las habilidades musicales del alumnado.

El resultado del coeficiente Alfa de Cronbach indica que la consistencia interna de la escala era alta. En otras palabras, los ítems/criterios estaban interrelacionados y era adecuado sumar las puntuaciones de los distintos ítems/criterios. Este estudio emplea una rúbrica analítica para evaluar las interpretaciones musicales del alumnado, lo que implica evaluar subcomportamientos específicos, como el tempo, la precisión melódica, la precisión rítmica, la precisión armónica y la digitación. Cuando se evalúan conjuntamente los resultados del AFC AFE y el valor alfa de Cronbach, puede interpretarse que estos subcomportamientos predichos teóricamente se reúnen de forma coherente y constituyen la interpretación musical. En otras palabras, en cierto sentido se ha confirmado esta hipótesis de la bibliografía.

Los resultados del análisis de discriminación de ítems revelaron que el PPT-PT era bueno para discriminar a los alumnos que tenían niveles altos y bajos de habilidad para tocar un instrumento musical. Esto demuestra una fuerte correlación entre las expresiones delineadas para los subcomponentes de la interpretación musical y su relevancia directa para el arte de la interpretación musical en sí. En consecuencia, el empleo de esta rúbrica puede facilitar la evaluación de las interpretaciones pianísticas de los alumnos utilizando repertorios similares a los ejemplos que aquí se ofrecen.

Los resultados de los estudios de generalizabilidad mostraron que la prueba era capaz de definir las diferencias individuales entre estudiantes. Además, la cantidad de varianza de los calificadores fue muy baja en total, lo que significa que los calificadores fueron coherentes. En otras palabras, hubo acuerdo entre los calificadores cuando puntuaron de forma independiente el rendimiento del alumnado. Sin embargo, la varianza del error aleatorio, que muestra la cantidad de error que interfirió en el proceso de medición, fue relativamente alta. Esto significa que

deberían tenerse en cuenta algunas otras variables (por ejemplo, el tiempo, el tipo de rúbrica, etc.) a la hora de realizar análisis de generalizabilidad en estudios posteriores. Además, el coeficiente *G* indicó un nivel de fiabilidad aceptable para la prueba investigada, que también coincide con el coeficiente Alfa de Cronbach. Sin embargo, los estudios de decisión utilizados indicaban que, para aumentar la fiabilidad, era mejor utilizar tres calificadores. En este estudio, tuvimos dos calificadores; por lo tanto, se recomienda a futuros instructores o investigadores que apliquen el PPT-PT que utilicen al menos tres evaluadores para obtener resultados más fiables.

Los resultados relacionados con el CCI indicaron una fiabilidad entre evaluadores moderada. El nivel de fiabilidad entre evaluadores fue relativamente aceptable. Por otra parte, de forma similar a los resultados del estudio de decisión, se consideró mejor utilizar tres evaluadores que dos. Cuando se evalúan conjuntamente los análisis de generalizabilidad y del coeficiente de correlación intraclase, resulta evidente que las explicaciones proporcionadas para los subcomportamientos en la interpretación pianística fueron comprendidas de forma coherente por varios calificadores. En consecuencia, esto sugiere un alto nivel de precisión en la introducción y definición de los criterios (subcomportamientos) durante el desarrollo de la clave de puntuación.

Los resultados del estudio de fijación de normas sugieren que la puntuación de corte del PPT-PT es relativamente alta. Esto demuestra la importancia de que la interpretación pianística se evalúe como un todo. Por ejemplo, si un alumno toca el ritmo, las notas, los acordes, el tempo, etc., con una precisión parcial, puede obtener una puntuación de 13 sobre 25 (52%) en esta escala. Por lo tanto, un alumno que obtiene una puntuación de 13 sobre 25 (52%) puede considerarse aprobado. Sin embargo, en este caso, la pieza interpretada sería incomprensible para el público. Además, en este estudio, los participantes eran candidatos a profesores, de quienes se espera que enseñen estas piezas a sus alumnos. Por lo tanto, para aprobar el examen era necesario obtener al menos 20 de 25 puntos (80%) en el PPT-PT.

Este estudio demuestra que el PPT-PT es válido y fiable y puede utilizarse para evaluar la interpretación pianística de estudiantes universitarios con una puntuación de corte de 20 de 25 (80%). Por lo tanto, se aconseja que los profesores utilicen esta escala para evaluar el rendimiento del alumnado. Sin embargo, se recomienda encarecidamente utilizar un mínimo de 20 como puntuación de corte. Además, los instructores obtendrán resultados más fiables si utilizan al menos tres calificadores. No obstante, no hay que olvidar que los calificadores deben tener experiencia y estar bien cualificados. De lo contrario, la fiabilidad de la escala puede disminuir (Kamış y Doğan, 2018).

Los futuros investigadores pueden repetir estudios similares con más muestras y calificadores. Además, pueden utilizar otras técnicas para los análisis de establecimiento de normas, como Bookmark, Nedelsky, etc. (Cizek y Bunch, 2007). Además, si los investigadores disponen de una muestra mayor (más de 250 participantes), se aconseja utilizar el modelo Rasch multifacético para probar la validez y fiabilidad del PPT-PT u otras escalas desarrolladas.

Referencias

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

Crocker, L. y 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. y Silverman, M. (2015). *Music matters: A Philosophy of music education*. Oxford University Press.

Fraenkel, J.R., Wallen, N.E. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y Kaufman, B. (2019). *Music Learning as Youth Development*. https://doi.org/10.4324/9780429436956-11

Shavelson, R.J. y 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. y 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. y 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. y 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. y Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson Education.

Thompson, S. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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

ANEXO A - Prueba de interpretación pianística para profesorado en formación

Rúbrica analítica

Nombre del/de la estudiante: **Departamento:** **..../..../ 20..**

	5 Puntos	4 Puntos	3 Puntos	2 Puntos	1 Punto
Tempo	**Tempo considerado y uniforme** *Tocó al tempo adecuado de principio a fin.*	**El ritmo es uniforme** *Tocó al ritmo adecuado, con algunas vacilaciones.*	**El ritmo es casi constante** *Hace pausas mientras toca o vacila mucho*	**El tempo es mayormente errático y desigual** *Pausa muchas veces mientras toca o el Tempo era inestable debido a la vacilación.*	**No pude completar la pieza.**
Precisión rítmica	**El ritmo es coherente y tiene sentido musical** *Tocó los ritmos completamente bien*	**El ritmo es coherente en su mayor parte y tiene sentido musical** *Comete pequeños errores rítmicos en uno o dos compases (menos del 10%) durante la interpretación*	**El ritmo es a menudo coherente y tiene poco sentido musical** *Cometió de tres a seis errores menores (11% a 30%) o se detuvo una vez durante la interpretación*	**El ritmo es incoherente y no tiene sentido musical.** *Cometió de siete a diez errores rítmicos menores (31% a 50%) o se detuvo varias veces (más de dos veces) durante la interpretación.*	**El ritmo era casi imperceptible**
Precisión melódica	**La melodía se siente completa y coherente; la pieza tiene sentido musical** *Tocó la melodía completamente bien*	**La melodía parece menos completa y coherente; la pieza tiene menos sentido musical.** *Ha tocado una o dos notas incorrectas (menos del 10%) durante la interpretación.*	**La melodía parece musicalmente incompleta** *Tocó tres-cinco notas incorrectas (11% a 30%) durante la interpretación.*	**La melodía no es completa ni coherente** *Tocó de seis a diez (31% a 50%) notas incorrectas durante la interpretación.*	**La melodía era casi imperceptible**
Precisión armónica	**Los acordes son evidentes y se tocan musicalmente** *Tocó todos los acordes completamente bien*	**Los acordes se utilizan de manera que apoyen musicalmente la pieza** *Toca uno o dos acordes incorrectos (menos de 10%) o voces incorrectas en los acordes.*	**Los acordes no siempre se alinean con la melodía** *Ha tocado de tres a cinco acordes incorrectos (11% a 30%) o voces incorrectas en los acordes.*	**Se utilizan acordes, pero no se alinean adecuadamente con la melodía** *Tocó de seis a diez acordes incorrectos (31% a 50%) o voces incorrectas en los acordes.*	**Los acordes eran casi imperceptibles**
Digitación	**Tocó toda la pieza con la digitación correcta** *Tocó toda la pieza con la digitación correcta.*	**Casi toda la digitación se tocó correctamente** *Tocó con una o dos digitaciones incorrectas (menos del 10%) durante la actuación*	**Parte de la digitación fue correcta, pero la impresión general fue incoherente** *Tocó de tres a cinco digitaciones incorrectas (11% a 30%) durante la interpretación*	**No se respetó la digitación** *Tocó de seis a diez digitaciones incorrectas (31% a 50%) durante la interpretación*	**Demostró una falta de comprensión de la digitación adecuada**

Rendimiento excepcional que no puede ser evaluado por el PPT-PT:

Debilidades del alumno que no pueden ser evaluadas por el PPT-PT:

ANEXO –B FORMULARİO DE DİCTAMEN PERİCİAL DEL MÉTODO ANGOFF AMPLİADO

Experto ID: _____

RONDA 1

Puntúe de 1 a 5 puntos la actuación del examinando límite para cada uno de los criterios que se indican a continuación

CRITERIOS	**Tempo (Crıterıos 1)**	**Ryhtmic Accuracy (Crıterıos 2)**	**Melodic Accuracy (Crıterıos 3)**	**Harmonic Accuracy (Crıterıos 4)**	**Fingering (Crıterıos 5)**
PUNTOS					

Revele sus estimaciones del rendimiento del examinando límite y discuta cualquier desacuerdo entre evaluadores.

RONDA 2

Evalúe de 1 a 5 puntos la actuación del examinando fronterizo con respecto a la discusión anterior para cada uno de los criterios que se indican a continuación.

CRITERIOS	**Tempo (Crıterıos 1)**	**Ryhtmic Accuracy (Crıterıos 2)**	**Melodic Accuracy (Crıterıos 3)**	**Harmonic Accuracy (Crıterıos 4)**	**Fingering (Crıterıos 5)**
PUNTOS					

Estadísticas descriptivas de la 1ª y 2ª ronda (Esta parte será cumplimentada por el investigador)

	Puntuación mínima		Puntuación máxima		Moda		Mediana		Promedio	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
Crıterıos 1										
Crıterıos 2										
Crıterıos 3										
Crıterıos 4										
Crıterıos 5										

R1: Ronda 1 R2: Ronda 2