Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Audience evaluation of heritage-derived cultural-creative products: a multilevel analysis of rater-level and product-level variance

Zetian Wang
Yunnan Nationalities University
Fangming Zhao
Yunnan Nationalities University
Wentao Tang
Yunnan Nationalities University
Mingfeng Xiao
Yunnan Nationalities University
Lingling Zheng
zheng10091ing@163.com
Yunnan Nationalities University

Article

Keywords:

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11229891/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Studies of heritage-derived cultural-creative products often interpret audience ratings as differences between products, yet each rating also reflects the person who gives it. This study examined how these two sources of variance contribute to evaluations of six designs derived from the Wa jicong top, a spinning top from Yunnan, China. Each of 228 respondents rated two randomly assigned designs, giving 456 evaluations. Crossed random-effects models estimated the rater and product shares of each response, and within-rater and between-rater analyses separated two kinds of association among evaluative judgements. Raters accounted for 81.6% of the variance in design perception, 81.8% in perceived cultural authenticity and 56.1% in aesthetic evaluation. Average differences between products accounted for less than 1% of each judgement; these shares were similar under alternative sample screenings. Cultural familiarity, measured through self-rated knowledge of the source culture and interest in ethnic styling, explained 7.2% to 22.8% of rater-level variance. The ranking of products on authenticity also varied with familiarity. Of the descriptive perceptions, dynamism showed the clearest product difference (product share 4.9%), although its observed ordering did not follow the designs' intended static and dynamic forms. The three evaluative judgements were strongly associated between raters. Within raters, no association between authenticity and aesthetic evaluation was detected, including after correction for measurement error. The indirect association of design perception with aesthetic evaluation through authenticity was 0.31 between raters and 0.00 within them. In this design system, evaluative scores primarily distinguished audience members. Repeated-stimulus studies should report how variance divides between raters and products and distinguish within-rater from between-rater associations before interpreting pooled coefficients.

Introduction

Intangible heritage increasingly reaches the public through designed goods. Cultural-creative products that take their form, colour and meaning from a heritage source now circulate through museum shops, scenic areas and online stores, and a growing body of research asks how audiences receive them (Lin, 2007; Qin et al., 2019; Yu et al., 2022; Li and Li, 2022; Cheng et al., 2026). The prevailing design of these studies is consistent: a sample of respondents rates one or more product stimuli on perception scales, and the findings are reported as differences between products and as associations among constructs such as design perception, perceived cultural authenticity and aesthetic evaluation. Read this way, a rating describes a product, and an association describes how a product's qualities combine in the mind of its audience.

A rating reflects both the person who gives it and the product being rated. Their relative contributions determine how the result can be interpreted. When product variance predominates, scores support comparisons among stimuli. When rater variance predominates, scores chiefly describe differences among audience members; an association across pooled ratings may then show that people who favour one quality also favour another. Studies of facial attractiveness find substantial observer-specific taste alongside shared taste (Hönekopp, 2006; Germine et al., 2015). Agreement is also stronger for natural scenes and faces than for architecture and artworks (Vessel et al., 2018). Heritage-derived products belong to the latter domain of cultural artefacts, and their meanings may be familiar to some audiences more than others. The balance of rater and product variance in their evaluation remains uncertain.

An exploratory audit identified 31 recent empirical studies in heritage, tourism and design in which each participant rated several stimuli (Supplementary Note 1). Twenty aggregated ratings to stimulus means, three analysed individual ratings as independent observations, six accounted for the dependence among a participant's ratings, and two used person-level data or did not state the unit of analysis. One study, of the haptic appeal of visually presented materials, reported participant and stimulus variance components from crossed random-effects models that also contained fixed effects for material properties (Campagna and Chamberlain, 2024). None of the other 30 treated both raters and stimuli as random factors, the specification that estimates how variance divides between them. Methods that separate the two have long been available: crossed random-effects models treat participants and stimuli as samples from their respective populations and estimate the variance attributable to each (Clark, 1973; Baayen et al., 2008; Judd et al., 2012).

These studies also examine relations among constructs. Questionnaire-based regression and structural equation models have linked design, cultural and aesthetic perceptions to consumer outcomes, although stimulus and response structures vary across studies (Qin et al., 2019; Yu et al., 2022; Li and Li, 2022; Cheng et al., 2026). If respondents rate several products, a pooled association combines two relations: whether people who give high ratings on one quality also give high ratings on another, and whether a given person rates the same product higher on both qualities (Enders and Tofighi, 2007; Bell and Jones, 2015). Three studies in the audit treated individual ratings of multiple stimuli as independent observations (e.g. Wu et al., 2023). It remains unclear which relation accounts for reported associations among evaluations of heritage-derived products.

This study applies both decompositions to audience evaluations of cultural-creative products derived from the Wa jicong top (□□□□), a spinning top from Wa communities in south-western Yunnan whose form echoes the termite mushroom after which it is named. Six original concept designs—four pieces of metal jewellery, a bookmark and a plush pendant—translate the top into a shared visual language and take three visual forms: static, dynamic and anthropomorphic. Each of 228 respondents rated two randomly assigned designs, yielding 456 evaluations in which raters and products are partially crossed. The instrument measured two kinds of response to each product. Descriptive perceptions (modernity, refinement, dynamism and distinctiveness) record what a product looks like; evaluative judgements (design perception, perceived cultural authenticity and aesthetic evaluation) record how well it succeeds. The crossed structure allows the variance of each response to be partitioned between raters and products, and the two kinds of response allow the partition to be compared within the same set of stimuli.

Three research questions organise the analysis:

RQ1. How is variance in audience evaluative judgements of heritage-derived cultural-creative products distributed between raters and products, and which rater characteristics account for the rater-level share?

RQ2. Which descriptive perceptions of the same products carry product-level variance that the evaluative judgements do not?

RQ3. Do the associations among the evaluative judgements hold within raters, across the products a rater evaluates, or only between raters?

The results show that evaluative judgements varied mainly between raters. Raters accounted for 81.6% of variance in design perception, 81.8% in perceived cultural authenticity and 56.1% in aesthetic evaluation; average product differences accounted for less than 1% of each. These shares were stable across alternative sample screenings. Cultural familiarity was the measured characteristic most closely associated with rater differences, and the ranking of products on authenticity varied with familiarity. Among descriptive perceptions, dynamism varied more by product than the evaluative judgements did, but its observed ordering did not match the intended static and dynamic forms. The evaluative judgements were closely associated between raters. Within raters, we detected no association of authenticity with aesthetic evaluation; the indirect association of design perception with aesthetic evaluation through authenticity was 0.31 between raters and 0.00 within them. Thus, the pooled evaluations chiefly reflected differences among audience members rather than a shared ranking of these six products.

The literature review sets out the research on heritage-derived products, individual differences in aesthetic judgement and repeated-stimulus analysis. The analytical framework then defines the variance shares and the two components of an association. After describing the stimuli, respondents, measures and models, we report the estimates and consider their implications and limits.

Literature review

Heritage-derived cultural-creative products and their audiences

Cultural-creative products bring heritage into everyday consumption. Design research has described this process as a translation from cultural source to product. Lin (2007) distinguished an object's outer features (colour, texture, form and decoration), its function and use, and its stories, emotions and cultural meaning. The model proposes moving among these levels when adapting a cultural object for contemporary products. Later frameworks developed the process through metaphor (Qin and Ng, 2020) and through study of local context before design begins (Zhang and Wen, 2020). Across these frameworks, the heritage source supplies the content and the product supplies the form in which audiences meet it.

Audience research asks how that encounter is received. Three responses recur. Design perception concerns how well a product's form, colour, ornament and craft are executed. Perceived cultural authenticity concerns whether the product is taken as a genuine expression of its source. Aesthetic evaluation concerns overall appeal. Authenticity is the least settled of the three. Wang (1999) distinguished objective authenticity, located in the original object, from constructive authenticity, which audiences project onto an object according to their own expectations and knowledge. Psychological accounts treat the authenticity of a cultural product as a judgement about whether its perceivable features faithfully represent the cultural knowledge and intentions of those who made it (Kreuzbauer and Keller, 2017). The judgement varies with

product type (Park et al., 2016), and at a cultural event perceived authenticity explains differences in how often visitors return (Castéran and Roederer, 2013).

Two kinds of instrument capture these responses. Semantic-differential scales used in Kansei engineering ask respondents to place a product between adjectives such as traditional–modern or static–dynamic; the resulting scores describe how it appears (Osgood et al., 1957; Nagamachi, 1995). Likert-type items ask whether a product is well designed, authentic or appealing, and therefore record an evaluation. Studies of cultural-creative products have used both formats to examine cultural perception, design aesthetics, emotional response and consumer outcomes (Qin et al., 2019; Yu et al., 2022; Li and Li, 2022; Cheng et al., 2026). Their scores are commonly interpreted as responses to the products that elicited them.

Rater and stimulus variance in aesthetic evaluation

A rating of an object reflects both the object and the person rating it. Research on aesthetic judgement has estimated how much of each. For facial attractiveness, the taste particular to each observer is about as powerful as the taste observers share (Hönekopp, 2006). A twin study of face preferences found that individual differences in taste arise mainly from each person's environment rather than from genes (Germine et al., 2015). The balance also depends on the domain. Across four visual domains, shared taste was strongest for faces, followed by natural landscapes, and was weak for architecture and artworks; observers responded to artefacts of human culture in highly individual ways (Vessel et al., 2018). In ratings of the haptic appeal of visually presented materials, participant variance likewise exceeded material variance in models that included material properties (Campagna and Chamberlain, 2024).

Individual differences in aesthetic judgement are structured rather than random. Models of aesthetic appreciation assign a central role to the knowledge, expertise and familiarity that a viewer brings to an object, which shape how its content is classified and interpreted (Leder et al., 2004). For heritage-derived products, the most relevant form of such knowledge is familiarity with the source culture. Constructive authenticity implies the same dependence from the heritage side: if authenticity is attributed according to what a viewer knows and expects, viewers with different knowledge of the source will judge the same product differently (Wang, 1999).

The literature leaves two questions open for heritage-derived products. First, does evaluation vary more by product or by person? Evidence from other cultural artefacts suggests substantial individual variation, but the studies in our audit did not partition variance in evaluations of these products. Second, do descriptive and evaluative responses distinguish the same products to the same extent? Judging that a product looks dynamic differs from judging that it is well designed, even when both ratings concern the same object. Their respective product shares have not been compared within one stimulus set.

Analysing repeated-stimulus designs

When each participant responds to several stimuli, responses are grouped by participant and, where stimuli are shared, by stimulus as well. Clark (1973) showed that analyses treating stimuli as fixed allow

conclusions about the particular stimuli used but not about the population from which they were drawn. Mixed-effects models with crossed random effects for participants and items address both sources of dependence at once and estimate the variance associated with each (Baayen et al., 2008). Judd et al. (2012) set out this approach for social psychology, and Westfall et al. (2014) showed that the number of stimuli, not only the number of participants, governs the precision of stimulus-level conclusions.

The choice of model also determines what an estimated association means. When a predictor varies both between and within participants, a coefficient from a pooled model is a weighted combination of the within-participant and between-participant associations, and the weight of the between-participant association is the share of the predictor's variation that lies between participants (Enders and Tofighi, 2007; Bell and Jones, 2015). That share grows with the participant-level variance of the predictor and as the number of ratings per participant falls. Explained variance in mixed models can be expressed for the fixed effects alone and for the full model (Nakagawa and Schielzeth, 2013), and can be decomposed by level, so that a predictor's contribution to the participant level is distinguished from its contribution to the observation level (Rights and Sterba, 2019).

These methods are established in psychology and linguistics. To see how they are used in heritage, tourism and design research, we conducted an exploratory audit: Crossref and web searches identified 43 empirical studies published between September 2021 and September 2026, of which 31 had participants rate several stimuli or conditions on scales (Supplementary Note 1). Of these, 20 analysed stimulus means, typically to rank stimuli or to build predictive models from stimulus features. Three analysed individual ratings with single-level procedures. Six accounted for the dependence among a participant's ratings, through crossed random effects (Campagna and Chamberlain, 2024) or paired and repeated-measures comparisons (Sun and Bao, 2025). Two used person-level data or did not state the unit of analysis. The designs themselves are common in the journal literature, including studies in this journal of historical architectural ornament, in-vehicle interfaces and colonial built heritage (Huo et al., 2023; Wu et al., 2023; Kang and Wang, 2025). Only Campagna and Chamberlain (2024) treated both participants and stimuli as random factors; its variance components were estimated conditionally on material-property predictors and for material samples rather than heritage-derived products. None of the audited analyses yields an unconditional partition of evaluative ratings.

Studies of heritage-derived products commonly interpret ratings as product responses. Research on aesthetic judgement shows that evaluations of cultural artefacts can depend strongly on the observer, while crossed models provide a way to estimate both contributions. These methods remain uncommon in the studies we audited. Nor has it been established whether associations among evaluative judgements hold when the same rater compares different heritage-derived products. The following framework defines the quantities used to address these questions.

Analytical framework

Each rating in the present design is given by one rater to one product, and each rater evaluated two of the six products. For any construct, the rating y given by rater r to product p can be written as

$$y_{rp} = \mu + u_r + v_p + e_{rp},$$

where *μ* is the grand mean, *ur* is the rater's deviation from it, *vp* is the product's deviation, and *erp* is the residual. Raters and products are crossed random factors with variances *σ²u* and *σ²v*, and the residual has variance *σ²e*. Each variance divided by their sum gives the share of the total located at that level.

Each variance share answers a different question. The rater share captures a person's tendency to rate the designs highly or poorly across products. The product share captures average differences among designs on which raters agree. The residual includes variation specific to a particular rater–product encounter and measurement error. Because each rater evaluated each assigned product once, those two sources cannot be separated. RQ1 first compares the rater and product shares of the three evaluative judgements.

The second part of RQ1 asks what the rater share consists of. Adding rater characteristics to the model reduces the rater-level variance by the part those characteristics account for. The proportional reduction in rater-level variance expresses this part as a fraction of the rater-level variance in the model without predictors. Three characteristics enter the model: cultural familiarity, a three-item scale covering knowledge of Wa culture, knowledge of the intangible heritage of Yunnan's ethnic minorities and interest in products with ethnic or guochao ('national trend') styling; whether the respondent had visited ethnic-minority areas of Yunnan; and region of residence (Yunnan, other Chinese provinces or overseas). In the analysis, the familiarity scale is also divided into a knowledge component (the first two items) and an interest component (the third).

RQ2 compares the product shares of two sets of responses. Modernity, refinement, dynamism and distinctiveness were measured with bipolar adjective pairs and describe how the products appear. Design perception, perceived cultural authenticity and aesthetic evaluation were measured with agreement statements and express an evaluation. We compare their individual product shares and the mean share for each set. Differences between the sets may reflect item format as well as content.

RQ3 concerns the associations among the evaluative judgements. When each rater evaluates several products, a predictor x can be written as the rater's mean, $\bar{x}$_r, plus the deviation of a particular rating from that mean, x_rp − $\bar{x}$_r. The association of an outcome with $\bar{x}$_r is the between-rater association, β_b: whether raters who rate higher on x also rate higher on the outcome. The association with x_rp − $\bar{x}$_r is the within-rater association, β_w: whether, for the same rater, the product rated higher on x is also rated higher on the outcome. A coefficient estimated by pooling all ratings combines the two. In a bivariate least-squares regression the combination is exact,

β_pooled = β_w(1 − ρ_x) + β_b ρ_x,

where ρ_x is the share of the sum of squares of x that lies between raters (Enders and Tofighi, 2007; Bell and Jones, 2015). With further predictors or product fixed effects the relation holds approximately, and the two components are estimated directly. ρ_x rises with the rater-level share of x and as the number of ratings per rater falls; with two ratings per rater, ρ_x ≈ ICC + (1 − ICC)/2. For design perception in the present data, a rater-level share of 0.816 gives ρ_x = 0.908. The indirect association of design perception with aesthetic evaluation through perceived cultural authenticity is decomposed in the same way, as the product of the

within-rater paths and the product of the between-rater paths. RQ3 is answered by comparing β_w and β_b for each association. Measurement error in a rating varies from one rating to the next and therefore lies mainly within raters, where it attenuates within-rater coefficients. The within-rater associations are therefore also estimated with correction for within-level unreliability and between latent variables in a two-level model.

Methods

Stimuli

The heritage source is the jicong top (□□□□), a spinning top played in Wa communities of south-western Yunnan and named for its resemblance to the jizong (termite) mushroom. Tops are among the characteristic sports implements of Yunnan's ethnic minorities, used by Wa, Yi, Derung and Dai communities alike (Lai and Zhao, 2006). The jicong top offers a compact visual grammar: a pointed crown, a rounded body and a handle beneath.

Six original concept designs were developed from this source (Fig. 1a): a brooch (P1), earrings (P2), a necklace pendant (P3), a ring (P4), a bookmark (P5) and a plush pendant (P6). They span three product categories—metal jewellery (P1–P4), a functional cultural product (P5) and novelty merchandise (P6)—and three visual forms. The static designs (P2, P4) present the top at rest. The dynamic designs (P1, P3, P5) add a golden spiral halo and star points that render the top in motion, and the bookmark adds a Wa buffalo-head motif. The anthropomorphic design (P6) gives the top a simplified face and limbs.

The designs use a common visual language. Each reduces the top's crown–body–handle profile to geometric form, draws on cloisonné enamel, and uses black, red and yellow associated with Wa culture, with gold outlines and accents. This shared treatment allows us to examine audience responses to variation within one design system.

Each product was presented as a design-proposal board combining a rendering of the product, images of the product worn or in use, and a one-sentence caption in Chinese naming the product and its design intent. The captions are translated in Supplementary Note 2.

Participants and procedure

Data were collected through a purpose-built online survey from 12 to 19 May 2026. Recruitment was mainly online, through student communities at Yunnan Minzu University and other universities, urban youth groups, social media and communities interested in cultural-creative products. Other respondents entered through QR codes at heritage sites, scenic areas, exhibitions and an ethnic traditional sports games event. All respondents gave electronic informed consent before starting the survey.

The survey first recorded background characteristics and cultural familiarity, then showed an overview of all six products. Each respondent was randomly assigned two of the six stimuli by the survey system, with assignment probabilities adjusted during data collection to balance coverage across products. Respondents rated each assigned product on all descriptive and evaluative items.

Responses were screened with four criteria specified in the study protocol: respondent under 18 years of age, missing required answers, completion time under 240 seconds, and identical responses across 35 or more scale items. The 240-second threshold lies between the 25th percentile (177 s) and the median (255 s) of raw completion times. Applied in that order, the criteria excluded 15 respondents under 18, none for missing answers, 183 for completion time and 6 for identical responses, leaving 456 evaluations from 228 of the 432 respondents who had submitted 864 evaluations (single-criterion counts and overlaps in Supplementary Table 1). The protocol also included two attention-check items, one direct and one reverse-keyed, recorded as quality indicators alongside the four criteria rather than as grounds for exclusion. Among the 228 respondents, 144 passed the direct item, 132 passed the reverse-keyed item and 127 passed both. The variance decompositions were re-estimated under six alternative screening definitions: both attention items passed, each item passed, completion time of at least 255 s (the median), both items passed with at least 255 s, and a completion-time threshold lowered to 177 s (the 25th percentile) with the other criteria unchanged (Supplementary Table 2).

Valid evaluations per product were 98 (P1), 87 (P2), 59 (P3), 97 (P4), 55 (P5) and 60 (P6). Every respondent rated exactly two products, so raters and products are partially crossed (Fig. 1b). Table 1 describes the sample. Respondents were predominantly aged 18–25 (94.7%) and 76.3% reported education at bachelor's level or higher; 50.4% were women. Most lived in Yunnan (65.8%). Han respondents made up 65.4% of the sample, members of Yunnan's other ethnic minorities 20.6% and Wa respondents 1.8%. In the preceding year, 55.7% had bought between one and three cultural-creative products.

Table 1
Sample characteristics (*N* = 228 respondents)

Characteristic	Category	*n*	%
Age	18–25 years	216	94.7
	26–35 years	3	1.3
	36 years and over	9	3.9
Gender	Male	105	46.1
	Female	115	50.4
	Prefer not to say	8	3.5
Ethnicity	Han	149	65.4
	Wa	4	1.8
	Other ethnic minority of Yunnan	47	20.6
	Other ethnic minority	28	12.3
Education	Bachelor's degree or higher	174	76.3
	Junior college, senior high school or lower	54	23.7
Occupation	Student or company employee	211	92.5
	Other	17	7.5
Region of residence	Yunnan	150	65.8
	Other Chinese provinces	74	32.5
	Overseas	4	1.8
Cultural-creative purchases in preceding year	None	72	31.6
	1–3	127	55.7
	4–10	22	9.6
	11 or more	7	3.1
Monthly income	Lowest band	193	84.6
	Higher bands	35	15.4

Percentages may not sum to 100 because of rounding.

Measures

All items used seven-point scales (Table 2). Items were written for this study and drew on the literatures reviewed above; full wording in Chinese and English is given in Supplementary Note 2.

Table 2
Measurement properties (*n* = 456 evaluations)

Construct	Family	Items	Cronbach's α	Composite reliability	Average variance extracted
Modernity	Descriptive	3	0.758	0.766	0.528
Refinement	Descriptive	3	0.806	0.807	0.582
Dynamism	Descriptive	3	0.736	0.758	0.516
Distinctiveness	Descriptive	3	0.641	0.638	0.376
Design perception	Evaluative	6	0.911	0.912	0.634
Perceived cultural authenticity	Evaluative	3	0.851	0.859	0.672
Aesthetic evaluation	Evaluative	3	0.910	0.910	0.771
Cultural familiarity (respondent level)	—	3	0.767	—	—

Evaluative measurement model	Robust CFI	Robust TLI	Robust RMSEA	SRMR
Three factors	0.966	0.956	0.078	0.035
Two factors (design perception and authenticity merged)	0.949	0.937	0.093	0.039
One factor	0.842	0.807	0.162	0.067

Heterotrait–monotrait ratios: design perception–perceived cultural authenticity 0.912; design perception–aesthetic evaluation 0.699; perceived cultural authenticity–aesthetic evaluation 0.746. Measurement models were estimated with robust maximum likelihood and standard errors clustered by respondent. Three factors versus two: Satorra–Bentler $\Delta\chi^2 = 39.09$, df = 2, $P = 3.25 \times 10^{-9}$. Two factors versus one: $\Delta\chi^2 = 46.02$, df = 1, $P = 1.17 \times 10^{-11}$. All items used seven-point scales.

Descriptive perceptions were measured with twelve semantic-differential pairs (Osgood et al., 1957) selected with reference to Kansei engineering practice (Nagamachi, 1995). Three pairs each measured modernity (traditional–modern, vintage–fashionable, retro–trendy), refinement (rustic–refined, coarse–delicate, ordinary–premium), dynamism (static–dynamic, stable–lively, serious–playful) and distinctiveness (common–unique, superficial–profound, abstract–concrete).

Evaluative judgements were measured with agreement statements. Design perception (six items) followed the outer level of Lin's (2007) model and covered colour, form, ornament, craft, clarity of cultural symbols and overall coherence. Perceived cultural authenticity (three items) drew on the account of authenticity as a judgement of faithful cultural representation (Kreuzbauer and Keller, 2017; Park et al., 2016) and asked

whether the product reflected the essence of Wa culture, conveyed the atmosphere of Yunnan's mountain-forest culture and presented its culture credibly and without exaggeration. Aesthetic evaluation (three items) covered visual appeal, willingness to recommend and overall liking (Li and Li, 2022).

Cultural familiarity was measured at the respondent level with three items on knowledge of Wa culture, knowledge of the intangible heritage of Yunnan's ethnic minorities and interest in products with ethnic or guochao styling. Respondents also reported whether they had visited ethnic-minority areas of Yunnan, their region of residence and their purchases of cultural-creative products in the preceding year. The two knowledge items correlated at $r = 0.65$ (Spearman–Brown reliability 0.79).

Reliability and convergent validity were assessed with Cronbach's α, composite reliability and average variance extracted (Fornell and Larcker, 1981), and discriminant validity with the heterotrait–monotrait ratio (Henseler et al., 2015). For the three evaluative judgements, three measurement models were compared: a three-factor model, a two-factor model merging design perception and perceived cultural authenticity, and a one-factor model. The models were estimated with robust maximum likelihood and standard errors clustered by respondent, and compared with Satorra–Bentler scaled difference tests.

Analysis

Variance in each construct was partitioned with linear mixed models containing crossed random intercepts for respondents and products, estimated by restricted maximum likelihood in lme4 (Bates et al., 2015). Each variance component was expressed as a share of the total. Profile-likelihood intervals were computed for each component, and 1,000 parametric bootstrap replicates provided a second set of intervals. The product component was tested by comparing the full model with a model containing only the respondent intercept; because the null value lies on the boundary of the parameter space, the likelihood-ratio P value was halved (Self and Liang, 1987; Stram and Lee, 1994). The boundary-corrected test corresponds to a $\bar{\chi}^2$ critical value of 2.71, whereas profile-likelihood intervals use the χ^2_1 value of 3.84. As a complementary estimate of product differences, products were also entered as fixed effects and their marginal R^2 computed (Nakagawa and Schielzeth, 2013).

The same decomposition was applied to all seven constructs. Percentile intervals for each product share came from 1,000 parametric bootstrap replicates. The difference between the mean product shares of the descriptive and evaluative families was estimated with 1,000 respondent-cluster bootstrap replicates, each resampling respondents together with both of their ratings. Pairwise product contrasts were Tukey-adjusted comparisons from models with product fixed effects and a respondent random intercept. To assess how far the family difference depended on single constructs, it was recomputed without dynamism and without distinctiveness.

The composition of rater-level variance was examined by adding respondent characteristics to the crossed model in blocks: cultural familiarity, grand-mean centred across respondents (M1); visits to ethnic-minority areas and region of residence (M2); and purchase frequency and gender (M3). The proportional reduction in rater-level variance was computed against a model without predictors fitted to the same respondents. Nested models were compared by likelihood-ratio tests after refitting with maximum likelihood. M3 used the

220 respondents who reported a gender; M1 and M2 used all 228. Age, education and income were not entered because their distributions were concentrated in single categories (Table 1). The knowledge and interest components of familiarity were also entered separately and together, and a familiarity × product interaction was tested in a model with product fixed effects and a respondent random intercept (likelihood-ratio test, 5 df).

To answer RQ3, the associations among design perception, perceived cultural authenticity and aesthetic evaluation were first estimated in the specification common in the literature, by pooling all 456 evaluations in ordinary least-squares regressions with product fixed effects, cultural familiarity and presentation order as covariates, and respondent-clustered standard errors with a small-sample correction (Cameron and Miller, 2015). Each predictor was then split into its respondent mean and the deviation from that mean, and both parts were entered together, giving between-rater and within-rater coefficients in the same specification; their difference was tested directly. The decomposition was repeated in hybrid mixed models with crossed random intercepts for respondents and products, estimated by restricted maximum likelihood. Indirect associations, pooled and separately within and between raters, were estimated with 2,000 respondent-cluster bootstrap replicates (Preacher and Hayes, 2008). Standardised coefficients multiply each coefficient by the standard deviation of the corresponding predictor component and divide it by the standard deviation of the outcome. Within-level and between-level reliability of the three judgements was estimated as Cronbach's α on respondent-centred and respondent-mean item scores and as ω from two-level confirmatory factor analyses (Geldhof et al., 2014). The decomposition was then repeated in two ways. First, the predictors were corrected for unreliability at each level, with reliabilities re-estimated in each of 2,000 respondent-cluster bootstrap replicates; replicates in which a reliability estimate fell outside 0 to 1 or the corrected cross-product matrix was not positive definite were excluded, and the number of valid replicates is reported with each interval. Second, a two-level structural equation model was fitted with latent design perception, authenticity and aesthetic evaluation at both levels, product dummy variables at the within level and cultural familiarity at the between level (lavaan; Rosseel, 2012), with Monte Carlo intervals from 20,000 draws for indirect associations. The observed decomposition was also repeated in three screening subsamples: both attention items passed, completion time of at least 255 s, and a completion-time threshold of 177 s.

The dataset was frozen on 23 May 2026 at 14:35:16 (UTC + 8), and its SHA-256 checksum was recorded. The variance decompositions, measurement-model comparisons, rater-level models, screening sensitivity analyses and within–between decompositions reported here were added after the freeze and run on the unmodified snapshot, whose checksum was verified before and after each run. Analyses used R 4.6.0.

Results

Measurement

Table 2 reports reliability and convergent validity for the seven constructs. Cronbach's α ranged from 0.641 (distinctiveness) to 0.911 (design perception), composite reliability from 0.638 to 0.912, and average

variance extracted from 0.376 to 0.771. The three evaluative judgements reached α values of 0.911 (design perception), 0.851 (perceived cultural authenticity) and 0.910 (aesthetic evaluation).

For the evaluative judgements, the three-factor model fitted the data well (robust CFI = 0.966, TLI = 0.956, RMSEA = 0.078, SRMR = 0.035; *n* = 456 evaluations clustered in 228 respondents). It fitted better than the two-factor model merging design perception and perceived cultural authenticity (robust CFI = 0.949; Satorra–Bentler $\Delta\chi^2 = 39.09$, df = 2, $P = 3.25 \times 10^{-9}$; ΔCFI = 0.017), which in turn fitted better than the one-factor model (robust CFI = 0.842; $\Delta\chi^2 = 46.02$, df = 1, $P = 1.17 \times 10^{-11}$). The heterotrait–monotrait ratio between design perception and perceived cultural authenticity was 0.912, reflecting the close co-variation of audiences' judgements of design quality and cultural authenticity; the corresponding ratios with aesthetic evaluation were 0.699 and 0.746. The three judgements are therefore analysed as distinct constructs.

Rater-level and product-level variance in evaluative judgements

Table 3 reports the variance decomposition of the three evaluative judgements. The rater level accounted for 81.6% of the variance in design perception, 81.8% in perceived cultural authenticity and 56.1% in aesthetic evaluation. The product level accounted for 0.65%, 0.35% and 0.98%, respectively, and the residual for 17.7%, 17.8% and 42.9%.

Table 3
Variance decomposition of evaluative judgements (*n* = 456 evaluations, 228 respondents, 6 products)

Rater share (%)	**Design perception**	**Perceived cultural authenticity**	**Aesthetic evaluation**
	81.6	**81.8**	**56.1**
Product share (%)	0.65	0.35	0.98
Residual share (%)	17.7	17.8	42.9
Product variance, profile-likelihood 95% interval	0 to 0.055	0 to 0.036	0 to 0.110
Product variance, parametric bootstrap 95% interval	0 to 0.035	0 to 0.021	0 to 0.073
Likelihood-ratio test of product component, *P* (boundary-corrected)	0.038	0.127	0.062
Products as fixed effects, marginal R^2 (%)	0.82	0.59	1.42

Linear mixed models with crossed random intercepts for respondents and products, estimated by restricted maximum likelihood. Boundary-corrected *P* values halve the χ^2 (df = 1) *P* value of the comparison with a respondent-only model; uncorrected values were 0.077, 0.255 and 0.125. No model was singular; all 1,000 bootstrap replicates converged for each judgement. Profile-likelihood intervals use the χ^2_1 critical value (3.84); the boundary-corrected test uses the $\bar{\chi}^2$ critical value (2.71). With the boundary critical value, the interval for the design-perception product variance is 0.00009 to 0.041.

The profile-likelihood intervals for the product variance component had a lower bound of zero for all three judgements, with upper bounds of 0.055, 0.036 and 0.110 on the variance scale. Likelihood-ratio tests of the product component, with *P* values halved for the boundary, gave *P* = 0.038 for design perception, *P* = 0.127 for perceived cultural authenticity and *P* = 0.062 for aesthetic evaluation. Entering products as fixed effects instead yielded marginal R^2 values of 0.82%, 0.59% and 1.42%. None of the three models was singular, and all 1,000 parametric bootstrap replicates converged for each judgement.

The rater-level share did not depend on how products were modelled or how variance was estimated (Table 5). With products as fixed effects, the intraclass correlations for raters were 0.823, 0.824 and 0.573. Maximum-likelihood estimation reproduced the restricted-maximum-likelihood shares to within 0.2 percentage points at both the rater and the product level.

Table 5
Robustness of the variance decomposition

Rater share, restricted maximum likelihood (%)	Design perception	Perceived cultural authenticity	Aesthetic evaluation
	81.6	**81.8**	**56.1**
Rater share, maximum likelihood (%)	81.6	81.8	56.0
Product share, restricted maximum likelihood (%)	0.65	0.35	0.98
Product share, maximum likelihood (%)	0.58	0.31	0.81
Rater intraclass correlation, products as fixed effects	0.823	0.824	0.573
Products as fixed effects, marginal R^2 (%)	0.82	0.59	1.42
Rater-level reduction, M2 on 228 respondents (%)	9.4	15.2	24.0
Rater-level reduction, M3 on 220 respondents (%)	9.1	16.0	23.0

Across six alternative screening definitions (118 to 307 respondents), the rater share ranged from 80.4% to 84.8% for design perception, 79.4% to 87.2% for perceived cultural authenticity and 52.1% to 61.2% for aesthetic evaluation, and the product share remained below 1% for all three (Supplementary Table 2).

Composition of rater-level variance

Table 4 reports how much of the rater-level variance respondent characteristics accounted for. Cultural familiarity alone reduced rater-level variance by 7.2% for design perception, 14.5% for perceived cultural authenticity and 22.8% for aesthetic evaluation. Having visited ethnic-minority areas of Yunnan alone reduced it by 4.3%, 4.0% and 3.0%. Region of residence, purchase frequency and gender each accounted for 2% or less, and region and gender produced slight negative reductions. Ethnicity, compared as Han versus

all other groups in a separate model, accounted for none of the rater-level variance (P = 0.72, 0.60 and 0.36 for the ethnicity effect).

Table 4
Proportional reduction in rater-level variance by respondent characteristics (%)

Model	Design perception	Perceived cultural authenticity	Aesthetic evaluation
Cultural familiarity alone (M1)	7.2	14.5	22.8
Visited ethnic-minority areas of Yunnan, alone	4.3	4.0	3.0
Purchase frequency, alone	0.2	0.4	2.0
Region of residence, alone	−0.4	−0.2	−0.4
Gender, alone (n = 220)	−0.5	−0.3	−0.6
Ethnicity (Han vs other), alone	−0.4	−0.4	−0.1
Knowledge + interest (both entered)	8.3	15.7	23.5
Familiarity + visits + region (M2)	9.4	15.2	24.0
M2 + purchase frequency + gender (M3; n = 220)	9.1	16.0	23.0

Reductions are relative to a model without respondent characteristics fitted to the same respondents; negative values indicate a slight increase. Cultural familiarity was grand-mean centred across respondents. Likelihood-ratio tests (maximum likelihood, 220 respondents): adding familiarity, $P = 1.4 \times 10^{-4}$, 2.8×10^{-8} and 4.3×10^{-10}; adding visits and region, P = 0.039, 0.132 and 0.207; adding purchase frequency and gender, P = 0.431, 0.190 and 0.474. Age, education and income were not modelled (Table 1).

The final model (M2), combining cultural familiarity, visits and region, reduced rater-level variance by 9.4%, 15.2% and 24.0%. In likelihood-ratio tests on the 220 respondents who reported a gender, adding cultural familiarity improved fit for all three judgements (design perception $\chi^2 = 14.51$, df = 1, $P = 1.4 \times 10^{-4}$; perceived cultural authenticity $\chi^2 = 30.84$, df = 1, $P = 2.8 \times 10^{-8}$; aesthetic evaluation $\chi^2 = 38.95$, df = 1, $P = 4.3 \times 10^{-10}$). Adding visits and region improved fit for design perception ($\chi^2 = 8.38$, df = 3, P = 0.039) but not for perceived cultural authenticity (P = 0.132) or aesthetic evaluation (P = 0.207). Adding purchase frequency and gender (M3) did not improve fit for any judgement (P = 0.431, 0.190 and 0.474), and the reductions on the 220-respondent sample (9.1%, 16.0% and 23.0%) were close to those of M2 on the full sample.

The composite familiarity coefficients were positive for all three judgements (standardised β = 0.25, 0.35 and 0.36; Supplementary Table 3). For design perception, familiarity acted through interest: entered together with interest, knowledge added nothing (likelihood-ratio P = 0.27). For perceived cultural authenticity and aesthetic evaluation, knowledge added to interest (P = 0.026 and P = 0.0023), and the two components together reduced rater-level variance by 15.7% and 23.5% (Table 4). The familiarity × product interaction was significant for perceived cultural authenticity ($\chi^2 = 17.61$, df = 5, P = 0.0035) and not for design perception (P

= 0.354) or aesthetic evaluation (P = 0.912). For perceived cultural authenticity, the six products' predicted ratings spanned 0.55 points one standard deviation below mean familiarity and 0.36 points one standard deviation above it, while each product's rating rose by 0.50 to 1.22 points (mean 0.79) between these two levels and the rank order changed (Spearman ρ = −0.14; Supplementary Table 3). Among respondents who passed both attention checks, familiarity reduced rater-level variance by 2.0%, 12.1% and 19.7% (Supplementary Table 2).

Descriptive perceptions and evaluative judgements

Figure 2a shows the product-level share for all seven constructs. The four descriptive perceptions had product shares of 4.87% (dynamism), 1.83% (refinement), 1.71% (modernity) and 0% (distinctiveness, a boundary estimate). The three evaluative judgements had shares of 0.98% (aesthetic evaluation), 0.65% (design perception) and 0.35% (perceived cultural authenticity). The profile-likelihood intervals for the product variance component excluded zero for dynamism (lower bound 0.021) and refinement (lower bound 0.003) and included zero for the other five constructs. The rater level remained the largest systematic source for every descriptive construct, with shares from 42.5% (dynamism) to 67.2% (refinement). Among respondents who passed both attention checks, the product share of dynamism was 9.83%; its profile-likelihood interval excluded zero under all seven screening definitions, whereas that of refinement included zero under the four definitions based on the attention items (Supplementary Table 2).

The mean product share was 2.10% across the descriptive family and 0.66% across the evaluative family (Fig. 2b). The difference of 1.45 percentage points (95% respondent-cluster bootstrap interval 0.06 to 3.67) was carried by dynamism: without it the difference was 0.52 percentage points (− 0.85 to 3.24), and without distinctiveness 2.15 (0.47 to 5.17; Supplementary Table 4). With products as fixed effects, the family means were 2.39% and 0.94%. Among respondents who passed both attention checks, the difference was 3.27 percentage points (1.15 to 6.40).

Pairwise comparisons located these differences. Across the three evaluative judgements, 2 of 45 Tukey-adjusted product contrasts reached $P < 0.05$: the plush pendant (P6) was rated higher than the ring (P4) on design perception (difference 0.32 points, $P = 0.018$) and higher than the brooch (P1) on aesthetic evaluation (0.54 points, $P = 0.027$). Across all seven constructs, 11 of 105 contrasts reached $P < 0.05$, five of them for dynamism.

Figure 3 shows the observed product means for dynamism and aesthetic evaluation. On dynamism, the plush pendant (P6) was rated higher than the brooch (P1; 0.75 points, $P = 0.002$), the ring (P4; 0.88 points, $P = 1.0 \times 10^{-4}$) and the bookmark (P5; 0.78 points, $P = 0.008$), and the earrings (P2) were rated higher than the brooch (0.54 points, $P = 0.021$) and the ring (0.67 points, $P = 0.002$). The highest observed means belonged to the anthropomorphic design (P6, 4.81) and to a static design (P2, 4.59); the three dynamic designs (P1, P3, P5) had means between 4.19 and 4.32, and the other static design (P4) the lowest mean (3.99). On aesthetic evaluation, the six observed means lay between 4.76 and 5.31.

Associations within and between raters

Pooling all 456 evaluations in the specification common in the literature, design perception was associated with perceived cultural authenticity (standardised β = 0.761), perceived cultural authenticity with aesthetic evaluation when design perception was held constant (β = 0.342), and design perception with aesthetic evaluation when perceived cultural authenticity was held constant (β = 0.321) and in total (β = 0.581). The standardised indirect association through perceived cultural authenticity was 0.260 (95% bootstrap interval 0.151 to 0.369; Supplementary Table 5). In these evaluations, 90.8% of the variation in each predictor lay between raters; in a bivariate regression this share is the weight that the between-rater association receives in a pooled coefficient, and it approximates that weight in the models reported here.

Table 6 and Fig. 4 separate the two components. Between raters, the associations closely matched the pooled estimates: design perception with perceived cultural authenticity, β = 0.767 (95% CI 0.688 to 0.846); perceived cultural authenticity with aesthetic evaluation, β = 0.391 (0.224 to 0.557); design perception with aesthetic evaluation, β = 0.277 (0.112 to 0.441) directly and 0.591 (0.500 to 0.682) in total. Within raters, comparing the two products that the same person evaluated, the associations were weaker or absent. A product whose design a rater judged more favourably was also judged somewhat more authentic (β = 0.103, 0.022 to 0.184), but no association of authenticity with aesthetic evaluation was detected (β = 0.005, − 0.079 to 0.090). Between-rater and within-rater associations differed for design perception and perceived cultural authenticity (P < 0.001), for perceived cultural authenticity and aesthetic evaluation (P = 0.024) and for the total association of design perception with aesthetic evaluation (P = 0.042), but not for the direct association of design perception with aesthetic evaluation (P = 0.66).

The indirect association followed the same division. Between raters it was 0.314 (95% bootstrap interval 0.178 to 0.442); within raters it was 0.002 (− 0.034 to 0.039). The pooled estimate lay between the two, close to the between-rater value. Hybrid mixed models with crossed random intercepts for raters and products gave the same pattern (Table 6). In these models the within-rater total association of design perception with aesthetic evaluation was small but distinguishable from zero (β = 0.066, 0.006 to 0.127). The pooled coefficients therefore describe how raters who judge the designs favourably also rate them highly on authenticity and aesthetic evaluation. Across the products that a rater evaluated, the indirect association through authenticity was close to zero.

The pattern held with measurement error taken into account (Supplementary Table 6). Reliability was lower within raters than between them for all three judgements (ω = 0.71, 0.56 and 0.84 within; 0.97, 0.94 and 0.97 between). With the predictors corrected for unreliability, the within-rater association of authenticity with aesthetic evaluation stayed at − 0.01 (95% bootstrap interval − 0.29 to 0.16; 1,278 valid replicates of 2,000), against 0.47 (0.23 to 0.76) between raters, and in the two-level latent-variable model it was − 0.08 (95% CI − 0.26 to 0.10). Neither interval excludes a modest positive within-rater association, and both lie below the between-rater estimate. The within-rater association of design perception with authenticity was 0.13 after correction and 0.37 (− 0.12 to 0.85) between latent variables. Between raters, latent design perception and authenticity were nearly indistinguishable (β = 0.93, 0.86 to 1.00), and their separate paths to aesthetic evaluation were not identified at that level. In the three screening subsamples (127 to 307 respondents), the within-rater indirect association stayed between 0.002 and 0.005 and the between-rater indirect association

ranged from 0.31 to 0.38; among respondents who passed both attention checks, design perception was also weakly associated with aesthetic evaluation within raters (β = 0.16, 0.00 to 0.32).

Table 6
Associations within and between raters

Panel	Association	Pooled standardised β [95% CI]	Between-rater standardised β [95% CI]	Within-rater standardised β [95% CI]	Between − within (unstandardised)	P
A: OLS with product fixed effects	Design perception → perceived cultural authenticity	0.761	0.767 [0.688, 0.846]	0.103 [0.022, 0.184]	0.451	< 0.001
A: OLS with product fixed effects	Perceived cultural authenticity → aesthetic evaluation (design perception controlled)	0.342	0.391 [0.224, 0.557]	0.005 [− 0.079, 0.090]	0.464	0.024
A: OLS with product fixed effects	Design perception → aesthetic evaluation (authenticity controlled)	0.321	0.277 [0.112, 0.441]	0.060 [− 0.053, 0.173]	0.106	0.657
A: OLS with product fixed effects	Design perception → aesthetic evaluation (total association)	0.581	0.591 [0.500, 0.682]	0.061 [− 0.060, 0.182]	0.482	0.042
A: OLS with product fixed effects	Indirect: design perception → authenticity → aesthetic evaluation	0.260 [0.151, 0.369]	0.314 [0.178, 0.442]	0.002 [− 0.034, 0.039]	—	—
B: crossed random-intercept hybrid REML	Design perception → perceived cultural authenticity	—	0.768 [0.703, 0.834]	0.109 [0.072, 0.146]	0.434	< 0.001

Note. Pooled coefficients are shown only for Panel A. The expanded between-rater sum-of-squares shares are 0.908 for design perception and 0.908 for authenticity. Component standardisation is coefficient × SD(predictor component) / SD(outcome). Both panels control cultural familiarity and presentation order; Panel A includes product fixed effects and respondent-cluster robust standard errors. Indirect-effect intervals are percentile intervals from 2,000 respondent-cluster bootstrap replicates. In the hybrid bootstrap, 631 of 2,000 fits were singular and retained in the interval calculation. P tests the unstandardised between-minus-within contextual effect; indirect rows omit that contrast and P. Supplementary Table 6 reports the same associations corrected for measurement error, from a two-level latent-variable model and in three screening subsamples.

Panel	Association	Pooled standardised β [95% CI]	Between-rater standardised β [95% CI]	Within-rater standardised β [95% CI]	Between − within (unstandardised)	P
B: crossed random-intercept hybrid REML	Perceived cultural authenticity → aesthetic evaluation (design perception controlled)	—	0.394 [0.257, 0.531]	0.007 [−0.058, 0.072]	0.463	0.003
B: crossed random-intercept hybrid REML	Design perception → aesthetic evaluation (authenticity controlled)	—	0.274 [0.142, 0.405]	0.065 [−0.0003, 0.130]	0.084	0.575
B: crossed random-intercept hybrid REML	Design perception → aesthetic evaluation (total association)	—	0.591 [0.514, 0.668]	0.066 [0.006, 0.127]	0.462	< 0.001
B: crossed random-intercept hybrid REML	Indirect: design perception → authenticity → aesthetic evaluation	—	0.318 [0.180, 0.444]	0.002 [−0.034, 0.039]	—	—

Note. Pooled coefficients are shown only for Panel A. The expanded between-rater sum-of-squares shares are 0.908 for design perception and 0.908 for authenticity. Component standardisation is coefficient × SD(predictor component) / SD(outcome). Both panels control cultural familiarity and presentation order; Panel A includes product fixed effects and respondent-cluster robust standard errors. Indirect-effect intervals are percentile intervals from 2,000 respondent-cluster bootstrap replicates. In the hybrid bootstrap, 631 of 2,000 fits were singular and retained in the interval calculation. P tests the unstandardised between-minus-within contextual effect; indirect rows omit that contrast and P. Supplementary Table 6 reports the same associations corrected for measurement error, from a two-level latent-variable model and in three screening subsamples.

Discussion

Evaluative judgements were organised around the rater

The three evaluative judgements varied mainly by rater. Raters accounted for 81.6% of variance in design perception, 81.8% in perceived cultural authenticity and 56.1% in aesthetic evaluation, whereas average differences among products accounted for less than 1% of each. The pattern persisted under alternative attention-check and completion-time screens. Aesthetic evaluation had a larger residual (42.9%), which combines rater-specific preferences for particular products with measurement error. In this set of six

designs, the measures commonly used to compare heritage products chiefly distinguished the people rating them.

This result is consistent with work in empirical aesthetics. Observers agree more about faces and natural landscapes than about architecture and artworks (Vessel et al., 2018). The present products were closer to the latter pattern. Audiences showed little agreement on which design was better, more culturally authentic or more appealing, but differed substantially in their overall evaluations of the six-design system.

The finding affects two common interpretations of product ratings. Mean-score rankings draw on the small product component: 0.35% to 0.98% of variance for the three evaluative judgements in these data. By contrast, the large rater component was consistent across the two products each person saw and was partly associated with cultural familiarity. Analyses of heritage-product reception therefore need to examine who rates the products favourably as well as which products receive higher average scores.

Cultural familiarity structures the rater level

Cultural familiarity explained more of the rater-level variance than any other measured characteristic: 7.2% in design perception, 14.5% in perceived cultural authenticity and 22.8% in aesthetic evaluation. Visits to ethnic-minority areas of Yunnan added a small amount. Region of residence, purchase frequency and gender added little in these data.

The ordering of these figures is informative. Familiarity explained least of design perception, more of authenticity and most of aesthetic appeal. Authenticity that depends on the viewer's knowledge is what Wang (1999) called constructive authenticity: the same product is judged more or less authentic according to what the viewer brings to it. Models of aesthetic appreciation give prior knowledge the same role in shaping how an object is classified and valued (Leder et al., 2004). The present data locate most of this dependence at the rater level: higher familiarity went with higher evaluations of all six products. For perceived cultural authenticity, which products were judged more authentic also differed with familiarity (familiarity × product, $P = 0.0035$), the product-specific dependence that constructive authenticity implies. The difference in each product's predicted rating between low and high familiarity (0.50 to 1.22 points) exceeded the spread among products at either level of familiarity (0.36 to 0.55 points). Who judged a product was more closely associated with its rating than which product was judged, and which product was judged more authentic depended on who judged it. The knowledge items carried an association with authenticity and aesthetic evaluation beyond interest in ethnic and guochao styling, whereas design perception was associated mainly with interest.

The sample was concentrated in younger and more educated respondents, yet raters still accounted for about 82% of variance in two of the three judgements. Measured background characteristics left 76% to 91% of rater-level variance unexplained. That remaining share may include differences in how readily people value heritage-derived design, as well as other characteristics we did not measure. Cultural familiarity was the clearest measured correlate, but it accounted for only part of the rater effect.

Audiences distinguished the products more in appearance than in evaluation

Dynamism distinguished the products more than any other descriptive measure. Its product-level interval excluded zero under every sample screen, and the product share reached 9.8% among respondents who passed both attention checks. Five of the eleven significant pairwise contrasts concerned dynamism. Refinement also had a nonzero product share in the full sample (1.83%). The 1.45-percentage-point difference between the descriptive and evaluative sets was driven by dynamism. Thus, respondents agreed more about how dynamic the six products looked than about their evaluative standing.

Perceived dynamism did not follow the intended visual categories. The three designs labelled dynamic used a golden spiral halo and star points, yet received middle-range dynamism ratings. The earrings, described as static on their presentation board, ranked second; the other static design, the ring, ranked last. The plush pendant ranked first. Its board noted that twisting the handle could move the figure's arms. Respondents therefore appear to have judged each complete proposal—including category, form, presentation and caption—rather than the intended static or dynamic treatment alone.

Product differences in appearance seldom translated into differences in evaluation. Five of 15 dynamism contrasts reached $P < 0.05$, compared with two of 45 contrasts across the evaluative judgements, whose product shares remained below 1%. Within this design system, products differed more in how they appeared to respondents than in how respondents valued them on average.

For designers, the distinction matters. Changes in form, category and visual treatment altered the character respondents perceived, and descriptive scales registered those changes. Average evaluative standing depended more on the audience than on differences among the six proposals. Comparisons across design systems and heritage sources could test whether this pattern extends beyond the present case. Within a system, descriptive scales remain useful for identifying product differences.

Associations among evaluations describe differences between people

The three evaluative judgements were associated chiefly between people. Raters who scored the designs highly on design perception also tended to score them highly on authenticity and aesthetic evaluation. At the between-rater level, latent design perception and authenticity were nearly indistinguishable; pooled coefficients closely followed the between-rater estimates. Within a rater's two product evaluations, the pattern was different. A design rated higher on design perception was rated somewhat more authentic, but we detected no association between authenticity and aesthetic evaluation in observed-score, reliability-corrected or latent-variable analyses. The within-rater estimates were below the between-rater estimate, although their intervals still allow a modest positive association. The data therefore support a common evaluative tendency between raters more clearly than a within-rater sequence from design through authenticity to aesthetic appeal.

The two levels address different claims. Between raters, the association concerns people who consistently view heritage-derived design favourably; cultural familiarity explains part of that tendency. Within raters, it concerns whether the product judged more faithful to its source is also the product the same person finds more appealing. That within-rater comparison is relevant to the proposed role of authenticity in product

appeal. We found no clear evidence for it in these data. The high heterotrait–monotrait ratio between design perception and authenticity (0.912) likewise reflects their near overlap between raters, although they were distinguishable within raters.

This distinction changes how pooled coefficients should be read. With one product per respondent, associations are necessarily between respondents. With repeated ratings, a pooled coefficient weights the between-respondent component according to the share of predictor variation between respondents; that share was 0.908 for design perception here. Unless the within-respondent component is estimated separately, the result may mainly describe audience differences. The pooled design–authenticity–aesthetic pathway described how raters differed, while the within-rater analysis detected no association between authenticity and aesthetic evaluation for the two products each person saw.

Implications for analysing audience evaluations

Three analytical steps follow from these findings for repeated-stimulus studies. First, report the variance attributable to raters and stimuli. In this study, that partition showed that evaluative scores varied mainly by rater, whereas product differences were most apparent in perceived dynamism.

Second, align analyses of stimulus means with the size of the product component. Twenty of the 31 studies in our audit analysed means by stimulus. This is appropriate when the research question concerns differences among stimuli, but the variance partition indicates their relative size: below 1% for the evaluative judgements here, compared with 4.87% for dynamism.

Third, separate within-rater and between-rater associations when interpreting repeated ratings. Pooled coefficients here resembled the between-rater estimates, while the within-rater indirect association was near zero. Adding a respondent mean alongside the respondent-centred predictor estimates both components (Enders and Tofighi, 2007; Bell and Jones, 2015). Supplementary Note 3 simulates conditions in which pooled and decomposed estimators diverge; the simulation illustrates estimator behaviour and does not validate the empirical coefficients reported here.

These analytical recommendations can be applied beyond the present case. The numerical estimates, however, describe this audience and these six designs.

Conclusions

Ratings of six cultural-creative products derived from the Wa jicong top varied principally across raters. Among 456 evaluations, the rater level accounted for 81.6% of variance in design perception, 81.8% in perceived cultural authenticity and 56.1% in aesthetic evaluation; average differences among products accounted for less than 1% of each. Cultural familiarity was the strongest measured correlate of rater-level differences, and the ranking of products on authenticity varied with familiarity. Perceived dynamism showed a detectable product difference, although its ordering did not match the intended static and dynamic forms. The evaluative judgements were closely associated between raters. Within raters, design perception and

authenticity were weakly associated, and we detected no association between authenticity and aesthetic evaluation.

For this design system, evaluative scores primarily distinguished audience members, and pooled associations among those scores mainly reflected differences between people. Studies of heritage-product reception can examine cultural familiarity alongside product comparisons. When respondents rate multiple stimuli, reporting the rater–stimulus variance partition and separating within-rater from between-rater associations helps establish what each estimate describes.

These findings concern one heritage source and one design system. Respondents viewed proposal images rather than physical products, and the sample was young and highly educated. Each person rated only two products once, so rater–product interaction cannot be separated from measurement error. Within-rater estimates also rely on two ratings per person and on measures with lower within-rater than between-rater reliability. The descriptive and evaluative sets differed in item format as well as content.

Future studies could repeat the partition across heritage sources and design systems, use physical products and broader samples, and ask each rater to evaluate more products.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The survey data were collected anonymously in May 2026. The Ethics Review Committee for Life Sciences and Medical Research Involving Humans, Yunnan Minzu University, retrospectively reviewed and approved the use of these data for research, publication and public data sharing on 23 September 2026 (approval no. 2026LL013). All procedures were performed in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki.

Informed consent

Informed consent was obtained electronically from all participants, who selected "I agree, start the questionnaire" after reading the study information. Participation was voluntary and anonymous, and no directly identifying information was collected.

Funding

This research was funded by the Scientific Research Fund for Postgraduates of the Yunnan Provincial Department of Education (grant no. 2026Y0715) and by the 2025 project of the Bashu Sports Culture Research Base, Sports Culture Development Center of the General Administration of Sport of China, "Bashu

Traditional Sports Culture Empowering the Cultural-Creative Development and Industrialisation of Ethnic Sports Implements" (grant no. 25BSE10).

Author Contribution

Conceptualization: Z.W. and L.Z.; methodology: Z.W. and L.Z.; software: Z.W.; formal analysis: Z.W.; investigation: Z.W., F.Z., W.T. and M.X.; data curation: Z.W.; writing—original draft: Z.W.; writing—review and editing: Z.W., F.Z., W.T., M.X. and L.Z.; visualization: Z.W.; supervision: L.Z.; project administration: L.Z.; funding acquisition: Z.W. All authors read and approved the final manuscript.

Data Availability

The de-identified evaluation data, screening log, codebook, analysis scripts, software environment (renv lock file), simulation reproducibility package and full questionnaire (Chinese original with English translation) are openly available in Figshare at https://doi.org/10.6084/m9.figshare.32531145 (direct link: https://figshare.com/articles/dataset/_/32531145) under a CC BY 4.0 licence. The public dataset replaces response identifiers with sequential anonymous codes, omits exact timestamps and open-ended answers, and withholds the answers of respondents under 18, who were excluded from all analyses.

References

1. Baayen RH, Davidson DJ, Bates DM (2008) Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang 59:390–412. https://doi.org/10.1016/j.jml.2007.12.005
2. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67:1–48. https://doi.org/10.18637/jss.v067.i01
3. Bell A, Jones K (2015) Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. Political Sci Res Methods 3:133–153. https://doi.org/10.1017/psrm.2014.7
4. Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. J Hum Resour 50:317–372. https://doi.org/10.3368/jhr.50.2.317
5. Campagna M, Chamberlain R (2024) How material sensory properties and individual differences influence the haptic aesthetic appeal of visually presented stimuli. Sci Rep 14:13690. https://doi.org/10.1038/s41598-024-63925-9
6. Castéran H, Roederer C (2013) Does authenticity really affect behavior? The case of the Strasbourg Christmas Market. Tour Manag 36:153–163. https://doi.org/10.1016/j.tourman.2012.11.012
7. Cheng H, Qiu X, Liu B et al (2026) What makes a museum product creative? Unpacking aesthetic, emotional, and cultural dimensions from the consumer perspective. Think Skills Creativity 59:101952. https://doi.org/10.1016/j.tsc.2025.101952
8. Clark HH (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. J Verbal Learn Verbal Behav 12:335–359. https://doi.org/10.1016/s0022-5371(73)80014-3

9. Enders CK, Tofighi D (2007) Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. Psychol Methods 12:121–138. https://doi.org/10.1037/1082-989x.12.2.121
10. Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. J Mark Res 18:39–50. https://doi.org/10.1177/002224378101800104
11. Geldhof GJ, Preacher KJ, Zyphur MJ (2014) Reliability estimation in a multilevel confirmatory factor analysis framework. Psychol Methods 19:72–91. https://doi.org/10.1037/a0032138
12. Germine L, Russell R, Bronstad PM et al (2015) Individual aesthetic preferences for faces are shaped mostly by environments, not genes. Curr Biol 25:2684–2689. https://doi.org/10.1016/j.cub.2015.08.048
13. Henseler J, Ringle CM, Sarstedt M (2015) A new criterion for assessing discriminant validity in variance-based structural equation modeling. J Acad Mark Sci 43:115–135. https://doi.org/10.1007/s11747-014-0403-8
14. Hönekopp J (2006) Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. J Exp Psychol Hum Percept Perform 32:199–209. https://doi.org/10.1037/0096-1523.32.2.199
15. Huo F, Zhao Y, Chai C, Fang F (2023) A user experience map design method based on emotional quantification of in-vehicle HMI. Humanit Social Sci Commun 10:264. https://doi.org/10.1057/s41599-023-01761-4
16. Judd CM, Westfall J, Kenny DA (2012) Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. J Personal Soc Psychol 103:54–69. https://doi.org/10.1037/a0028347
17. Kang X, Wang J (2025) Design optimization of wood-carved window grilles in historical architectures using stable diffusion model and intuitionistic fuzzy VIKOR. Humanit Social Sci Commun 12:972. https://doi.org/10.1057/s41599-025-05388-5
18. Kreuzbauer R, Keller J (2017) The authenticity of cultural products: A psychological perspective. Curr Dir Psychol Sci 26:417–421. https://doi.org/10.1177/0963721417702104
19. Lai Y, Zhao J (2006) A study of the modelling culture of ethnic sports implements in Yunnan. J Yunnan Minzu Univ (Philosophy Social Sci Ed) (5):128–131 (in Chinese)
20. Leder H, Belke B, Oeberst A, Augustin D (2004) A model of aesthetic appreciation and aesthetic judgments. Br J Psychol 95:489–508. https://doi.org/10.1348/0007126042369811
21. Li Y, Li J (2022) The influence of design aesthetics on consumers' purchase intention toward cultural and creative products: Evidence from the Palace Museum in China. Front Psychol 13:939403. https://doi.org/10.3389/fpsyg.2022.939403
22. Lin R (2007) Transforming Taiwan Aboriginal cultural features into modern product design: A case study of a cross-cultural product design model. Int J Des 1:45–53
23. Nagamachi M (1995) Kansei engineering: A new ergonomic consumer-oriented technology for product development. Int J Ind Ergon 15:3–11. https://doi.org/10.1016/0169-8141(94)00052-5
24. Nakagawa S, Schielzeth H (2013) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods Ecol Evol 4:133–142. https://doi.org/10.1111/j.2041-

210x.2012.00261.x

25. Osgood CE, Suci GJ, Tannenbaum PH (1957) The measurement of meaning. University of Illinois Press, Urbana, IL, USA
26. Park J, Javalgi R, Wachter M (2016) Product ethnicity and perceived consumer authenticity: The moderating role of product type. J Consumer Mark 33:458–468. https://doi.org/10.1108/JCM-01-2015-1272
27. Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 40:879–891. https://doi.org/10.3758/BRM.40.3.879
28. Qin Z, Ng S (2020) Culture as inspiration: A metaphorical framework for designing products with traditional cultural properties (TCPs). Sustainability 12:7171. https://doi.org/10.3390/su12177171
29. Qin Z, Song Y, Tian Y (2019) The impact of product design with traditional cultural properties (TCPs) on consumer behavior through cultural perceptions: Evidence from the young Chinese generation. Sustainability 11:426. https://doi.org/10.3390/su11020426
30. Rights JD, Sterba SK (2019) Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. Psychol Methods 24:309–338. https://doi.org/10.1037/met0000184
31. Rosseel Y (2012) lavaan: An R package for structural equation modeling. J Stat Softw 48:1–36. https://doi.org/10.18637/jss.v048.i02
32. Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610. https://doi.org/10.1080/01621459.1987.10478472
33. Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177. https://doi.org/10.2307/2533455
34. Sun M, Bao Y (2025) Effects of reflecting water on viewers' visual attention, spatial perception, and emotional responses: A case study in Chinese heritage garden. npj Herit Sci 13:26. https://doi.org/10.1038/s40494-025-01650-w
35. Vessel EA, Maurer N, Denker AH, Starr GG (2018) Stronger shared taste for natural aesthetic domains than for artifacts of human culture. Cognition 179:121–131. https://doi.org/10.1016/j.cognition.2018.06.009
36. Wang N (1999) Rethinking authenticity in tourism experience. Annals Tourism Res 26:349–370. https://doi.org/10.1016/s0160-7383(98)00103-0
37. Westfall J, Kenny DA, Judd CM (2014) Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. J Exp Psychol Gen 143:2020–2045. https://doi.org/10.1037/xge0000014
38. Wu Y, Li N, Xia L et al (2023) Visual attention predictive model of built colonial heritage based on visual behaviour and subjective evaluation. Humanit Social Sci Commun 10:869. https://doi.org/10.1057/s41599-023-02399-y

39. Yu L, Feng X, Wang J et al (2022) Research on the mechanism of emotional design in Chinese cultural and creative products. Herit Sci 10:119. https://doi.org/10.1186/s40494-022-00751-0
40. Zhang X, Wen K-H (2020) A model process of integrating context of local culture for pre-development stage in the design of cultural and creative products: Using Macao's historical buildings as an example. Sustainability 12:6263. https://doi.org/10.3390/su12156263

Figures

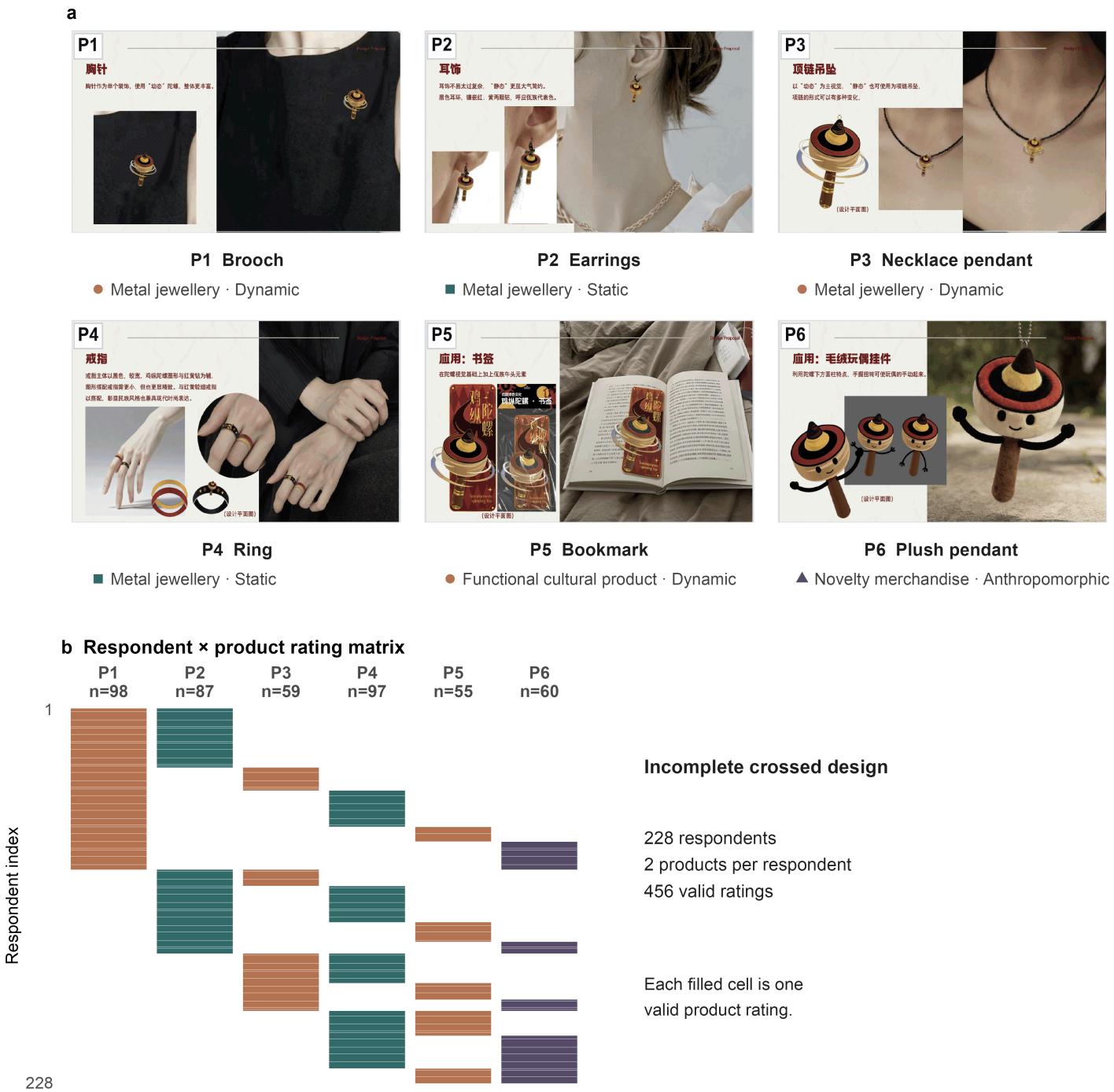


Figure 1

Stimuli and rating design. a The six concept designs derived from the Wa jicong top, as presented to respondents on design-proposal boards, with product category and visual form. Board captions are translated in Supplementary Note 2. b Rating matrix of the 228 respondents (rows) by six products (columns); each filled cell is one valid evaluation. Each respondent rated two products, giving 456 evaluations (P1 = 98, P2 = 87, P3 = 59, P4 = 97, P5 = 55, P6 = 60).

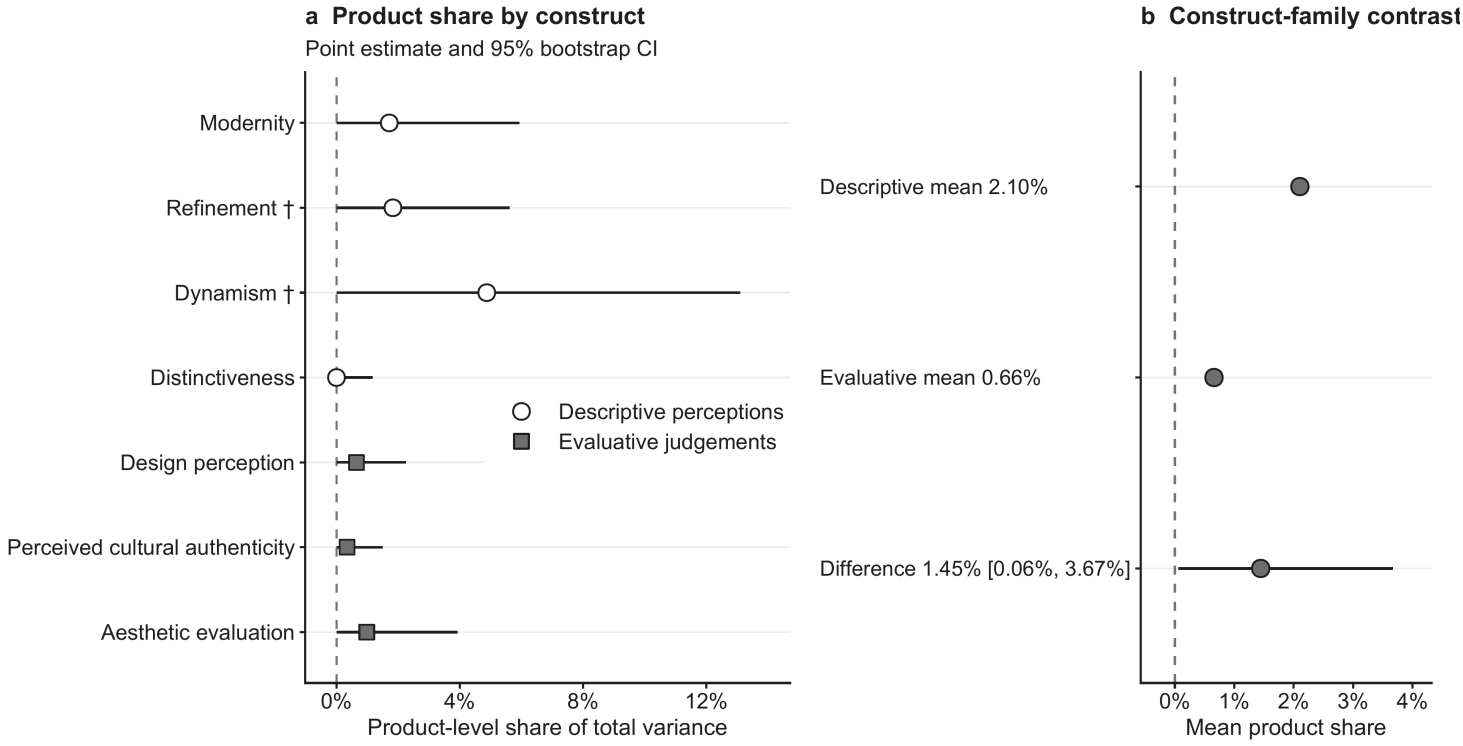


Figure 2

Product-level share of variance for descriptive perceptions and evaluative judgements. a Share of total variance located at the product level for each of the seven constructs, from linear mixed models with crossed random intercepts for respondents and products (restricted maximum likelihood). Points are estimates; bars are 95% percentile intervals from 1,000 parametric bootstrap replicates. **† The profile-likelihood interval for the product variance component excludes zero. The bars and the † marker use different criteria: bars show bootstrap intervals for the share of variance, and † marks components whose profile-likelihood interval for the variance excludes zero. b** Mean product share for the descriptive family (modernity, refinement, dynamism, distinctiveness) and the evaluative family (design perception, perceived cultural authenticity, aesthetic evaluation), and their difference with a 95% interval from 1,000 respondent-cluster bootstrap replicates. *n* = 456 evaluations from 228 respondents.

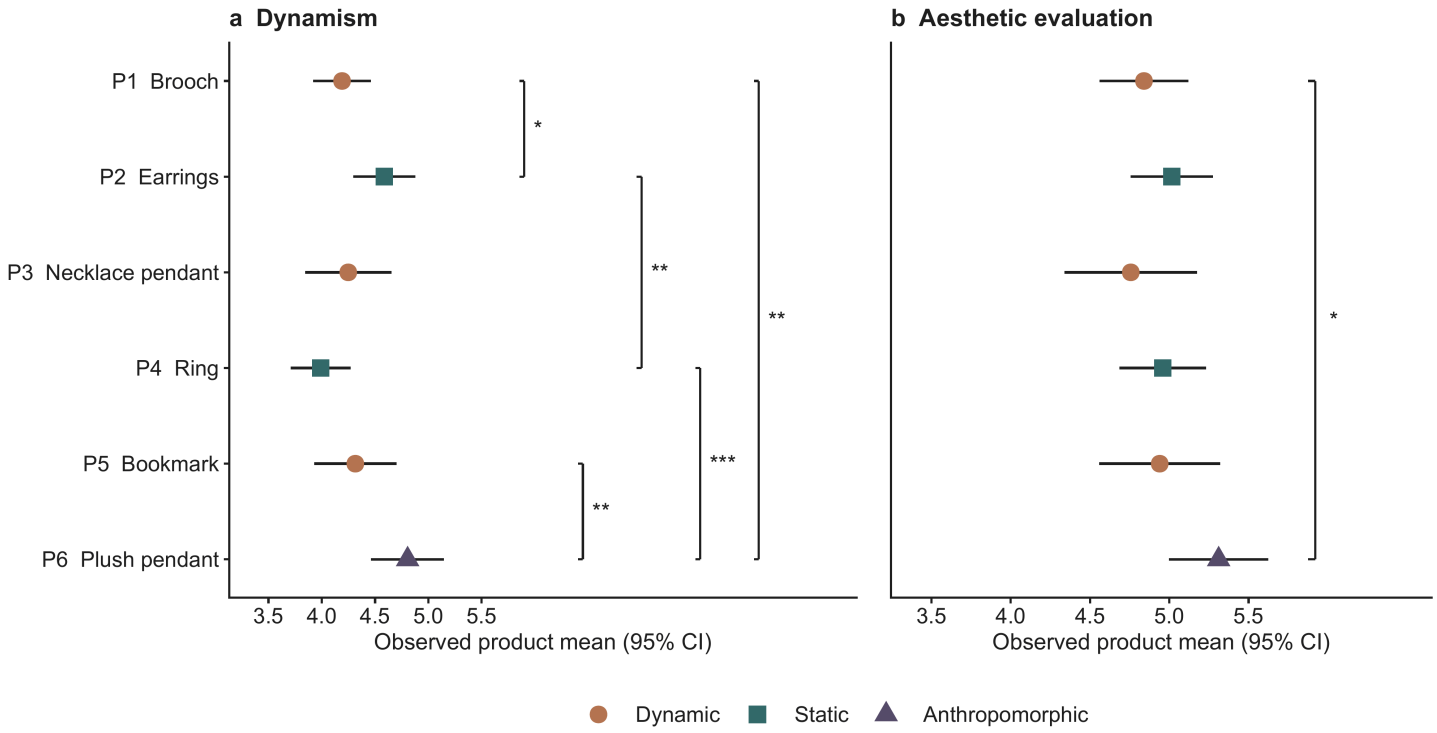


Figure 3

Observed product means for dynamism and aesthetic evaluation. a Dynamism. **b** Aesthetic evaluation. Points are observed means; bars are 95% confidence intervals (mean ± 1.96 × s.d./√*n*). Colour and shape indicate visual form: dynamic (P1, P3, P5), static (P2, P4) and anthropomorphic (P6). Brackets mark Tukey-adjusted pairwise contrasts with $P < 0.05$ from linear mixed models with product fixed effects and a respondent random intercept: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Evaluations per product: P1 = 98, P2 = 87, P3 = 59, P4 = 97, P5 = 55, P6 = 60.

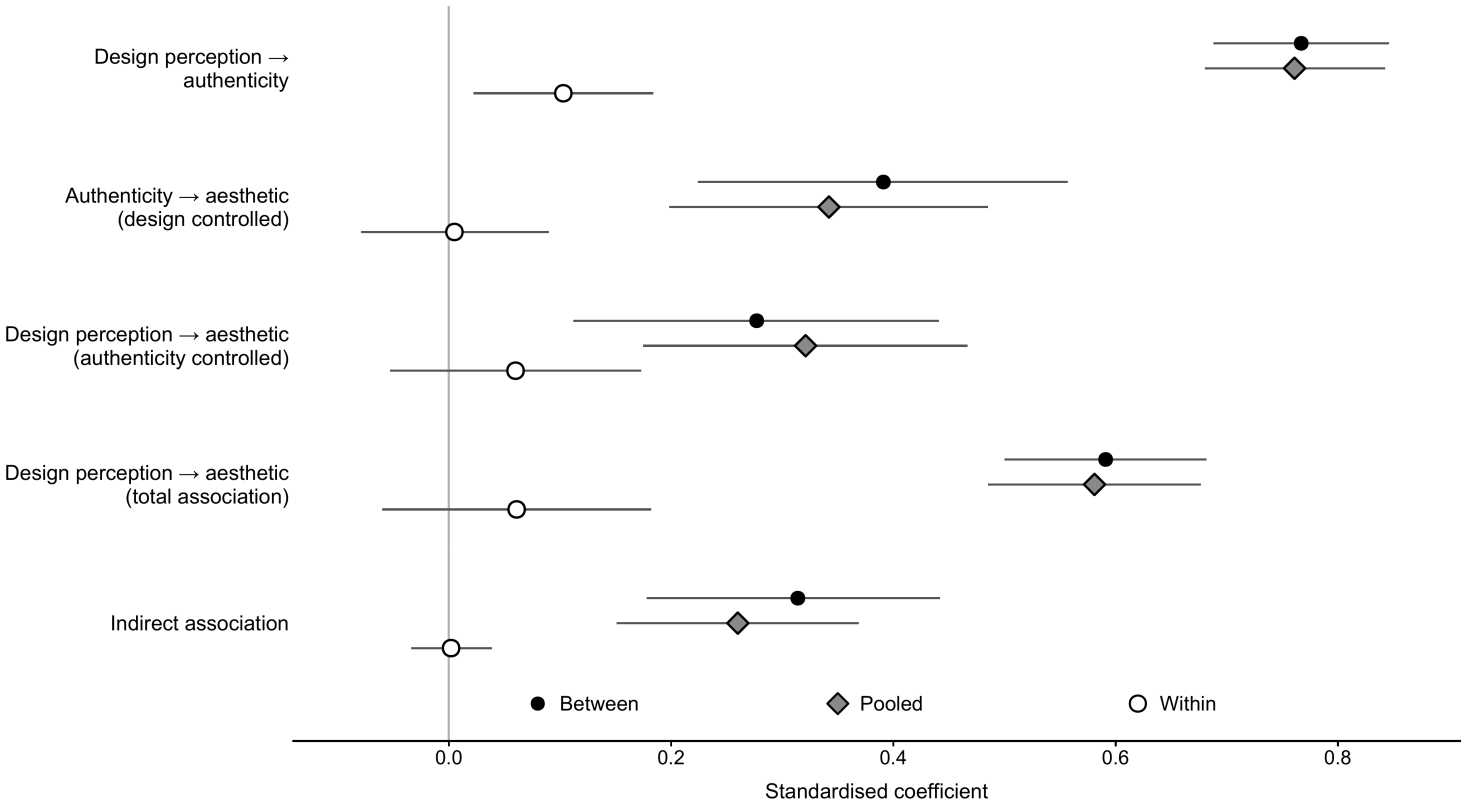


Figure 4

Within-rater and between-rater associations among evaluative judgements. Standardised coefficients from least-squares models with product fixed effects and respondent-clustered standard errors (Table 6, Panel A). Filled circles show between-rater associations, grey diamonds pooled associations from the same models without the within–between split, and open circles within-rater associations. Bars are 95% confidence intervals; for the indirect association they are 95% percentile intervals from 2,000 respondent-cluster bootstrap replicates. **Each coefficient is standardised by the standard deviation of its own predictor component, so positions compare associations within each level; the tests of between–within differences in Table 6 use unstandardised coefficients. Models control for cultural familiarity and presentation order.** *n* = 456 evaluations from 228 respondents.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf