Revista de Psicodidáctica, 2019, 24 (1), 31–38

Revista de Psicodidáctica

www.elsevier.es/psicod

Original

Evaluación de la calidad de programas de educación patrimonial: construcción y calibración de la escala Q-Edutage

Olaia Fontal Merillas [a], Silvia García Ceballos [a], Benito Arias [b], y Víctor B. Arias [c,d,*]

[a] Departamento de Didáctica de la Expresión Musical, Plástica y Corporal, Facultad de Educación y Trabajo Social, Universidad de Valladolid, Valladolid, España
[b] Departamento de Psicología, Facultad de Educación y Trabajo Social, Universidad de Valladolid, Valladolid, España
[c] Departamento de Evaluación, Personalidad y Tratamiento Psicológico, Facultad de Psicología, Universidad de Salamanca, Salamanca, España
[d] Instituto Universitario de Integración en la Comunidad (INICO), Facultad de Psicología, Universidad de Salamanca, Salamanca, España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 29 de mayo de 2018
Aceptado el 1 de agosto de 2018
On-line el 9 de octubre de 2018

Palabras clave:
Evaluación
Educación patrimonial
Calidad
Programas educativos

RESUMEN

Mejorar la evaluación de la calidad de los programas educativos es uno de los principales objetivos de investigación en educación patrimonial. Sin embargo, no se dispone de un instrumento que sea breve, objetivo y que permita el uso de un estándar común para la comparación insesgada de la calidad entre distintos programas. El objetivo de este estudio ha sido el diseño y construcción de un instrumento para la evaluación de la calidad de programas de educación patrimonial, que mantenga un equilibrio adecuado entre precisión y brevedad, y pueda ser utilizado tanto en solitario (p. ej., con propósitos de cribado cuando el número de programas a evaluar es elevado), como de apoyo a sistemas de evaluación más amplios. Se identifican indicadores de calidad relevantes, de acuerdo a la investigación previa y a las valoraciones de 17 expertos, dando como resultado 14 indicadores de calidad que son calibrados mediante modelos de la Teoría de la Respuesta al Ítem, a partir de la evaluación de 330 programas de educación patrimonial. La escala es capaz de discriminar con precisión entre varios niveles de calidad (i.e., muy bajo, bajo, medio, alto y muy alto), aporta un buen nivel de información a lo largo de una zona amplia de la variable, y produce puntuaciones insesgadas entre distintos evaluadores. La escala Q-Edutage supone un aporte relevante que contribuye a mejorar el rigor de la evaluación y la planificación de programas en el ámbito de la educación patrimonial.

Assessing the Quality of Heritage Education Programs: Construction and Calibration of the Q-Edutage Scale

Keywords:
Assessment
Heritage education
Quality
Educational programs

ABSTRACT

Improving the assessment of the quality of educational programs is one of the main objectives of research in heritage education. However, we do not have an instrument that is brief, objective and allows the use of a common standard for unbiased quality comparison between different programs. The objective of this study has been to design and develop a tool for the quality assessment of heritage education programs, which maintains an appropriate balance between accuracy and brevity, and can be used both on its own (e.g., for screening purposes when the number of programs to be evaluated is high) and to support broader assessment systems. Relevant quality indicators were identified, according to previous research and evaluations by 17 experts, resulting in 14 quality indicators that were calibrated using Item Response Theory models from the assessment of 330 heritage education programs. The scale was able to discriminate with high precision between various levels of quality (i.e., very low, low, medium, high and very high), provided a good level of information over a wide area of the variable, and produced unbiased scores among different evaluators. The Q-Edutage scale is an relevant addition that contributes to improving the rigour of evaluation and program planning in the field of heritage education.

* Autor para correspondencia.
Correo electrónico: vbarias@usal.es (V.B. Arias).

https://doi.org/10.1016/j.psicod.2018.07.003
1136-1034/

Introducción

La educación patrimonial se ha desarrollado en las dos últimas décadas como una disciplina de investigación transversal y con proyección internacional, atendiendo al número y calidad de las tesis doctorales, proyectos competitivos y artículos científicos (Fontal e Ibáñez-Etxeberria, 2017). Las líneas de investigación han evolucionado superando los enfoques iniciales, centrados en la didáctica del patrimonio y de carácter fundamentalmente descriptivo, hasta la educación patrimonial (Cobaleda, 2016; Fontal, 2003) con una proyección evaluativa centrada en la medición de la calidad de los programas de educación patrimonial, tanto en su diseño como en los resultados de su implementación (Fontal, 2016b; Martín-Cáceres y Cuenca, 2016).

Como resultado del desarrollo de la disciplina, en los últimos diez años se ha producido un aumento sustancial de investigaciones en educación patrimonial (Fontal e Ibáñez-Etxeberria, 2017) en tres grandes áreas temáticas: (1) el análisis de la normativa educativa (Fontal, Ibáñez-Etxeberria, Martínez, y Rivero, 2017; Potočnik, 2017); (2) la accesibilidad, diversidad e inclusión (Deng, 2015; Marín-Cepeda, García-Ceballos, Vicent, Gillate, y Gómez-Redondo, 2017); y (3) el uso de las Tecnologías de la Información y la Comunicación como recurso y como contexto para la educación patrimonial (Agrusti, Poce, y Re, 2017; Cozzani, Pozzi, Dagnino, Katos, y Katsouli, 2017; Ibáñez-Etxeberria, Fontal, y Rivero, en prensa).

Según su enfoque metodológico, los estudios se pueden clasificar en tres grandes genealogías (Fontal e Ibáñez-Etxeberria, 2017): (a) *Genealogía de investigación (Re)Conceptualizante*, centrada en la epistemología de la educación patrimonial (Fontal y Juanola, 2015; Klein y Van Boxtel, 2011) en la que se recogen modelos, procesos o definiciones teóricas cuyo objetivo es configurar un corpus conceptual (Gürçayir, 2013); (b) *Genealogía de investigación didáctico-contextual*, centrada en la didáctica del patrimonio en contextos formales (Fontal, 2016c) y no formales (Calaf, Gillate, y Gutiérrez, 2015) con atención a los procesos de interpretación, comunicación y difusión del patrimonio (Kitungulu, 2015; Martín-Cáceres y Cuenca, 2011); y (c) *Genealogía de investigación evaluativa*, que se ocupa de la evaluación de programas educativos y, con menor frecuencia, de los aprendizajes (Domínguez y López, 2017; Tsai, 2011).

El presente estudio se enmarca en el ámbito de la genealogía de la investigación evaluativa, dado que su objetivo es la construcción de un instrumento para la evaluación objetiva de la calidad de programas de educación patrimonial. Esta labor se ha articulado desde el Observatorio de Educación Patrimonial en España (OEPE), en el marco del «método secuencial de análisis y evaluación de los programas de educación patrimonial» (SAEPEP-OEPE; Fontal, 2016a), que ya ha sido implementado en diversos estudios (Gómez-Redondo, Calaf, y Fontal, 2017; Marín-Cepeda et al., 2017; Rivero, Fontal, García-Ceballos, y Martínez, 2018).

El método SAEPEP–OEPE para la evaluación de programas de educación patrimonial

El SAEPEP-OEPE es un método secuencial de análisis y evaluación de programas de educación patrimonial (Fontal, 2016a). Su objetivo es evaluar la calidad de los programas a través de un sistema de filtros secuenciados que permite seleccionar los mejores casos sobre la base de estándares definidos a partir de: (1) textos normativos, (2) resultados de evaluaciones previas e (3) indicadores de éxito extraídos de estudios de caso. El método se articula en ocho fases, cada una asociada al uso de determinados instrumentos de evaluación (Fontal y Juanola, 2015; para una descripción detallada del método, ver Fontal, 2016a). La fase 1 consiste en la búsqueda y localización de programas. En la fase 2, los programas que cumplan los criterios de inclusión son incorporados a la base de datos del OEPE. En la fase 3, se recopila información de los programas seleccionados conforme a 42 campos de registro relacionados con la identificación, localización, descripción del proyecto, diseño educativo y anexo documental. En la fase 4 se realiza un análisis descriptivo de los datos inventariados. Posteriormente, se suceden dos fases relacionadas con la evaluación de programas: la fase 5, que atiende a estándares básicos, y la fase 6, que atiende a estándares específicos. Por último, las fases 7 y 8 consisten en una evaluación pormenorizada de los programas con mejores resultados en las anteriores fases, mediante estudios de caso únicos o múltiples, según corresponda (Simons, 2011; Stake, 2010) y/o la evaluación de aprendizajes (Stake y Munson, 2008). El método SAEPEP-OEPE contribuye a suplir la actual escasez de procedimientos estructurados para la evaluación de la calidad de los diseños en educación patrimonial (Fontal, 2016a).

El presente estudio

Si bien existen antecedentes solventes en evaluación de programas (Calaf, San Fabián, y Gutiérrez, 2017; Vicent, Ibáñez-Etxeberria, y Asensio, 2015), no existe un instrumento para la evaluación de programas de educación patrimonial que sea breve, objetivo, no dependiente de las características individuales del evaluador, con propiedades métricas robustas, y que permita el uso de un estándar común para la comparación insesgada de la calidad entre programas de educación patrimonial. Disponer de un instrumento como el descrito facilitaría (a) la evaluación precisa y objetiva de programas, (b) la realización rápida de cribados de los mejores programas para su posterior evaluación en profundidad y (c) la comunicación entre investigadores e instituciones dedicadas a la evaluación de la calidad de la educación patrimonial.

El objetivo de este estudio es el diseño y construcción de un instrumento para la evaluación de la calidad de programas de educación patrimonial que mantenga un equilibrio adecuado entre precisión y brevedad, y que pueda ser utilizado tanto en solitario (p. ej., con propósitos de cribado cuando el número de programas a evaluar es elevado), como en apoyo a sistemas de evaluación más amplios (tales como el método SAEPEP-OEPE descrito en la sección anterior). Para ello, se diseña la escala Q-Edutage en tres pasos: (1) identificación mediante la revisión de la literatura relevante de los principales indicadores básicos de calidad de los programas de educación patrimonial; (2) selección de los indicadores con mayores garantías de validez de contenido mediante un estudio de jueces expertos; y (3) la calibración y construcción de la versión final del instrumento mediante procedimientos enmarcados en la Teoría de la Respuesta al Ítem (TRI).

Método

Participantes

La muestra de calibración del instrumento consiste en 330 programas seleccionados aleatoriamente a partir de los 1719 programas de educación patrimonial actualmente registrados en la base de datos de OEPE. En la muestra se recogen 16 tipos de programa, siendo los más frecuentes los proyectos educativos (20.6%), los diseños didácticos (14.5%), las herramientas didácticas (13%) y los proyectos de investigación (9.1%). Se extraen tres submuestras aleatorias que son asignadas a tres evaluadores expertos, quienes reciben previamente un breve entrenamiento en el uso de la rúbrica de evaluación de los ítems. Los evaluadores son académicos (dos catedráticos y un titular) del ámbito de la Didáctica de la Expresión Plástica, la Didáctica de las Ciencias Sociales y Psicología.

Tabla 1
Codificación de las variables en función de los estándares de calidad

Ítem	Estándar	Codificación
i01	Datos de contacto con la dirección y/o equipo de diseño, planificación e implementación	Contacto
i02	Descriptores que definen el programa	Descriptores
i03	Concepción holística del patrimonio en su naturaleza (material e inmaterial) y en sus cualidades (arqueológico, histórico, documental, artístico, etc.)	Patrimonio
i04	Especificación del tipo/tipología de proyecto desarrollado (programa educativo, proyecto educativo, diseño educativo, acción educativa, actividad aislada etc.)	Tipología
i05	Descripción de las bases, principios y criterios sobre los que se establece el programa	Criterios
i06	Concreción del público al que va dirigido	Público
i07	Incorporación de anexos documentales (memoria, imágenes, vídeos, materiales didácticos empleados, etc.)	Anexos
i08	Justificación del proyecto	Justificación
i09	Descripción de los objetivos a lograr en el desarrollo del programa	Objetivos
i10	Presentación de contenidos abordados en el programa	Contenidos
i11	Orientación metodológica y estrategias de enseñanza aprendizaje	Metodología
i12	Definición de recursos, formatos, soportes y tecnología empleados	Recursos
i13	Determinación de los sistemas o herramientas de evaluación	Evaluación
i14	Medición del impacto y repercusión de la propuesta	Impacto y repercusión

Instrumento

La primera fase de construcción del instrumento consiste en la revisión de la literatura sobre evaluación de la calidad en educación patrimonial (Web of Science y Scopus). Se revisan un total de 311 artículos encontrados mediante los descriptores «heritage», «education» y «quality». Esta búsqueda es posteriormente acotada por el descriptor «standard», obteniendo un total de 29 artículos, de los cuales solo seis son relevantes para los objetivos del estudio. Para la selección del primer conjunto de indicadores se sigue el modelo propuesto por Stake (2006), donde se busca el valor ideal de los objetivos a conseguir, el enfoque racional por encima del intuitivo y la especificidad de los estándares mediante una hoja de control que permite el control del sesgo. Así mismo, se realizan tres análisis de contenido sobre diferentes muestras extraídas de la base de datos del OEPE (N = 350, 644 y 1120 programas) teniendo en consideración los criterios metodológicos derivados del Plan Nacional de Educación y Patrimonio. De esta fase se construye un primer conjunto de 21 indicadores de calidad.

En la segunda fase, 17 jueces expertos evalúan la relevancia de los 21 indicadores. Los expertos son académicos pertenecientes a áreas relacionadas directa o transversalmente con la educación patrimonial (Didáctica de la Expresión Plástica, Didáctica de las Ciencias Sociales, Psicología, Didáctica y Organización Escolar, Didáctica de la Expresión Corporal, Didáctica de la Lengua y la Literatura, Música, Sociología y Expresión gráfica arquitectónica). Los expertos evalúan los indicadores de acuerdo a su coherencia, relevancia y congruencia respecto del objeto de evaluación en una escala de 4 puntos, así como la claridad en su redacción, formato y extensión (Bolívar, 2013; Corral, 2009). Dada la naturaleza ordinal de la escala de medida, se calculan las medianas de las puntuaciones otorgadas por los jueces a la coherencia (Md = 4), relevancia (Md = 4) y congruencia (Md = 3) de los ítems. La concordancia de los jueces se evalúa mediante el coeficiente ponderado de Bangdiwala B^W_N (Bangdiwala, 1987) obteniéndose valores muy satisfactorios (B^W_N = .879 para coherencia, B^W_N = .912 para relevancia y B^W_N = .889 para congruencia). A partir de los resultados de esta fase se eliminan seis indicadores. Se considera que cuatro son altamente redundantes en el contenido, y que dos corresponden no a estándares generales de calidad sino a estándares específicos. El contenido de los 14 indicadores seleccionados para la primera versión de la escala se resume en la Tabla 1 (el formato completo de los ítems y la rúbrica de puntuación pueden ser solicitados al primer autor). Para puntuar cada indicador se construye una escala de clasificación de cuatro categorías ordenadas («*no se alcanza*», «*se alcanza con condiciones*», «*se alcanza*» y «*se alcanza con calidad*»).

Análisis de datos

Los datos se analizan con el programa IRTPRO 4.0 (Cai, Thissen, y du Toit, 2015), utilizando el modelo de respuesta graduada o *Graded Response Model* (*GRM*) (Samejima, 1997). El GRM asume, además de los supuestos usuales en la TRI, que las categorías a las que el individuo responde (o en las que el programa es calificado, como es el caso) pueden ordenarse o jerarquizarse como sucede, por ejemplo, con las escalas probabilísticas de estimaciones sumatorias o «tipo Likert». El GRM especifica la probabilidad de un programa de ser calificado con una categoría i_k como función del nivel del programa en la variable latente (θ_j), el parámetro de localización de la categoría de respuesta k (β_{ik}), y el parámetro de discriminación del ítem (α_i).

El objetivo de la calibración es garantizar que el test fuera máximamente preciso en zonas medias y altas de la variable latente (calidad). Los criterios de retención de ítems son: (a) presentar una capacidad de discriminación al menos moderada (i.e., parámetro alfa mayor a .65, de acuerdo a la clasificación de Baker (2001); (b) que el cumplimiento del ítem no resultase en exceso fácil (i.e., que el parámetro $\beta 2$ –correspondiente al umbral de paso de la categoría de respuesta «se alcanza con condiciones» hacia la categoría «se alcanza»– no presentase un valor theta sustancialmente inferior a -1); (c) que no presentasen problemas de independencia condicional; (d) que el ajuste del ítem al modelo fuera adecuado (i.e., que sus frecuencias observada y esperada no fueran significativamente distintas (p < .01); y (e) que la estimación de los parámetros del ítem fuera suficientemente precisa (i.e., con un error estándar de estimación menor a .30, según lo sugerido por Tay, Meade, y Cao, 2014).

Resultados

Comprobación de la unidimensionalidad e independencia local

La unidimensionalidad y la independencia local son dos requisitos básicos en la TRI. Para garantizar el suficiente cumplimiento de ambos supuestos, se siguen las siguientes estrategias: (a) se realiza un análisis paralelo optimizado (Timmerman y Lorenzo-Seva, 2011) basado en análisis factorial de rango mínimo (MRFA, implementado en el programa FACTOR 10.8.03; Lorenzo-Seva y Ferrando, 2006), donde se compara la estructura de la matriz de correlaciones policóricas de los 14 ítems con los resultados de 500 matrices permutadas a partir de los datos brutos; (b) se estiman dos de los índices de cercanía a la unidimensionalidad recomendados por Ferrando y Lorenzo-Seva (2017): la varianza común

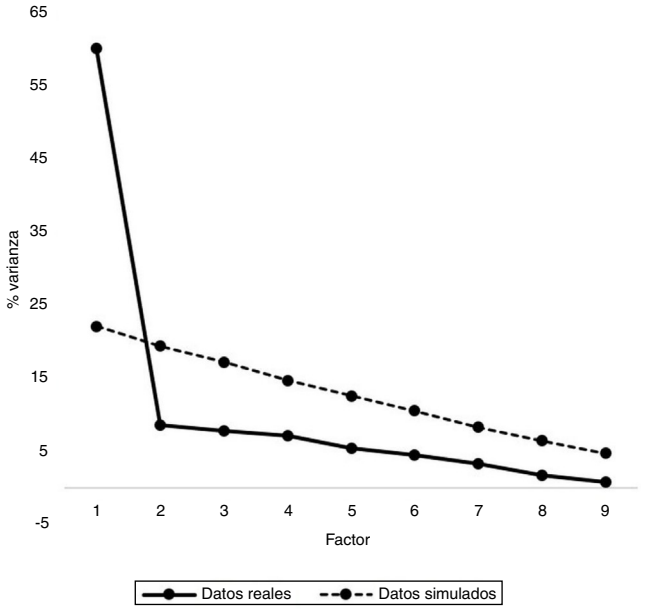

Figura 1. Resultados del análisis paralelo.

explicada (ECV) y la media de las cargas residuales absolutas de los ítems (MIREAL). ECV estima el tamaño del factor dominante respecto al total de varianza común; valores comprendidos entre .70 y .85 son indicadores de que la estructura de los datos es básicamente unidimensional (Rodríguez, Reise, y Haviland, 2016). MIREAL es la media de las cargas absolutas de un potencial segundo factor MRFA residual, ortogonal al factor primario. En consecuencia, MIREAL es un estimador del grado en que la estructura de los datos se desvía de la unidimensionalidad, dado que la presencia de un factor dominante no necesariamente equivale a la ausencia de factores residuales con potencial relevancia sustantiva. Como regla general, un MIREAL menor a .30 sugiere ausencia de un factor residual relevante (Ferrando y Lorenzo-Seva, 2017); y (c) la inspección de los valores estandarizados LD χ^2 para cada par de ítems. La independencia condicional requiere que la mayor parte de los valores LD χ^2 sean menores a 10 (Cai, Thissen, y du Toit, 2011).

La escala adquiere un valor ECV de .88, sugiriendo la presencia de un factor claramente dominante. El valor MIREAL fue de .23, sugiriendo que no es plausible la presencia de varianza sistemática relevante más allá del factor principal. El análisis paralelo (Figura 1) sugiere la retención de un único factor, en cuanto que la varianza capturada por el primer factor es superior a la derivada de las matrices simuladas (centil 95), y la varianza capturada por el segundo factor en los datos reales es en todos los casos inferior a la calculada a partir de las matrices aleatorias. Los valores LD χ^2 estandarizados son en todos los casos inferiores a 10, excepto en el par de ítems 5 («descripción de las bases, principios y criterios sobre los que se establece el programa») y 8 («justificación del proyecto»), con un valor LD χ^2 de 48. La escala adquiere un nivel adecuado de consistencia interna (alfa de Cronbach = .89; alfa ordinal = .91; theta ordinal = .92).

Estimación de los parámetros del modelo

Para estimar los parámetros α_i y β_{ik} de los ítems se utiliza un método de máxima verosimilitud marginal, cuyos resultados se muestran en la Tabla 2.

Los ítems presentan parámetros de discriminación comprendidos entre 1.00 (ítem 4) y 2.63 (ítem 8), de los que tres están en un rango de discriminación moderado (ítems 1, 2 y 4), tres alto

Tabla 2
Parámetros del modelo

Item	α	β1	β2	β3
1	1.04 (.14)	-2.58	-0.53	1.41
2	1.09 (.15)	-3.08	-1.02	1.00
3	1.14 (.15)	-0.82	.88	2.69
4	1.00 (.14)	-2.96	-0.97	.93
5	2.37 (.28)	-0.44	.44	1.32
6	1.18 (.15)	-1.36	.85	2.03
7	1.35 (.17)	0.52	1.45	2.28
8	2.63 (.29)	-0.36	.46	1.46
9	2.06 (.24)	-0.56	.39	1.89
10	2.57 (.29)	-0.47	.40	1.08
11	2.45 (.28)	-0.27	.45	1.25
12	2.10 (.25)	-0.22	.83	1.48
13	1.13 (.17)	1.30	2.22	3.59
14	1.19 (.16)	.18	1.68	3.41

El error estándar de estimación se muestra entre paréntesis.

(ítems 6, 3 y 14) y ocho muy alto (13, 7, 5, 9, 12, 8, 11 y 10). Los errores estándar (M = .20) sugieren que los parámetros de discriminación son en esta muestra estimados con alta precisión (Tay et al., 2014). Atendiendo al parámetro b1 (umbral entre las categorías «*no cumple*» y «*cumple con condiciones*»), los ítems se distribuyen entre zonas muy bajas de la variable (β1= -3.08, ítem 2) y relativamente altas (β1 = 1.30, ítem 13). Los parámetros β2 (umbral entre las categorías «*cumple con condiciones*» y «*cumple*») se localizan entre zonas bajas de la variable (β2 = -1.02, ítem 2) y muy altas (β2 = 2.22, ítem 13). Los errores de estimación de los parámetros de localización son reducidos (M = .19 para β1 y M = .14 para β2, y M = .20 para β3).

Se considera eliminar uno de los ítems del par que presenta problemas de independencia condicional (ítems 5 y 8). Sin embargo, considerando que (a) ambos ítems presentan diferencias de contenido conceptualmente relevantes, y que (b) aportan información no redundante a la medida, se opta por conservar ambos.

La Figura 2 muestra la curva de información del test y la distribución del error estándar de medida. La curva de información indica en qué rango de la variable latente theta (M = 0, DT = 1) el test es máximamente informativo. La zona productiva para la medida se encuentra aproximadamente entre -.1.3 y +2.4 DT (puntos en los que se cortan las curvas de información y del error de medida), con un pico de máxima información comprendido entre aproximadamente -.5 y +1.7 desviaciones típicas.

La Figura 3 muestra la curva característica del test. La curva representa la relación entre las puntuaciones directas esperadas y la puntuación en la variable latente. Como se esperaba dados los objetivos del instrumento, la escala no discrimina adecuadamente en niveles bajos de la variable, en cuanto que en un rango aproximado de -3 DT a -1.5 DT los cambios en la variable latente prácticamente no producen cambios en la puntuación observada. Por el contrario, a partir de la media hasta aproximadamente 2.5 DT, la pendiente se hace sensiblemente más pronunciada.

Ajuste de los datos al modelo de respuesta graduada

Examinamos la magnitud de los errores estándar (menores valores indican mejor precisión en la estimación de los parámetros), los estadísticos M_2 y su valor RMSEA asociado (valores M_2 no significativos y valores RMSEA próximos a cero sugieren buen ajuste de los datos al modelo; Maydeu-Olivares y Joe, 2006), y la significatividad de las diferencias entre las frecuencias de respuesta al ítem observadas y esperadas para cada ítem mediante S-χ^2 (para un buen ajuste, se espera que la mayor parte de los ítems adquieran valores S-χ^2 no significativos; Orlando y Thissen, 2000).

M_2 (1505, gl = 805) resulta estadísticamente significativo (p < .0001), sugiriendo la presencia de cierto grado de desajuste. Sin embargo, el valor RMSEA asociado (.03) sugiere que el desajuste se

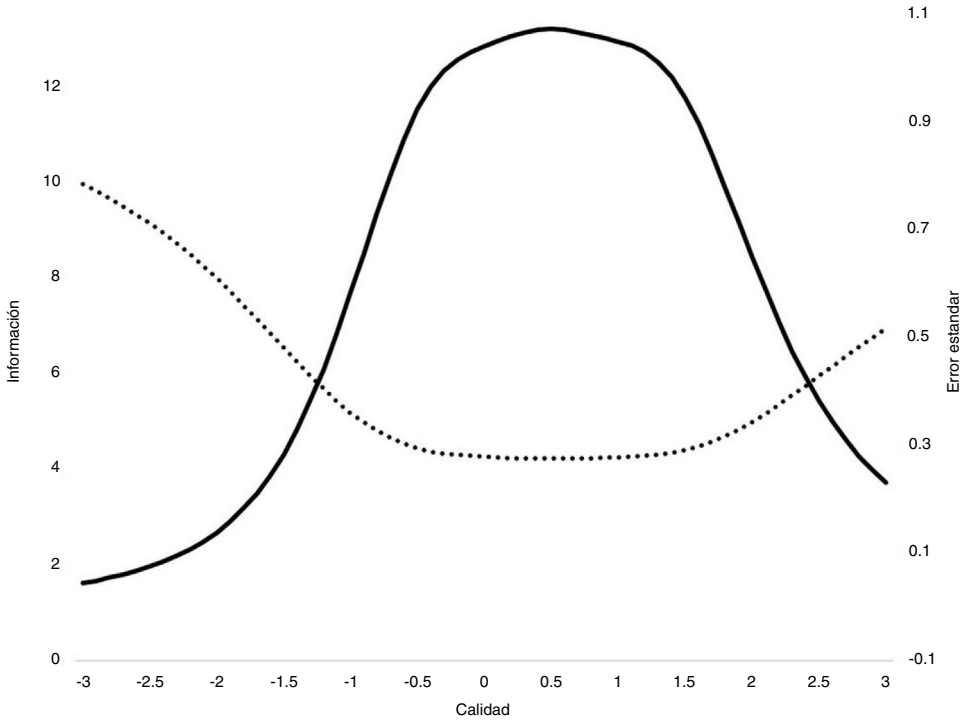

Figura 2. Curva de información del test.

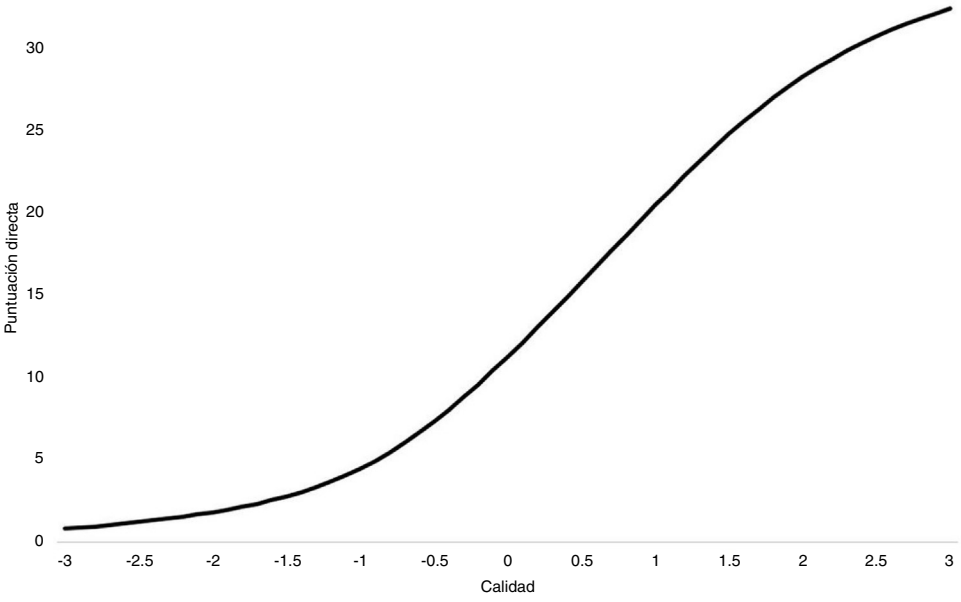

Figura 3. Curva característica del test.

debe a la presencia de una cantidad limitada de error no modelado (Cai et al., 2011). Los errores estándar de estimación son pequeños, indicando que los parámetros son estimados con precisión alta en el caso del parámetro alfa o muy alta en los parámetros beta (Tay et al., 2014). Por último, no se observan diferencias significativas entre las frecuencias observadas y las esperadas por el modelo, en cuanto que ningún valor $S\text{-}\chi^2$ resulta significativo ($p < .05$).

Independencia de la medida respecto al evaluador

Una característica indispensable en un instrumento de medida de la calidad es que, asumiendo un uso correcto del test, este opere de forma similar independientemente del evaluador. En consecuencia, las puntuaciones en el test deberían ser función de la interacción entre las propiedades evaluadas (calidad de los programas en este caso) y las propiedades métricas de los ítems, dependiendo en menor medida de elementos ajenos al constructo de interés. A fin de valorar hasta qué punto el instrumento opera de forma similar entre evaluadores, se estima el funcionamiento diferencial de los ítems (DIF) entre las puntuaciones obtenidas por cada uno de los tres expertos que participan en la toma de datos. Para ello, utilizamos el test de Wald. En este procedimiento primero se obtiene la significatividad estadística ($p < .05$) de las diferencias entre los parámetros estimados a partir de los datos obtenidos por

Tabla 3
Resultados del análisis del funcionamiento diferencial del ítem

		Primera iteración		Segunda iteración	
Ítem	Contraste	Total χ^2	p	Total χ^2	p
1	Experto 2 vs. experto 3	1.4	.7149	0	1
2		1.4	.7073	0	1
3		4	.2579	0	1
4		2.4	.4923	0	1
5		2.1	.5474	0	1
6		5.9	.1143	0	1
7		0.9	.8273	0	1
8		2.9	.4119	0	1
9		1.2	.7532	0	1
10		8.6	.0356	8.1	.0435
11		6.6	.0843	0	1
1	Experto 1 vs. experto 2	15.1	.0018	14.2	.0026
2		3.8	.2861	0	1
3		3.7	.3029	0	1
4		1.3	.7334	0	1
5		9.1	.0283	8.9	.0302
6		3.7	.2999	0	1
7		1.9	.5843	0	1
8		3.3	.3514	0	1
9		3.5	.3185	0	1
10		1.2	.7444	0	1
11		1.2	.7453	0	1
1	Experto 1 vs. experto 3	11.2	.0107	1.9	.5876
2		5.2	.1571	0	1
3		9.7	.021	5.6	.1308
4		2.3	.5076	0	1
5		12.5	.0058	5.7	.1254
6		9.8	.0208	5.9	.1167
7		0.5	.9103	0	1
8		2	.5726	0	1
9		2.2	.5373	0	1
10		9.8	.0206	9.2	.0268
11		4.5	.2147	0	1

Los ítems sospechosos de DIF están marcados en negrita.

cada experto. Los ítems que en esta fase han resultado invariantes se emplean en una segunda iteración para estimar de nuevo las diferencias de parámetros en aquellos ítems sospechosos de DIF. Las iteraciones continúan hasta obtener un set estable de ítems con DIF. Dado que un DIF estadísticamente significativo puede resultar irrelevante si su tamaño de efecto es muy pequeño, también estimamos el tamaño del efecto del DIF mediante el cálculo de la diferencia estandarizada esperada en las puntuaciones del test (ETSSD; Meade, 2010).

Los resultados del análisis del DIF se muestran en la Tabla 3. Se realizan 33 contrastes sobre los ítems que registraron observaciones en todas las categorías en los tres expertos, de los que cuatro resultan en valores χ^2 significativos ($p < .05$) en la segunda iteración. Sin embargo, no se observa ningún ítem que resultase sospechoso de DIF de forma simultánea en los tres contrastes. En la Figura 4 se muestran las curvas características del test para cada evaluador, obtenidas a partir de un modelo parcialmente invariante donde se estiman libremente los parámetros de los ítems sospechosos de DIF. Se puede observar que las curvas están muy cercanas entre sí, sugiriendo que la escala funciona de forma muy similar en los tres evaluadores. La mayor diferencia se observa entre el experto uno y el experto tres, con un valor ETSSD de .091. Dado que el ETSSD se puede interpretar de forma similar a una *d* de Cohen (Cohen, 1988; Meade, 2010), es posible concluir que el tamaño del DIF fue muy bajo.

Baremos de puntuación de la prueba

Para derivar las normas de puntuación de la escala final, primero se estiman las puntuaciones esperadas a posteriori (EAP), y posteriormente se transforman a una escala con media 100 y desviación típica 15 para mayor facilidad de corrección e interpretación. La Tabla 4 muestra las puntuaciones directas, las correspondientes puntuaciones EAP, junto con el error estándar de medida a partir del cual pueden obtenerse los intervalos de confianza. Así, por ejemplo, una puntuación directa de 26 en la escala corresponde a una puntuación theta de .99 y a una puntuación estandarizada de 115. En consecuencia, este programa presenta una calidad razonablemente alta (una desviación típica por encima de la media).

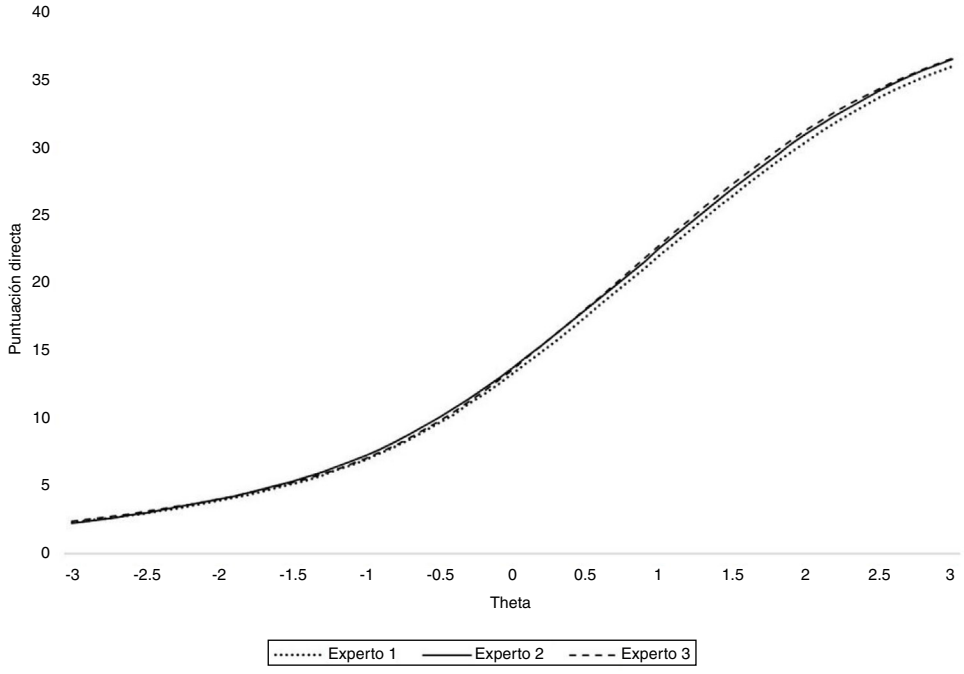

Figura 4. Curvas características del test de cada evaluador.

Tabla 4
Tabla de conversión de puntuaciones directas a puntuaciones theta y puntuaciones estandarizadas

Puntuación directa	EAP[θ|x]	EE[θ|x]	Puntuación estandarizada (*M* = 100, *DT* = 15)
0	-2.549	.620	62
1	-2.226	.569	67
2	-1.961	.530	71
3	-1.733	.497	74
4	-1.528	.469	77
5	-1.339	.442	80
6	-1.165	.418	83
7	-1.002	.395	85
8	-0.851	.375	87
9	-0.710	.357	89
10	-0.579	.342	91
11	-0.457	.330	93
12	-0.342	.321	95
13	-0.233	.313	97
14	-0.129	.308	98
15	-0.028	.304	100
16	0.069	.301	101
17	0.164	.299	102
18	0.258	.297	104
19	0.350	.296	105
20	0.442	.296	107
21	0.533	.296	108
22	0.623	.296	109
23	0.714	.297	111
24	0.805	.299	112
25	0.897	.301	113
26	0.99	.303	115
27	1.084	.306	116
28	1.180	.311	118
29	1.279	.316	119
30	1.381	.322	121
31	1.486	.329	122
32	1.595	.337	124
33	1.708	.347	126
34	1.826	.358	127
35	1.950	.371	129
36	2.081	.385	131
37	2.219	.401	133
38	2.368	.419	136
39	2.529	.438	138
40	2.710	.459	141
41	2.925	.489	144
42	3.194	.534	148

Nota. EAP: puntuaciones estimadas a posteriori, EE: error estándar de medida.

Discusión

El objetivo de este estudio ha sido la construcción y calibración de un instrumento breve para la evaluación de la calidad de programas de educación patrimonial. Para ello, se sigue una secuencia de pasos que comprende la identificación de indicadores de calidad relevantes, la selección mediante la participación de jueces expertos de un set de indicadores claros y coherentes con el contenido a evaluar, y la calibración de un set final de 14 indicadores mediante la aplicación de modelos de la Teoría de la Respuesta al Ítem a las evaluaciones de 330 programas de educación patrimonial realizadas por tres expertos independientes.

De acuerdo a las expectativas basadas en los modelos jerárquicos de calidad (Brady y Cronin, 2001), la escala resulta claramente unidimensional. En lo que respecta a la validez de contenido, los indicadores representan aspectos y facetas de la calidad variados y no redundantes, como se verifica por el hecho de que el instrumento aporta un buen nivel de información a lo largo de una zona de la variable latente suficientemente amplia. El grado de fiabilidad de la escala es elevado, alcanzando su precisión máxima en un rango de calidad comprendido entre zonas bajas (aproximadamente -1 desviaciones típicas) y muy altas (aproximadamente +2 desviaciones típicas). No se aprecia la existencia de un efecto techo relevante. Lo anterior sugiere que la escala es capaz de discriminar con elevada precisión entre varios niveles de la variable (i.e., muy bajo, bajo, medio, alto y muy alto), permitiendo una adecuada clasificación de los programas de acuerdo a su calidad.

Las puntuaciones obtenidas en la Q-Edutage resultan independientes del evaluador, dado que las propiedades de la medida son estables cuando se comparan las respuestas de tres observadores independientes. Esta característica resulta de gran utilidad, en cuanto que –asumiendo una correcta aplicación del instrumento– la no interferencia de características individuales del observador permite la obtención de estimaciones de la calidad insesgados y, en consecuencia, la comparación válida y justa de los programas. En resumen, la Q-Edutage se presenta como un instrumento breve pero preciso y útil para la evaluación de la calidad de los programas de educación patrimonial.

La Q-Edutage puede resultar de utilidad para la investigación evaluativa como responsabilidad prioritaria que se debe asumir desde cualquier ámbito educativo que trabaje con el patrimonio (Popham, 1983), en el marco de una «cultura evaluativa» (Pérez Juste, 2000) respaldada por órganos como el Consejo de Europa o el Plan Nacional de Educación y Patrimonio de España. La escala Q-Edutage se presenta como una herramienta sólida que permitirá (a) su uso interno por parte de las instituciones –gestores y equipos educativos– a quienes facilitará la evaluación de los programas mediante la aplicación de criterios de calidad previamente cotejados; y (b) su uso externo por parte de personas e instituciones –investigadores u organismos públicos– a quienes permitirá extraer información cuantificable y desarrollar estudios globales en relación con la calidad educativa en materia de patrimonio. Así mismo, permitirá confirmar la evaluación interna desde el uso externo, verificar su adecuación y constatar la objetividad e imparcialidad en la recolección y análisis de los datos extraídos del primer nivel. Este nivel superior de evaluación permite una metaevaluación que complementa las insuficiencias o la subjetividad que puede derivar de los miembros integrantes del proyecto.

En conclusión, la Q-Edutage supone un aporte relevante que contribuye al rigor de la planificación educativa como pilar fundamental en el campo de la investigación evaluativa en educación patrimonial, facilita la operativización del nivel de calidad de los proyectos de educación patrimonial y permite extraer información fiable sobre aquellos aspectos susceptibles de mejora.

Este estudio no está exento de limitaciones. Probablemente la más relevante sea la imposibilidad de contrastar las puntuaciones de la Q-Edutage con resultados que presumiblemente deberían estar relacionados con la calidad del programa (p. ej., satisfacción de los participantes, o metas de aprendizaje logradas). Un objetivo relevante de investigación futura es valorar la capacidad predictiva de la calidad del proceso (evaluado mediante la Q-Edutage) respecto a los resultados esperados de la implementación de programas de educación patrimonial.

Pese a sus limitaciones, este estudio es el primero en dedicarse a la construcción de un instrumento objetivo de evaluación de la calidad en programas de educación patrimonial a partir de estándares rigurosos y procedimientos enmarcados en la psicometría moderna. Esperamos que este instrumento contribuya tanto al progreso de la investigación como a la mejora de las prácticas en el ámbito de la educación patrimonial.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Agradecimientos

Este trabajo ha sido realizado en el marco del proyecto EDU2015-65716-C2-1-R, financiado por el Ministerio de Industria, Economía y Competitividad de España y el Fondo Europeo de Desarrollo Regional.

Referencias

Agrusti, F., Poce, A., y Re, M. R. (2017). Mooc design and heritage education. Developing soft and work-based skills in higher education students. *Journal of E-Learning and Knowledge Society*, *13*(3), 97–107, https://doi.org/10.20368/1971-8829/1385.

Baker, F. B. (2001). *The basics of item response theory. In ERIC clearinghouse on assessment and evaluation*. College Park: University of Maryland.

Bangdiwala, K. (1987). Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS User's Group International Conference*, *12*, 1083–1088.

Bolívar, C. R. (2013). *Instrumentos y técnicas de investigación educativa: Un enfoque cuantitativo y cualitativo para la recolección y análisis de datos* (3.rd. Ed.). Houston, Texas: Danaga.

Brady, M. K., y Cronin, J. J., Jr. (2001). Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *Journal of Marketing*, *65*(3), 34–49.

Cai, L., Thissen, D., y du Toit, S. (2011). *IRTPRO user's guide*. Lincolnwood, IL: Scientific Software International.

Cai, L., Thissen, D., y du Toit, S. (2015). *IRTPRO for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International.

Calaf, R., Gillate, I., y Gutiérrez, S. (2015). Transitando por la evaluación de los Programas Educativos de Museos de Arte del proyecto ECPEME. *Educatio Siglo XXI*, *33*(1), 129–150, https://doi.org/10.6018/j/222531.

Calaf, R., San Fabián, J. L., y Gutiérrez, S. (2017). Evaluación de programas educativos en museos: Una nueva perspectiva. *Bordón*, *69*(1), 45–65. http://dx.doi.org/10.13042/Bordon.2016.42686

Cobaleda, M. (2016). The «didactic guide»: A tool for the cultural goods and heritage program. *Opción*, *32*(11), 856–872.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.nd. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Corral, Y. (2009). Validez y confiabilidad de los instrumentos de investigación para la recolección de datos. *Revista Ciencias de la Educación*, *19*(33), 228–247.

Cozzani, G., Pozzi, F., Dagnino, F. M., Katos, A. V., y Katsouli, E. F. (2017). Innovative technologies for intangible cultural heritage education and preservation: The case of i-Treasures. *Personal and Ubiquitous Computing*, *21*(2), 253–265, https://doi.org/10.1007/s00779-016-0991-z.

Deng, L. (2015). Inclusive museum and its impact on learning of special needs children. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4, https://doi.org/10.1002/pra2.2015.1450520100110.

Domínguez, A., y López, R. (2017). Patrimonios en conflicto, competencias cívicas y formación profesional en educación primaria. *Revista de Educación*, *375*, 86–104, https://doi.org/10.4438/1988-592X-RE-2016-375-336.

Ferrando, P. J., y Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory ítem factor analysis. *Educational and Psychological Measurement*, https://doi.org/10.1177/0013164417719308.

Fontal, O. (2003). *La educación patrimonial. Teoría y práctica en el aula el museo e Internet*. Gijón: Trea.

Fontal, O. (2016a). The spanish heritage education observatory/El observatorio de educación patrimonial en España. *Culture and Education*, *28*(1), 254–266, https://doi.org/10.1080/11356405.2015.1110374.

Fontal, O. (2016b). Educación patrimonial: retrospectiva y prospectivas para la próxima década. *Estudios Pedagógicos*, *42*(2), 415–436.

Fontal, O. (2016c). El patrimonio a través de la educación artística en la etapa de primaria. *Arte, Individuo y Sociedad*, *28*(1), 105–120, https://doi.org/10.5209/rev_ARIS.2016.v28.n1.47683.

Fontal, O., y Ibáñez-Etxeberria, A. (2017). La investigación en educación patrimonial Evolución y estado actual a través del análisis de indicadores de alto impacto. *Revista de Educación*, *375*, 184–214.

Fontal, E. E., Ibáñez-Etxeberria, A., Martínez, M., y Rivero, M. P. (2017). El patrimonio como contenido en la etapa de Primaria: del currículum a la formación de maestros. *Revista Electrónica Interuniversitaria de Formación del Profesorado*, *20*(2), 79–94, https://doi.org/10.6018/reifop/20.2.286321.

Fontal, O., y Juanola, R. (2015). La educación patrimonial: Una disciplina útil y rentable en el ámbito de la gestión del patrimonio cultural Cadmo. *International Journal of Educational Research*, *23*(1), 254–266, https://doi.org/10.3280/CAD2015-001002.

Gómez-Redondo, C., Calaf, R., y Fontal, O. (2017). Design of an instrument of analysis for heritage educational resources. *Cadmo. International Journal of Educational Research*, *1*, 63–80, https://doi.org/10.3280/CAD2017-001008.

Gürçayir, S. (2013). Customary modes, modern ways: formal, non-formal education and intangible cultural heritage. *Milli Folklor*, *12*(100), 31–39.

Ibáñez-Etxeberria, A., Fontal, O., y Rivero, P. (en prensa). Educación patrimonial y TIC en España: marco normativo, variables estructurantes y programas referentes. *Arbor, 195.*

Kitungulu, L. (2015). Collaborating to enliven heritage collections. *Museum International*, *65*(1–4), 113–122, https://doi.org/10.1111/muse.12043.

Klein, S., y van Boxtel, M. G. (2011). 'See, think, feel, ask, talk, listen, and wonder'. Distance and proximity in history teaching and heritage education in the Netherlands. *Tijdschrift Voor Geschiedenis*, *124*(3), 381–395.

Lorenzo-Seva, U., y Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91.

Marín-Cepeda, S., García-Ceballos, S., Vicent, N., Gillate, I., y Gómez-Redondo, C. (2017). Educación patrimonial inclusiva en OEPE: un estudio prospectivo. *Revista de Educación*, *375*, 110–135.

Martín-Cáceres, M., y Cuenca, J. M. (2011). La enseñanza y el aprendizaje del patrimonio en los museos: la perspectiva de los gestores. *Revista de Psicodidáctica*, *16*(1), 99–122.

Martín-Cáceres, M., y Cuenca, J. M. (2016). Communicating heritage in museums: outlook, strategies and challenges through a SWOT analysis. *Museum Management and Curatorship*, *31*(3), 1–18, https://doi.org/10.1080/09647775.2016.1173576.

Maydeu-Olivares, A., y Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713, https://doi.org/10.1007/s11336-005-1295-9.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of ítems and scales. Journal of. *Applied Psychology*, *95*(4), 728–743, https://doi.org/10.1037/a0018966.

Orlando, M., y Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous ítem response theory models. *Applied Psychological. Measurement*, *24*(1), 50–64, https://doi.org/10.1177/01466216000241003.

Pérez Juste, R. (2000). La evaluación de programas educativos Conceptos básicos, planteamientos generales y problemática. *Revista de Investigación Educativa*, *18*(2), 261–287.

Popham, W. J. (1983). *Evaluación basada en criterios*. Madrid: Magisterio Español, S.A.

Potočnik, R. (2017). Effective approaches to heritage education: raising awareness through fine art practice. *International Journal of Education Through Art*, *13*(3), 285–294, https://doi.org/10.1386/eta.13.3.285_1.

Rivero, P., Fontal, O., García-Ceballos, S., y Martínez, M. (2018). A model for heritage education through archaeological sites: The case of the roman city of Bilbilis. *Curator, the Museum Journal*, *61*(2), 315–326, https://doi.org/10.1111/cura.12258.

Rodríguez, A., Reise, S. P., y Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150, https://doi.org/10.1037/met0000045.

Samejima, F. (1997). Graded response model. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern ítem response theory* (pp. 85–100). New York, NY: Springer-Verlag.

Simons, H. (2011). *El estudio de caso: Teoría y práctica*. Madrid: Ed. Morata.

Stake, R. (2006). *Evaluación comprensiva. Evaluación basada en estándares*. Barcelona: Grao.

Stake, R. (2010). *Investigación con estudios de casos* (5.ª ed). Madrid: Morata.

Stake, R., y Munson, A. (2008). Qualitative assessment of arts education. *Arts Education Policy Review*, *109*(6), 13–22, https://doi.org/10.3200/AEPR.109.6.13-22.

Tay, L., Meade, A. W., y Cao, M. (2014). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46, https://doi.org/10.1177/1094428114553062.

Timmerman, M. E., y Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous ítems with parallel analysis. *Psychological Methods*, *16*, 209–220, https://doi.org/10.1037/a0023353.

Tsai, S. C. (2011). Multimedia courseware development for world heritage sites and its trial integration into instruction in higher technical education. *Australasian Journal of Educational Technology*, *27*(7), 1171–1189, https://doi.org/10.14742/ajet.911.

Vicent, N., Ibáñez-Etxeberria, A., y Asensio, M. (2015). Evaluation of heritage education technology-based programs. *Virtual Archaeology Review*, *6*(13), 20–27.