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

psicodidáctica

Revista de Psicodidáctica

www.elsevier.es/psicod

Original

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

Olaia Fontal Merillas[a], Silvia García Ceballos[a], Benito Arias[b], and 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, Spain
[b] Departamento de Psicología, Facultad de Educación y Trabajo Social, Universidad de Valladolid, Valladolid, Spain
[c] Departamento de Evaluación, Personalidad y Tratamiento Psicológico, Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain
[d] Instituto Universitario de Integración en la Comunidad (INICO), Facultad de Psicología, Universidad de Salamanca, Salamanca, Spain

ARTICLE INFO

Article history:
Received 29 May 2017
Accepted 1 August 2018
Available online 22 November 2018

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 rigor of evaluation and program planning in the field of heritage education.

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

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.

PII of original article:S1136-1034(18)30155-2.
☆ Please cite this article as: Fontal Merillas O, García Ceballos S, Arias B, Arias VB. Evaluación de la calidad de programas de educación patrimonial: construcción y calibración de la escala Q-Edutage. Rev Psicodidact. 2019;24:31–38. https://doi.org/10.1016/j.psicod.2018.07.003
* Corresponding author.
E-mail address: vbarias@usal.es (V.B. Arias).

Introduction

Heritage education has developed over the last two decades as a cross-disciplinary research discipline with international reach. Research on this discipline involves examining the quantity and quality of doctoral theses, competitive projects and scientific articles (Fontal & Ibáñez-Etxeberria, 2017). Research lines have evolved from the initial approaches, centered on heritage didactics of a mainly descriptive nature, to heritage education (Cobaleda, 2016; Fontal, 2003), with evaluative goals focused on assessing the quality of the design and results of heritage education programs (Fontal, 2016b; Martín-Cáceres & Cuenca, 2016).

The development of this discipline has substantially increased research on heritage education in the last ten years (Fontal & Ibáñez-Etxeberria, 2017) in three major areas: (1) analysis of educational regulations (Fontal, Ibáñez-Etxeberria, Martínez, & Rivero, 2017; Potočnik, 2017); (2) accessibility, diversity and inclusion (Deng, 2015; Marín-Cepeda, García-Ceballos, Vicent, Gillate, & Gómez-Redondo, 2017); and (3) the use of Information and Communication Technologies (ICTs) as a resource and as context for heritage education (Agrusti, Poce, & Re, 2017; Cozzani, Pozzi, Dagnino, Katos, & Katsouli, 2017; Ibáñez-Etxeberria, Fontal, & Rivero, in press).

According to their methodological approach, studies can be classified into three major genealogies (Fontal & Ibáñez-Etxeberria, 2017): (a) *Re-conceptualizing research*, centered on the epistemology of heritage education (Fontal & Juanola, 2015; Klein & Van Boxtel, 2011), involving models, processes or theoretical definitions whose objective is to configure a conceptual corpus (Gürçayir, 2013); (b) *Didactic-contextual research*, focused on heritage teaching in formal (Fontal, 2016c) and non-formal (Calaf, Gillate, & Gutiérrez, 2015) contexts, highlighting the processes of heritage interpretation, communication, and dissemination (Kitungulu, 2015; Martín-Cáceres & Cuenca, 2011); and (c) *Evaluative research*, which assesses educational programs and, less frequently, apprenticeships (Domínguez & López, 2017; Tsai, 2011).

This study is framed within the genealogy of evaluative research because its objective is the construction of a tool to assess the quality of heritage education programs. This work has been developed by the Spanish Heritage Education Observatory (OEPE), within the framework of the "analysis and assessment sequential method for heritage education programs" (*Secuencia de Análisis y Evaluación de Programas de Educación Patrimonial-Observatorio de Educación en España*—SAEPEP-OEPE, Fontal, 2016a), which has been implemented in several studies (Gómez-Redondo, Calaf, & Fontal, 2017; Marín-Cepeda et al., 2017; Rivero, Fontal, García-Ceballos, & Martínez, 2018).

The SAEPEP-OEPE method for the evaluation of heritage education programs

The *SAEPEP-OEPE* (Spanish Heritage Education Observatory-Analysis and Assessment Sequential Method for Heritage Education Programs) is a sequential method of analysis and assessment for heritage education programs (Fontal, 2016a). Its objective is to assess the quality of the programs through a system of sequenced filters that allows for selecting the best cases based on standards defined from (1) normative texts, (2) results from previous assessments, and (3) success indicators extracted from case studies. This method consists of eight phases, each of which is associated with the use of certain evaluation tools (Fontal & Juanola, 2015; for a detailed description of the method, see Fontal, 2016a). Phase 1 consists of searching for and locating programs. Phase 2 incorporates the programs that meet the inclusion criteria into the SHEO database. Phase 3 collects information on the selected programs according to 42 registration fields related to identification, location, project description, educational design, and documentary annex. In phase 4, a descriptive analysis of the data is carried out. Subsequently, there are two phases related to the evaluation of programs: phase 5, related to basic standards, and phase 6, related to specific standards. Finally, phases 7 and 8 consist of a detailed evaluation of the programs that had the best results in previous phases, through single or multiple case studies, accordingly (Simons, 2011; Stake, 2010), and/or through the evaluation of learning (Stake & Munson, 2008). The SHEO-AASMHEP augments the current shortage of structured procedures that evaluate the quality of heritage education designs (Fontal, 2016a).

The present study

Although there are solid evaluation programs (Calaf, San Fabián, & Gutiérrez, 2017; Vicent, Ibáñez-Etxeberria, & Asensio, 2015), there is no instrument for assessing heritage education programs that is brief, objective, independent from the individual characteristics of the evaluator, with robust metric properties, which allows the use of a common standard for unbiased quality comparison between heritage education programs. Having an instrument such as that described would facilitate (a) the accurate and objective evaluation of programs, (b) the rapid screening of the best programs for further in-depth evaluation, and (c) communication between researchers and institutions focused on assessing the quality of heritage education.

The objective of this study is to design and develop a tool to assess the quality of heritage education programs maintaining an adequate balance between precision and brevity, which can be used both on its own (e.g., for screening purposes when the number of programs to be evaluated is high) or to support broader assessing systems (such as the SAEPEP-OEPE described in the previous section). To accomplish this goal, the Q-Edutage scale has been designed in three steps: (1) identification by reviewing the relevant literature regarding the main basic quality indicators for heritage education programs, (2) selection of the indicators with the highest standards of content validity through a study of expert evaluators, and (3) the calibration and construction of the final version of the instrument through procedures framed in Item Response Theory (IRT).

Method

Participants

The calibration sample of the instrument consists of 330 programs randomly selected from the 1719 heritage education programs currently registered in the OEPE database. The sample includes 16 types of program, the most frequent being educational projects (20.6%), didactic designs (14.5%), didactic tools (13%), and research projects (9.1%). Three random sub-samples were extracted and assigned to three expert reviewers, who had previously received a brief training on how to assess the items using the scoring rubric. The evaluators are didactic academics (two professors and a tenured professor) in the fields of Plastic Expression, Social Sciences and Psychology.

Instrument

The first phase of construction of the instrument consisted of reviewing the literature on quality assessment in heritage education (Web of Science and Scopus). A total of 311 articles found through the descriptors "heritage", "education", and "quality" were reviewed. This search was subsequently bounded by the descriptor "standard", obtaining a total of 29 articles, of which only 6 are relevant to the objectives of the study. The first set of indicators

Table 1
Coding of variables according to quality standards

Item	Standard	Coding
i01	Contact information with the management and/or design team, planning and implementation.	Contact
i02	Descriptors that define the program.	Descriptors
i03	Holistic conception of heritage in its nature (material and immaterial) and its qualities (archeological, historical, documentary, artistic...).	Heritage
i04	Specification of the type/typology of the project developed (educational program, educational project, educational design, educational action, isolated activity, etc.).	Typology
i05	Description of the bases, principles, and criteria on which the program is established.	Criteria
i06	Specification of the target audience.	Audience
i07	Incorporation of documentary annexes (memory, images, videos, teaching materials, etc.).	Annexes
i08	Project justification.	Justification
i09	Description of the objectives to be achieved in the development of the program.	Objectives
i10	Presentation of the contents of the program.	Contents
i11	Methodological approach and teaching-learning strategies.	Methodology
i12	Definition of resources, formats, supports, and technology used.	Resources
i13	Determination of evaluation systems or tools.	Evaluation
i14	Evaluation of the impact and repercussion of the proposal.	Impact and Repercussion

were selected according to the model proposed by Stake (2006) by seeking the optimal value of the objectives to be achieved, selecting the rational approach over the intuitive, and setting the specificity of the standards through a control sheet that allows for bias control. Likewise, three content analyses were carried out on different samples taken from the OEPE database (N = 350, 644 and 1120 programs), considering the methodological criteria derived from the National Education and Heritage Plan *(Plan Nacional de Educación y Patrimonio—PNEyP)*. From this phase, a first set of 21 quality indicators was constructed.

In the second phase, 17 expert evaluators assessed the relevance of the 21 indicators. The experts are academics from areas directly or transversally related to heritage education (Didactics of Plastic Expression, Didactics of Social Sciences, Psychology, Didactics and School Organization, Didactics of Body Expression, Didactics of Language and Literature, Music, Sociology, and Graphic Architectural Expression). The experts evaluated the indicators according to their coherence, relevance, and congruence in relation to the evaluation object on a 4-point scale, as well as the clarity in expression, format, and extension (Bolívar, 2013; Corral, 2009). Given the ordinal nature of the measurement scale, the medians of the scores awarded by the evaluators to the items' coherence (Md = 4), relevance (Md = 4), and congruence (Md = 3) were calculated. The evaluators' agreement was assessed by Bangdiwala's Weighted Agreement Coefficient B^W_N (Bangdiwala, 1987), obtaining very satisfactory values (B^W_N = .879 for coherence, B^W_N = .912 for relevance, and B^W_N = .889 for congruence). Six indicators were eliminated from the results of this phase. Four were considered highly redundant in content, and two corresponded to specific quality standards but not to general standards. The content of the 14 indicators selected for the first version of the scale is summarized in Table 1 (the complete format of the items and the scoring rubric can be requested from the first author). To score each indicator, a classification scale of four ordered categories was constructed ("*not achieved*", "*achieved with conditions*", "*achieved*", and "*achieved with quality*").

Data analysis

The data were analyzed with the IRTPRO 4.0 program (Cai, Thissen, & du Toit, 2015) using the *Graded Response Model* (GRM) (Samejima, 1997). The GRM assumes, in addition to the usual assumptions in IRT, that the categories to which the individual responds (or in which the program is qualified, as in this case) can be ordered or hierarchized as summative assessment probabilistic scales or "Likert type" scales. The GRM specifies the probability of a program being qualified with a category i_k as a function of the program level in the latent variable (θ_j), the location parameter of the response category k (β_{ik}), and the item discrimination parameter (α_i).

The purpose of calibration is to ensure that the test is maximally accurate in medium and high areas of the latent variable (program quality). The item retention criteria are as follows: (a) has, at least, a moderate discrimination capacity (i.e., alpha parameter greater than .65, according to the classification of Baker, 2001); (b) the compliance of the item does not result in easy excess (i.e., that the β2 parameter—corresponding to the passing threshold of the response category "achieved with conditions" to the category "achieved"—does not present a theta value substantially lower than -1); (c) does not have problems of conditional independence; (d) adequate fitting of the item to the model (i.e., that its observed and expected frequencies are not significantly different (p < .01); and (e) the estimation of the parameters of the item are sufficiently precise (i.e., with a standard estimation error less than .30, as suggested by Tay, Meade, & Cao, 2014).

Results

Assessing unidimensionality and local independence

Unidimensionality and local independence are two basic requirements in IRT. To ensure sufficient compliance with both assumptions, the following strategies are utilized. (a) An optimized parallel analysis is carried out (Timmerman & Lorenzo-Seva, 2011) based on minimum rank factor analysis (MRFA, implemented in the program FACTOR 10.8.03; Lorenzo-Seva & Ferrando, 2006) and comparing the structure of the polychoric correlation matrix of the 14 items with the results of 500 permuted raw data matrices; (b) Two of the unidimensionality proximity indexes recommended by Ferrando and Lorenzo-Seva (2017) are estimated: the explained common variance (ECV) and the mean of item residual absolute loadings (MIREAL). ECV estimates the size of the dominant factor in relation to the total common variance; values between .70 and .85 are indicators of the unidimensional structure of the data (Rodríguez, Reise, & Haviland, 2016). MIREAL is the mean of absolute loadings of a potential second residual MRFA factor, orthogonal to the primary factor. Consequently, MIREAL is an estimator of the degree to which the structure of the data deviates from unidimensionality, given that the presence of a dominant factor does not necessarily equate to the absence of residual factors with potential substantive relevance. As a general rule, a MIREAL less than .30 suggests the absence of a relevant residual factor (Ferrando & Lorenzo-Seva, 2017); and (c) The standardized values LD χ^2 are

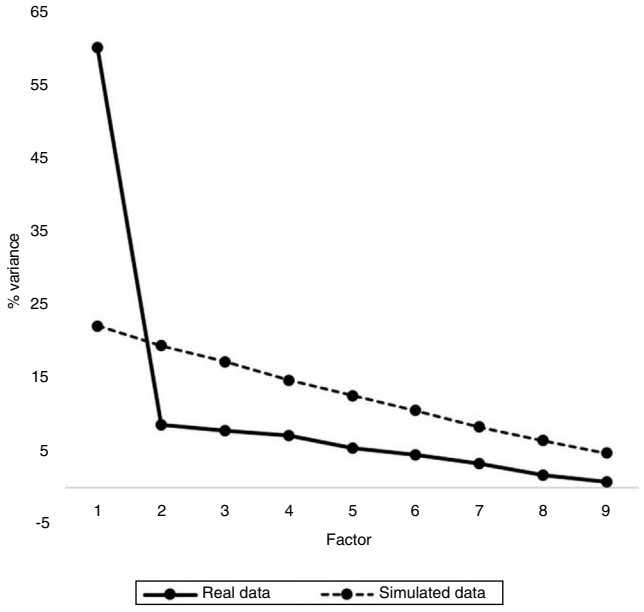

Figure 1. Results from parallel analysis.

inspected for each pair of items. Conditional independence requires that most LD χ^2 values be less than 10 (Cai, Thissen, & du Toit, 2011).

The scale obtained an ECV value of .88, suggesting the presence of a clearly dominant factor. The MIREAL value was .23, suggesting that the presence of relevant systematic variance beyond the main factor is not plausible. The parallel analysis (Figure 1) suggests the retention of a single factor because the variance captured by the first factor is higher than that derived from the simulated matrices (95th centile), and the variance captured by the second factor in the real data is in all cases less than that calculated from the random matrices. The standardized LD χ^2 values are in all cases less than 10, except in the pair of items 5 ("description of the bases, principles and criteria on which the program is established") and 8 ("project justification"), with an LD χ^2 value of 48. The scale demonstrates an adequate level of internal consistency (Cronbach's alpha = .89; ordinal alpha = .91, ordinal theta = .92).

Estimation of model parameters

To estimate the parameters α_i and β_{ik} of the items, a marginal maximum likelihood method is used, the results of which are shown in Table 2.

Table 2
Model parameters

Item	>α	β1	β2	β3
1	1.04 (.14)	−2.58	−.53	1.41
2	1.09 (.15)	−3.08	−1.02	1.00
3	1.14 (.15)	−.82	.88	2.69
4	1.00 (.14)	−2.96	−.97	.93
5	2.37 (.28)	−.44	.44	1.32
6	1.18 (.15)	−1.36	.85	2.03
7	1.35 (.17)	.52	1.45	2.28
8	2.63 (.29)	−.36	.46	1.46
9	2.06 (.24)	−.56	.39	1.89
10	2.57 (.29)	−.47	.40	1.08
11	2.45 (.28)	−.27	.45	1.25
12	2.10 (.25)	−.22	.83	1.48
13	1.13 (.17)	1.30	2.22	3.59
14	1.19 (.16)	.18	1.68	3.41

Note. The standard error of estimation is shown in parentheses.

The items have discrimination parameters between 1.00 (item 4) and 2.63 (item 8), of which three are at a moderate discrimination range (items 1, 2 and 4), three are at a high discrimination range (items 6, 3, and 14), and eight are at a very high discrimination range (13, 7, 5, 9, 12, 8, 11, and 10). Standard errors (M = .20) suggest that the discrimination parameters are estimated with high precision in this sample (Tay et al., 2014). Considering β1 parameters (threshold between the categories "*not achieved*" and "*meets the conditions*"), the items are distributed between very low (β1 = −3.08, item 2) and relatively high (β1 = 1.30, item 13) regions of the latent variable. β2 parameters (threshold between the categories "*meets the conditions*" and "*achieved*") are located between low (β2 = −1.02, item 2) and very high (β2 = 2.22, item 13) regions of the latent variable. The estimation errors of the location parameters are reduced (M = .19 for β1, M = .14 for β2, and M = .20 for β3).

We considered removing one of the items of the pair that presents conditional independence problems (items 5 and 8). However, given that (a) both items present conceptually relevant differences in content and that (b) they provide non-redundant information to the scale, we decided to keep both.

Figure 2 shows the information curve of the test and the distribution of the standard error of measurement. The information curve indicates at what range of the latent variable theta (M = 0, SD = 1) the test is maximally informative. The productive zone for the measurement is approximately between −.1.3 and +2.4 *SD* (points at which the information and error of measurement curves are cut), with a maximum information peak between approximately −.5 and +1.7 *SD*.

Figure 3 shows the characteristic curve of the test. The curve represents the relationship between the expected observed scores and the score in the latent variable. As expected, given the objectives of the instrument, the scale does not adequately discriminate at low levels of the variable because at a range of approximately −3 *SD* to −1.5 *SD*, changes in the latent variable produce practically no change in the observed score. In contrast, from the mean to approximately 2.5 *SD*, the slope becomes noticeably more pronounced.

Fitting the data to the GRM model

We examined the magnitude of the standard errors (lower values indicate higher precision in the estimation of the parameters), the M_2 statistics, and associated RMSEA values (non-significant M_2 values and RMSEA values close to zero suggest good fit of the data to the model; Maydeu-Olivares & Joe, 2006), as well as the significance of the differences between the frequencies of item response observed and expected for each item by S-χ^2 (for a good fit, most items are expected to obtain non-significant S-χ^2 values, Orlando & Thissen, 2000).

M_2 (1505, df = 805) is statistically significant (p < .0001), suggesting the presence of a certain level of misfit. However, the associated RMSEA value (.03) suggests that the misfit is due to the presence of a limited amount of unmodeled error (Cai et al., 2011). The standard estimation errors are small, indicating that the parameters are estimated with high precision for the alpha parameter or very high precision for beta parameters (Tay et al., 2014). Finally, no significant differences were found between the frequencies observed and those expected by the model because no S-χ^2 value was significant (p < .05).

Independence of the scale in relation to the evaluator

An indispensable characteristic in a quality assessment instrument is that, assuming a correct use of the test, it operates uniformly regardless of the evaluator. Consequently, the test scores should be a function of the interaction between the properties evaluated (in this case, the quality of the programs) and the metric properties

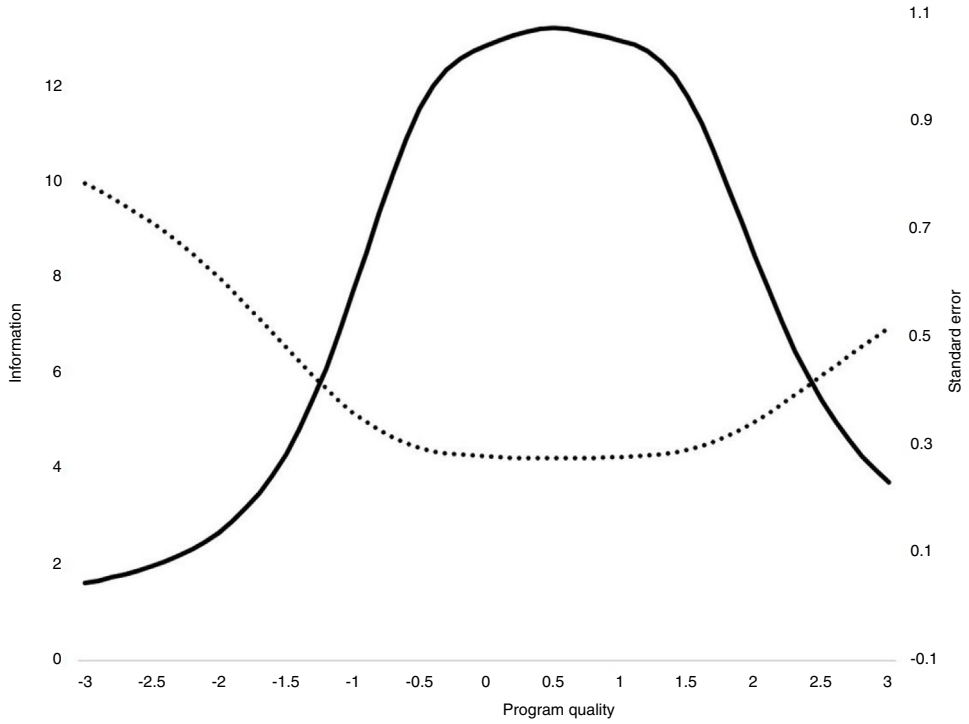

Figure 2. Test information curve.

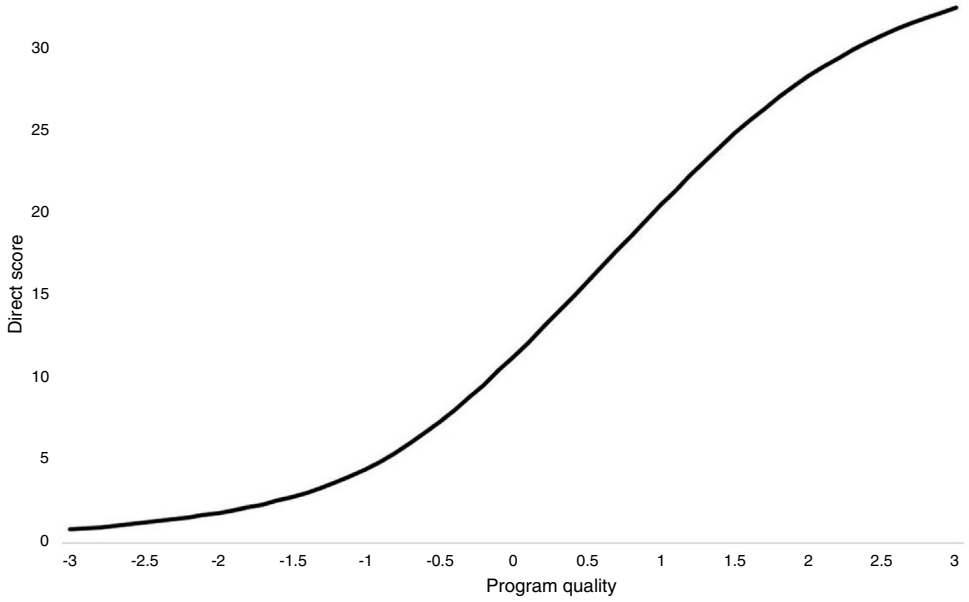

Figure 3. Test characteristic curve.

of the items, depending less on elements outside the construct of interest. To assess the uniformity level to which the instrument operates among evaluators, we estimated the differential item functioning (DIF) among the scores obtained by each of the three experts involved in the data collection. For this, the Wald test is applied. First, we obtain the statistical significance ($p < .05$) of the differences between the parameters estimated from the data obtained by each expert. The items that have been invariant in this phase are used in a second iteration to re-estimate the differences of parameters in suspicious DIF items. The iterations continue until a stable set of DIF items is obtained. Given that a statistically significant DIF may be irrelevant if its effect size is very small, we also estimated the DIF effect size by calculating the expected test scores standardized difference (ETSSD, Meade, 2010).

Table 3 shows the results of the DIF analysis. Thirty-three contrasts are performed on the items that recorded observations in all the categories by the three experts, of which four result in significant χ^2 values ($p < .05$) in the second iteration. However, no suspicious DIF item was observed in the three simultaneous contrasts. Figure 4 shows the characteristic curves of the test for each evaluator, obtained from a partially invariant model where the parameters of the suspicious DIF items are estimated freely. It can be seen that the curves are very close to one another, suggesting that the scale works very similarly for the three evaluators. The

Table 3
Results of the analysis of the differential item functioning

Item	Contrast	First iteration		Second iteration	
		Total c^2	p	Total c^2	p
1	Expert 2 vs. Expert 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		.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	Expert 1 vs Expert 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	Expert 1 vs. Expert 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		.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

Note. Suspicious items in DIF are marked in bold.

greatest difference is observed between expert 1 and expert 3, with an ETSSD value of .091. Since the ETSSD can be interpreted similarly to a Cohen *d* (Cohen, 1988; Meade, 2010), it is possible to conclude that the size of the DIF was very low.

Test scoring scales

To derive the scoring rubric of the final scale, the Expected a posteriori scores (EAP) are estimated first, and the scores are then transformed to a 100-mean scale and standard deviation of 15 for greater ease of correction and interpretation. Table 4 shows the direct scores and the corresponding EAP scores, together with the standard error of measurement from which the confidence intervals can be obtained. In this sense, a direct score of 26 corresponds to a theta score of .99 on the scale and a standardized score of 115. Consequently, this program presents a reasonably high quality (a standard deviation above the mean).

Discussion

The objective of this study is to develop and calibrate a brief instrument to assess the quality of heritage education programs. For this, a sequence of steps is followed that includes the identification of relevant quality indicators; the selection, through the participation of expert evaluators, of a set of clear and coherent indicators with content to be evaluated; and the calibration of a final set of 14 indicators through the application of Item Response Theory models on the evaluations of 330 heritage education programs carried out by three independent experts.

According to expectations based on hierarchical quality models (Brady & Cronin, 2001), the scale is clearly unidimensional. Regarding content validity, the indicators represent varied and non-redundant quality aspects and facets, as verified by the fact that the instrument provides a good level of information throughout a sufficiently broad area of the latent variable. The scale's reliability degree is high, reaching its maximum precision in a quality range between low (approximately -1 standard deviations) and very high (approximately + 2 standard deviations) zones. The existence of a relevant ceiling effect is not observed. This result suggests that the scale is able to discriminate with high precision between several levels of the variable (i.e., very low, low, medium, high, and very high), allowing an adequate classification of the programs according to their quality.

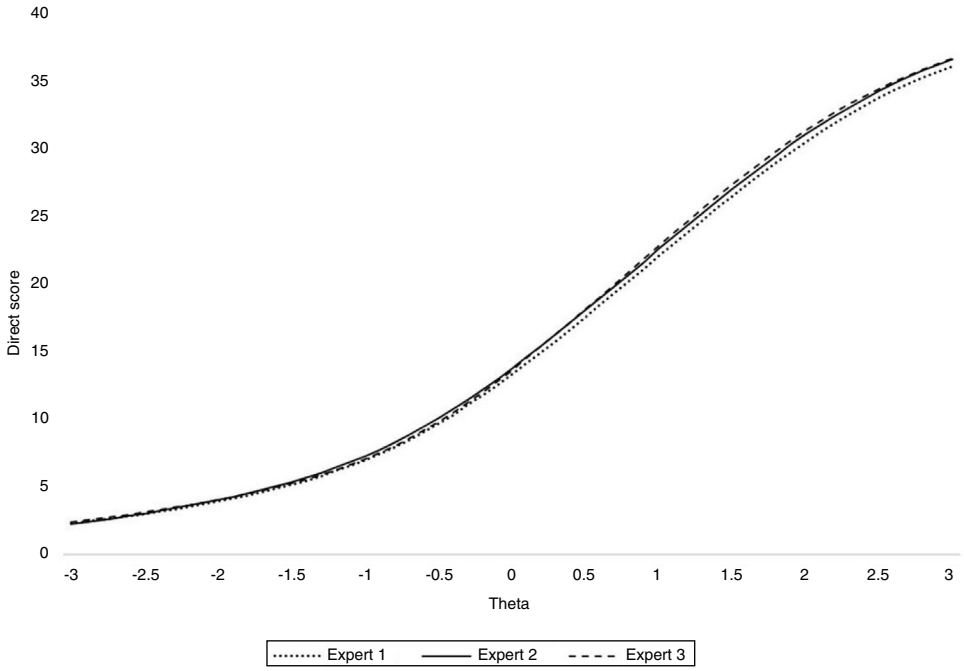

Figure 4. Characteristic test curves of each evaluator.

Table 4
Conversion table from direct scores to theta scores and standardized scores

Direct score	EAP[θ|x]	EE[θ|x]	Standardized score ($M = 100$, $SD = 15$)
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	−.851	.375	87
9	−.710	.357	89
10	−.579	.342	91
11	−.457	.330	93
12	−.342	.321	95
13	−.233	.313	97
14	−.129	.308	98
15	−.028	.304	100
16	.069	.301	101
17	.164	.299	102
18	.258	.297	104
19	.350	.296	105
20	.442	.296	107
21	.533	.296	108
22	.623	.296	109
23	.714	.297	111
24	.805	.299	112
25	.897	.301	113
26	.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

Note. EAP: Expected a posteriori scores, SE: standard error of measurement.

The scores obtained by Q-Edutage are independent from the evaluator because the properties of the scale are stable when the responses of three independent observers are compared. This characteristic is very useful since—assuming a correct application of the instrument—the non-interference of the observer's individual characteristics allows for obtaining unbiased quality estimates and, consequently, the valid and fair comparison of the programs. In summary, Q-Edutage is presented as a brief but precise and useful tool for assessing the quality of heritage education programs.

Q-Edutage can be useful for evaluative research as a priority responsibility that must be assumed by any educational field associated with heritage (Popham, 1983), within the framework of an "evaluative culture" (Pérez Juste, 2000) supported by institutions such as the Council of Europe or the Spanish National Plan for Education and Heritage (Plan Nacional de Educación y Patrimonio—PNEyP). The Q-Edutage scale is presented as a solid tool that will allow (a) its internal use by educational institutions—managers and educational teams—for whom it will facilitate the evaluation of programs through the application of previously collated quality criteria and (b) its external use by people and institutions—researchers or public bodies—for whom it will allow the extraction of quantifiable information and help developing global studies related to heritage educational quality. Likewise, it will allow the confirmation of internal evaluation from external use and verify its adequacy, objectivity and impartiality in the collection and analysis of the data extracted from the first level. This higher level of evaluation allows a meta-evaluation that complements the insufficiencies or subjectivity that arise from the project members.

In conclusion, Q-Edutage is a relevant contribution to the rigor of educational planning as a key aspect in the field of evaluative research in heritage education, facilitates the operationalization of quality level in heritage education projects, and allows the extraction of reliable information on those aspects susceptible to improvement.

One of the most relevant limitations of this study is the impossibility of contrasting Q-Edutage scores with results that presumably should be related to the quality of the program (e.g., satisfaction of the participants, or learning goals achieved). A relevant objective for future research is to assess the predictive capacity of the quality of the process (evaluated through Q-Edutage) in relation to the expected results by the implementation of heritage education programs.

Despite its limitations, this is the first study to focus on the construction of an objective instrument for assessing the quality of heritage education programs based on rigorous standards and procedures framed in modern psychometrics. We hope that this instrument contributes both to the progress of research and to the improvement of heritage education practices.

References

Agrusti, F., Poce, A., & 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. http://dx.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* (3rd. Ed.). Houston, Texas: Danaga.

Brady, M. K., & 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., & du Toit, S. (2011). *IRTPRO user's guide*. Lincolnwood, IL: Scientific Software International.

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

Calaf, R., Gillate, I., & 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. http://dx.doi.org/10.6018/j/222531

Calaf, R., San Fabián, J. L., & 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* (2nd 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., & 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. http://dx.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. http://dx.doi.org/10.1002/pra2.2015.1450520100110

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

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, http://dx.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. http://dx.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. http://dx.doi.org/10.5209/rev_ARIS.2016.v28.n1.47683

Fontal, O., & 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., & 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. http://dx.doi.org/10.6018/reifop/20.2.286321

Fontal, O., & 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. http://dx.doi.org/10.3280/CAD2015-001002

Gómez-Redondo, C., Calaf, R., & Fontal, O. (2017). Design of an instrument of analysis for heritage educational resources, Cadmo. *International Journal of Educational Research*, *1*, 63–80. http://dx.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., & Rivero, P. (in press). 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. http://dx.doi.org/10.1111/muse.12043

Klein, S., & 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., & 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., & Gómez-Redondo, C. (2017). Educación patrimonial inclusiva en OEPE: un estudio prospectivo. *Revista de Educación*, *375*, 110–135.

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

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

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

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous ítem response theory models. *Applied Psychological. Measurement*, *24*(1), 50–64. http://dx.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. http://dx.doi.org/10.1386/eta.13.3.285_1

Rivero, P., Fontal, O., García-Ceballos, S., & 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. http://dx.doi.org/10.1111/cura.12258

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

Samejima, F. (1997). Graded response model. In W. J. van der Linden, & 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* (5th ed.). Madrid: Morata.

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

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

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous ítems with parallel analysis. *Psychological Methods*, *16*, 209–220. http://dx.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. http://dx.doi.org/10.14742/ajet.911

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