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.

# Consumer Preferences for Museum Cultural and Creative Products: Best–Worst Scaling, Preference Heterogeneity, and Portfolio Decision Support

**Bin Zhang**
Shaanxi University of Science and Technology
**Ping Ji**
jiping@sust.edu.cn

Shaanxi University of Science and Technology
**Qi Jia**
Shaanxi University of Science and Technology
**Yefan Guo**
Shaanxi University of Science and Technology

**Article**

**Keywords:**

**Posted Date:** November 23rd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10886250/v1

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

**Additional Declarations:** No competing interests reported.

Consumer Preferences for Museum Cultural and Creative Products: Best–Worst Scaling, Preference Heterogeneity, and Portfolio Decision Support

Bin Zhang, Ping Ji *, Qi Jia, and Yefan Guo

College of Art and Design, Shaanxi University of Science and Technology, Xi'an, 710021, China

* Correspondence: jiping@sust.edu.cn

**Abstract**

Museum cultural and creative product development must reconcile heritage translation with consumer appeal, yet choice-based evidence on how consumers differentiate concrete design alternatives remains limited. This study developed 36 product concepts from 15 Western Zhou bronze prototypes and collected Best–Worst Scaling (BWS) data from 404 valid respondents. Conditional-logit MaxDiff models, design-balanced utilities, consumer-heterogeneity analyses, prototype-controlled tests, independent label validation, and constrained portfolio optimization were used to examine product preferences and decision robustness. Preferences differed markedly across products and remained stable under stricter data-quality filtering. Significant heterogeneity was associated with purchase experience, collection orientation, practical orientation, interaction orientation, and professional background, whereas cultural familiarity showed no independent effect after joint controls. After controlling for cultural prototypes, broad Product Concept and Translation Pathway categories did not significantly explain utility, and cultural-resource potential was not consistently associated with consumer preference. Label validation preserved the main behavioral inferences, while portfolio objective values changed by less than 0.4% despite changes in portfolio membership. These findings support product-level choice evidence and preference heterogeneity as key inputs for culturally informed museum product development and transparent portfolio decision support.

**Keywords:** museum cultural and creative products; consumer preferences; Best–Worst Scaling; preference heterogeneity; cultural heritage; portfolio decision support

## 1. Introduction

Museum cultural and creative products transform collection-based cultural resources into contemporary product forms that serve both cultural-communication and market-oriented functions [1,2]. Unlike conventional consumer products, their evaluation involves not only functionality, aesthetics, and user experience, but also cultural symbols, historical information, and the ways in which heritage elements are reinterpreted through design [3,4]. Consumers may therefore consider cultural meaning, creative expression, visual design, and practical functionality simultaneously when evaluating such products [5,6]. However, the historical or artistic value of a cultural resource does not necessarily determine consumer responses to the products developed from it. A central challenge is

therefore to understand how consumers differentiate among specific product alternatives derived from different cultural prototypes and design translations.

Existing research on museum cultural and creative consumption has examined perceived value, cultural meaning, creativity, functionality, aesthetics, authenticity-related experiences, and individual consumer differences [7]. Much of this evidence, however, has been generated through questionnaire ratings, purchase-intention measures, perceived-value models, user-requirement evaluations, or consumer reviews [8,9]. These approaches identify important attributes and predictors of product evaluation, but provide less direct evidence about how consumers trade off concrete alternatives that differ simultaneously in cultural prototypes, product forms, and design expressions. Choice-based evidence is therefore needed to establish the relative preference structure among specific museum product concepts.

This issue also concerns preference heterogeneity and the level at which such heterogeneity is analyzed. Choice-based methods such as Best–Worst Scaling (BWS) require consumers to make explicit trade-offs among concrete alternatives and therefore allow relative preference structures to be estimated across products [10]. However, an aggregate preference structure does not imply that all consumers evaluate the available alternatives in the same way [11]. Previous research has suggested that demographic characteristics, prior consumption experience, cultural experience, and product-related orientations may be associated with differences in evaluations of cultural and creative products [12]. Yet differences in overall evaluations do not necessarily imply that consumer characteristics alter the relative utilities of specific products in the same way. Moreover, product-level preference heterogeneity may not be captured adequately by broader design strategies. It is therefore important to distinguish heterogeneity across individual products from heterogeneity that can be summarized at the strategy level, thereby avoiding the premature interpretation of product-specific differences as general preferences for broad design categories.

Moving from preference identification to product development introduces a further decision problem. Structured multi-criteria approaches have increasingly been used in museum cultural and creative design to integrate user requirements, expert judgments, design attributes, and alternative evaluations [13]. However, these approaches have primarily focused on requirement weighting, attribute evaluation, or the ranking of individual alternatives [14]. Less attention has been given to how choice-derived consumer utility can be incorporated into portfolio-level selection together with cultural-resource translation requirements, risks of cultural misinterpretation, and design-strategy considerations. Selecting the most preferred individual product is not equivalent to constructing a portfolio that must also consider prototype duplication, design-strategy coverage, translation-pathway representation, and overall balance. Extending product-level consumer-choice evidence to constrained portfolio selection can therefore shift the decision problem from identifying preferred

products to determining how a coordinated product set should be constructed under multiple objectives and constraints [15].

Against this background, the present study uses Western Zhou bronze cultural resources associated with the Baoji Bronze Ware Museum to connect product-level consumer preferences, preference heterogeneity, design and cultural-resource characteristics, and constrained portfolio decision making.

RQ1. What relative preference structure do consumers exhibit across specific museum cultural and creative products, and to what extent are consumer characteristics associated with heterogeneity in product-level preferences?

RQ2. To what extent are product-level consumer utilities associated with broader Product Concept, Translation Pathway, and cultural-resource characteristics, and to what extent can product-level consumer heterogeneity be summarized by Product Concept strategies? Are the corresponding design-classification inferences robust to independent label validation?

RQ3. When consumer utility is incorporated together with cultural and design requirements into constrained portfolio decision making, how are individual-product preferences reflected in portfolio-level selection, and how robust are the resulting portfolios to alternative portfolio sizes, objective priorities, and classification boundaries?

Together, these analyses connect product-level consumer choice and preference heterogeneity with design and cultural-resource characteristics and extend the resulting behavioral evidence to constrained portfolio decision support. The overall research framework is provided in Supplementary Figure S1.

## 2. Materials and Methods

### 2.1. Research Design and Product Stimulus Development

#### 2.1.1. Cultural Prototype Selection and Formalization

Western Zhou bronze cultural resources associated with the Baoji Bronze Ware Museum were used as the case sampling frame because formal museum and related documentary sources provided traceable evidence for systematically verifying artifact identity, chronology, visual characteristics, and cultural information. Cultural prototypes were identified using a theory-driven stratified purposive sampling procedure. An initial pool of 25 candidate artifacts or artifact groups was screened against mandatory criteria concerning historical period, bronze material, artifact identity, formal documentary evidence, completeness of core information, consistency of key facts, and the availability of visual and cultural resources suitable for design translation. Eligible candidates were then balanced across Western Zhou chronological phases, archaeological provenance, primary

function, visual morphology, and cultural meaning, while inscription level and prior cultural-product development were monitored to avoid excessive concentration. This procedure yielded 15 Western Zhou bronze cultural prototypes, comprising both individual artifacts and artifact groups, as the basis for subsequent cultural-resource formalization and product-concept development.

After the cultural prototypes had been established, a museum-informed interdisciplinary co-design workshop was conducted in April 2026 at the College of Art and Design, Shaanxi University of Science and Technology, as a formative procedure preceding product-concept development. Six participants took part, including two professors and four doctoral researchers; two were museum-affiliated and four were university-based. The 60-min workshop followed a structured 15–30–15 min sequence. Participants first independently identified salient cultural anchors, potential design-translation resources, and risks of cultural misinterpretation; they then discussed which cultural characteristics should be preserved, could be translated, or should not be excessively transformed. The final stage consolidated these discussions into 15 artifact-level synthesis summaries covering cultural anchors, design-translation opportunities, allowable transformation ranges, and interpretive boundaries. The workshop did not directly generate the final product concepts; subsequent cultural-resource coding, prototype-anchor formalization, and product-level classifications were undertaken by the research team using these outputs and the corresponding cultural evidence. The workshop served as a formative expert design-development activity and was not part of the subsequent consumer behavioral study; no individual-level behavioral outcomes or personal data were collected or analyzed for the consumer research.

#### 2.1.2. Product Concept Generation and Screening

Recent cultural-product design research has similarly emphasized the explicit identification of cultural carriers and the systematic translation of cultural elements into product-level design features [16]. Building on this cultural evidence, the research team established a cultural-resource and prototype-anchor framework and generated 66 candidate product concepts. Each concept was linked to a single primary bronze cultural prototype and incorporated at least two traceable prototype anchors, including at least one primary-level anchor.

Two design-classification systems were prespecified during product development. Product Concept (PC) comprised form-and-collectible, functional-experiential, interactive-mechanism, and narrative-expressive products (PC01–PC04), whereas Translation Pathway (TP) comprised form, motif, functional, semantic/structural-relationship, and historical-narrative translation (TP01–TP05). The 66 candidate concepts were subsequently screened and reviewed for cultural compliance, conceptual redundancy, and engineering feasibility, yielding 36 formal product concepts. Because PC and TP were assigned before consumer data collection, their original labels were retained as prespecified

design classifications in the primary analyses rather than retrospectively adjusted according to consumer preferences.

2.1.3. Visual Stimulus Design, Pretesting, and Standardization

After the 36 formal product concepts had been established, the research team prepared a standardized design-control specification for each product. Each specification documented the corresponding cultural prototype, core cultural anchors, prespecified Product Concept (PC) and Translation Pathway (TP), product function and structural mechanism, prototype-identifying features to be retained, permissible design translations, and cultural-interpretive boundaries. These specifications were finalized and frozen before visual stimulus production and formal consumer data collection.

A single lead designer with a product-design background translated the frozen specifications into visual stimuli within predefined design constraints. The assigned product function, structural mechanism, prototype-identifying features, cultural anchors, and PC/TP classifications were not altered during visual realization. Presentation variables, including composition, background, viewing angle, product prominence, lighting, shadows, and rendering detail, were standardized to reduce non-target visual variation across stimuli.

Two research-team members who were not involved in visual realization independently reviewed all stimuli using predefined criteria derived from the frozen specifications. Reviews focused on prototype identifiability, cultural-anchor preservation, product structure and functionality, and compliance with cultural-interpretive boundaries. Discrepancies were resolved by reference to the frozen specifications and supporting cultural evidence, and revisions were restricted to presentation-level adjustments. All revisions were completed before the pretest, and the final stimulus set was version-frozen before formal BWS administration.

The 36 stimuli were then evaluated in an independent pretest involving 20 adult participants who did not participate in the subsequent BWS study. All pretest participants provided informed consent before participation. Participants evaluated visual clarity, product-concept comprehensibility, consistency between the image and the corresponding product name and description, and the adequacy of the presentation for product evaluation using four five-point Likert items ranging from 1 (“strongly disagree”) to 5 (“strongly agree”). Stimulus order was randomized for each participant. Measures of attractiveness, liking, purchase intention, or relative preference were not collected, thereby preventing the pretest from functioning as a preference-based screening stage.

The pretest supported the suitability of the stimulus set for the formal experiment. Mean ratings were 4.38 (SD = 0.70) for visual clarity, 4.23 (SD = 0.75) for product-concept comprehensibility, 4.41 (SD = 0.69) for image–description consistency, and 4.32 (SD = 0.70) for presentation adequacy.

Pretest feedback prompted targeted review and limited presentation-level adjustments to selected stimuli, but no product concepts were removed or replaced, and the prespecified PC/TP classifications, core cultural anchors, and cultural-interpretive boundaries remained unchanged. The final stimuli were standardized to 720 × 540-pixel RGB PNG files and accompanied by consistently formatted Chinese-language product names and descriptions. Prices, branding, human figures, usage scenes, and unrelated promotional cues were excluded. All stimuli were linked through unique Stimulus_IDs and frozen before formal BWS administration. Representative stimuli covering the four Product Concept categories and five Translation Pathways are shown in Figure 1; the examples were selected to illustrate methodological coverage rather than consumer-preference outcomes.




Figure 1. Representative standardized product stimuli used in the BWS experiment. The eight examples illustrate the four Product Concept categories and five Translation Pathways represented within the full 36-stimulus set. All stimuli shown were created specifically for this study under the standardized design-control procedure described in Section 2.1.3.

2.2. BWS Experimental Design, Participants, and Data Quality

Relative purchase preferences for the 36 formal product stimuli were elicited using Best–Worst Scaling (BWS) [10]. The formal experiment was conducted on site under supervised conditions after institutional ethical approval was obtained in June 2026, with adult participants recruited through on-site convenience sampling. Eligibility criteria required participants to be at least 18 years old, able to

understand the BWS task instructions and clearly view the visual stimuli, and not to have participated previously in the cultural-prototype screening, co-design workshop, product-concept or stimulus-development stages, or the pre-experimental stimulus pretest. Eligible participants who provided informed consent were allocated in a 1:1:1:1 ratio using block randomization to V1_S1, V1_S2, V2_S1, and V2_S2. Before the formal tasks, research personnel provided standardized instructions on the experimental procedure and Best/Worst choice rules and confirmed task comprehension. During the experiment, only procedural and operational assistance was provided, without interpretive guidance regarding product content, cultural meaning, or purchase choices.

The BWS design comprised two task-composition versions (V1 and V2) and two task-presentation sequences (S1 and S2). V1 and V2 varied product co-occurrence across choice sets. Because presentation order can introduce ordering effects in repeated choice experiments, S2 presented the same tasks as the corresponding S1 condition in reverse order while retaining the within-task A–D product positions, thereby allowing presentation-order effects to be explicitly evaluated [17]. Each participant completed 18 BWS tasks, each presenting four standardized product stimuli consisting of an image, product name, and concise description. Participants were instructed to assume that products within a task were similarly priced and within an acceptable price range and to select the product they would be most willing to purchase (Best) and the product they would be least willing to purchase (Worst); the same product could not be selected for both responses within a task. Participants completing all tasks therefore made 18 Best and 18 Worst choices and encountered each of the 36 products twice.

In addition to the BWS choices, gender, age, museum cultural-product purchases during the previous 12 months, familiarity with Chinese bronze culture, and art-, design-, archaeology-, or museum-related educational or occupational background were collected for sample description and preference-heterogeneity analyses. Purchase experience was recorded in five ordered categories (0, 1, 2–3, 4–6, and ≥7 purchases) and coded from 0 to 4 for analysis; familiarity with Chinese bronze culture was measured on a five-point scale, whereas professional background was recorded as yes, no, or uncertain. After completing all BWS tasks, participants rated three five-point items reflecting collection orientation toward cultural or artistic products, the importance attached to practical functionality, and the importance attached to interactive experience. Data were recorded anonymously without direct personal identifiers. All participants received information regarding the study purpose, procedures, data use, voluntary participation, and the right to withdraw and provided informed consent before participation.

A total of 433 participants began the formal experiment, of whom 407 completed all tasks and generated complete BWS records eligible for data-quality screening. Exclusion criteria were established before the primary analyses and included missing or invalid Best/Worst responses, within-task Best–Worst overlap, selection of the same Best or Worst position in at least 17 of the 18

tasks, and total completion time below 60 s; faster responses, positional concentration, or duplicated response vectors that did not meet the exclusion thresholds were retained as quality flags [18]. A pre-freeze integrity audit identified Best–Worst overlaps in a subset of early records, which were corrected only after the affected participants were re-contacted and their intended choices reconfirmed; consequently, no such conflicts remained among the 407 records entering formal screening. Three records were excluded for extreme positional responding, two of which also met the <60-s criterion, yielding a final analytical sample of 404 participants: 105 in V1_S1, 92 in V1_S2, 97 in V2_S1, and 110 in V2_S2. Exclusion criteria were not modified in response to subsequent utility estimates, product rankings, or significance results.

The study protocol was reviewed and approved by the Ethics Committee of the College of Art and Design, Shaanxi University of Science and Technology (Ethical Approval ID: SUST20260617; approved on 17 June 2026). The corresponding approval documentation was administratively processed and issued by the Office of Science and Technology, Shaanxi University of Science and Technology. The visual-stimulus pretest and the formal BWS study were conducted only after ethical approval had been obtained. All procedures involving human participants were conducted in accordance with relevant institutional guidelines and regulations and the Declaration of Helsinki. All participants were adults aged 18 years or older and were informed of the study purpose, procedures, data use, voluntary nature of participation, and their right to withdraw before participation. Informed consent was obtained from all participants involved in the visual-stimulus pretest and the formal Best–Worst Scaling study. Data were collected anonymously without direct personal identifiers.

### 2.3. Consumer Preference Estimation and Design-Balanced Utilities

Standardized Best–Worst scores were first calculated to provide a descriptive comparison of overall preferences across the 36 products. For product $i$, the standardized score was defined as:

$$BW_i = \frac{B_i - W_i}{E_i}$$

where $B_i$, $W_i$, and $E_i$ denote the numbers of times product $i$ was selected as Best, selected as Worst, and exposed, respectively. Because the experimental design provided equal exposure for all 36 products, this measure directly reflected the overall direction and descriptive ranking of product preferences but was not used as the primary consumer-utility measure in subsequent analyses.

Product utilities were subsequently estimated using a conditional-logit MaxDiff model within a random-utility framework [19]. For a four-product choice task, the systematic utility of any possible ordered Best–Worst pair $(b, w)$ was defined as the difference between the utilities of the two products:

$$V(b, w) = U_b - U_w$$

The corresponding choice probability was:

$$P(b,w) = \frac{\exp(U_b - U_w)}{\sum_{j \neq k} \exp(U_j - U_k)}$$

For model identification, one product was fixed as the reference alternative and the remaining product parameters were estimated. Utilities for all 36 products were then transformed to a zero-sum centered scale such that $\sum_i U_i = 0$, allowing direct comparison on a common utility scale. Because each participant completed 18 repeated choice tasks, statistical inference used respondent-clustered robust standard errors to account for within-participant dependence in choice responses [20].

To assess whether task composition and presentation order affected relative product preferences, the MaxDiff model was further extended to include Product × Version, Product × Sequence, and Product × Version × Sequence interactions. Because Version and Sequence remained constant within a given choice task, their standalone main effects were not separately identified in the conditional-logit framework; inference therefore focused on their interactions with product utility. Global design effects were evaluated using likelihood-ratio tests, while condition-specific product effects were assessed using respondent-clustered robust Wald tests. For multiple product-level comparisons, the Benjamini–Hochberg procedure was applied to control the false discovery rate [21].

Because the realized sample sizes of the four Version × Sequence experimental cells were not identical, a Design-Balanced Utility was constructed so that the four cells contributed equally to the overall consumer-utility estimate. For product $i$:

$$U_i^{DB} = \frac{U_{i,V1S1} + U_{i,V1S2} + U_{i,V2S1} + U_{i,V2S2}}{4}$$

Accordingly, each experimental cell contributed a target weight of $25\%$ to the final utility estimate, preventing differences in realized cell sizes from implicitly weighting overall product utilities. Design-Balanced Utilities were further transformed to the same zero-sum centered scale and served as the primary consumer-utility measure for subsequent product-preference comparisons, consumer-heterogeneity analyses, product-attribute tests, and portfolio optimization. Pooled MaxDiff utilities and standardized Best–Worst scores were retained as descriptive and robustness benchmarks. Where exponentiated utilities were used for positive normalization, they were interpreted only as relative comparison indices rather than actual purchase probabilities.

Finally, a stricter sensitivity sample was constructed to assess whether the estimated preference structure depended on the data-quality boundaries specified in Section 2.2. In addition to the primary exclusion criteria, records with completion times of 60–74 s or with the same Best or Worst position selected in 12–16 of the 18 tasks were removed, yielding a sensitivity sample of $n = 388$ participants.

The MaxDiff model and Design-Balanced Utilities were then re-estimated using the same analytical procedure as in the primary sample. Preference structures from the $n = 404$ primary sample and the $n = 388$ sensitivity sample were compared using Pearson correlations, Spearman rank correlations, changes in product rankings, and Top-$K$ overlap. This sensitivity analysis was used solely to assess robustness to stricter response-quality criteria and did not replace the $n = 404$ primary analytical sample.

2.4. Consumer Heterogeneity and Design-Attribute Analysis

To examine whether consumer characteristics systematically altered relative preferences across the 36 products, Product × Consumer interaction models were estimated within the conditional-logit MaxDiff framework described in Section 2.3 while retaining Product × Version, Product × Sequence, and the corresponding higher-order design-control terms [22]. Consumer variables comprised museum cultural-product purchase experience during the previous 12 months, familiarity with Chinese bronze culture, collection orientation toward cultural or artistic products, importance attached to practical functionality, importance attached to interactive experience, and relevant professional background. The first five variables were standardized so that their interaction coefficients represented changes in relative product utility associated with a one-standard-deviation increase in the corresponding consumer characteristic. Professional background was analyzed only among participants providing an unambiguous yes/no response ($n = 391$) and was symmetrically effect-coded. Overall heterogeneity associated with each consumer variable was evaluated using conditional likelihood-ratio tests comparing the joint model with a reduced model excluding all corresponding Product interactions. The six global heterogeneity tests constituted a predefined testing family and were adjusted using the Benjamini–Hochberg false discovery rate procedure. For product-specific simple effects, FDR adjustment was applied separately within the 36 product-level tests associated with each consumer variable. Because consumer characteristics may be correlated, primary inference was based on conditional effects from the multivariable joint model rather than separate univariable models.

Secondary product-level analyses were conducted to assess whether the prespecified Product Concept (PC) and Translation Pathway (TP) classifications were systematically associated with variation in Design-Balanced Utility. Mean utilities, median utilities, and rankings were first summarized across PC and TP categories to characterize descriptive preference patterns. Because multiple products could originate from the same cultural prototype and prototype composition differed across PC and TP categories, category effects were further evaluated using OLS models containing fixed effects for the 15 cultural prototypes and HC3 heteroskedasticity-robust standard errors [23]. Overall PC and TP effects were assessed using joint F-tests. Unadjusted category-level utility differences were therefore treated as descriptive patterns, whereas the prototype-controlled

models assessed whether PC or TP accounted for utility variation beyond differences attributable to the underlying cultural prototypes. Because these analyses used estimated utilities for the 36 products as the analytical units, they were treated as secondary explanatory analyses of the primary consumer-preference model rather than as independent primary utility models.

Associations between cultural-resource characteristics and consumer preference were explored at the level of the 15 cultural prototypes to avoid treating shared resource ratings attached to multiple products from the same prototype as independent observations. For each prototype, Design-Balanced Utilities were averaged across its associated products and correlated using Spearman coefficients with six frozen cultural-resource dimensions: form extractability, motif extractability, narrative potential, functional-analogy potential, interaction potential, and cultural-misinterpretation risk. The six correlations constituted a single testing family and were adjusted using the Benjamini–Hochberg FDR procedure. Given the prototype-level sample size of 15, these analyses were treated as exploratory resource-level association analyses, intended to examine whether consistent association patterns existed between cultural-resource translation potential and consumer preference rather than to provide confirmatory or causal evidence or to treat resource scores as substitutes for consumer-derived utility.

Consumer × PC interaction models were additionally estimated within the MaxDiff framework to examine whether consumer heterogeneity could be summarized by higher-level product-concept strategies. The five standardized consumer variables were jointly interacted with PC category, whereas professional background was evaluated in the same $n = 391$ subsample while controlling for the remaining consumer interactions. PC01 served as the reference category for identification, after which the four PC-specific slopes were centered to sum to zero. Global PC-level heterogeneity for each consumer variable was evaluated using respondent-clustered robust Wald tests with three degrees of freedom, and the six global tests constituted a single testing family adjusted using the Benjamini–Hochberg procedure. These analyses operated at a different level from the 36-product heterogeneity models: Product × Consumer models assessed whether consumer characteristics altered utilities for specific products, whereas Consumer × PC models examined whether such heterogeneity could be summarized by the four broader product-concept strategies. Consumer-level aggregation focused on PC because PC represented consumer-facing product strategies, whereas TP primarily characterized the underlying design-translation pathway.

### 2.5. Post-Hoc Independent Stimulus Validation and Label-Sensitivity Analysis

To assess the reliability of the PC and TP classifications and the sensitivity of related behavioral inferences to classification boundaries, the 36 final stimuli underwent independent post-hoc validation after BWS data collection. Two independent coders, blinded to the research team's original PC/TP labels, consumer-preference results, and subsequent portfolio-optimization outcomes, first

independently assigned one primary PC and one primary TP to each stimulus. After completing this classification stage, the coders were provided with the corresponding cultural-prototype evidence and independently rated prototype identifiability, cultural-anchor preservation, and boundary compliance on ordinal scales ranging from 0 to 2. Agreement for PC and TP classifications was assessed using Cohen's $\kappa$, whereas agreement for the three ordinal cultural-validity ratings was evaluated using quadratic-weighted Cohen's $\kappa$ [24,25]. Disagreements were reviewed against the prespecified operational definitions, final stimuli, and corresponding cultural evidence to determine the final validated classifications and cultural-validity ratings. The validation procedure therefore distinguished classification reliability from cultural validity: the former assessed whether PC/TP category boundaries could be independently identified, whereas the latter assessed whether the final stimuli retained traceable prototype characteristics, cultural anchors, and interpretive boundaries.

Because the research team's original PC and TP labels had been established during product development before consumer data were generated, the independent validation results were not used to retrospectively alter the primary analysis labels. The original classifications remained the prespecified labels for the primary behavioral analyses, whereas the independently validated labels were used only in a label-sensitivity branch. In this branch, consumer data, Design-Balanced Utilities, statistical models, and testing procedures were held fixed while the PC/TP labels alone were replaced. The corresponding PC/TP product-level analyses and Consumer × PC models were then repeated to determine whether the principal behavioral inferences depended on classification boundaries.

## 2.6. Product Portfolio Optimization and Algorithmic Validation

### 2.6.1. Portfolio Objective and Constraints

Following consumer-preference estimation, design-attribute analysis, and cultural-resource evaluation, the 36 formal product concepts were formulated as a constrained portfolio-decision problem. Unlike ranking individual products, portfolio selection requires simultaneous consideration of consumer preference, cultural-resource translation potential, cultural-expression safety, and design-strategy balance at the portfolio level. Therefore, consumer-derived utilities and cultural-design attributes were jointly incorporated as decision inputs.

For candidate product $i$, a binary decision variable $x_i$ was defined, where $x_i = 1$ indicates that product $i$ is selected and $x_i = 0$ otherwise. All 36 formal products entered the optimization candidate space without utility-based pre-screening. The Main-Balanced scenario was defined as the primary specification with a fixed portfolio size of $K = 8$, whereas $K = 6$ and $K = 10$ were examined only in portfolio-size sensitivity analyses. Optimization inputs included the Design-Balanced Utility estimated in Section 2.3, cultural-resource translation indicators, cultural-misinterpretation risk, and the prespecified PC and TP classifications established during product development. Consumer-

heterogeneity parameters were not directly incorporated into the optimization objective and were instead used for post-hoc profiling of selected portfolios, avoiding unsupported population weighting assumptions.

The portfolio objective was defined as:

$$J(X) = w_U F_U(X) + w_C F_C(X) + w_S F_S(X) + w_D F_D(X)$$

where $F_U$, $F_C$, $F_S$, and $F_D$ represent consumer preference, cultural-pathway fit, cultural safety, and design-strategy diversity, respectively.

The consumer-preference component $F_U$ was constructed from the Design-Balanced Utility:

$$F_U(X) = \frac{1}{K}\sum_{i=1}^{36} x_i \, U_i^{norm}, \qquad U_i^{norm} = \frac{U_i^{DB} - U_{\min}}{U_{\max} - U_{\min}}$$

where $U_i^{DB}$ denotes the Design-Balanced Utility estimated in Section 2.3, and $U_{\min}$ and $U_{\max}$ denote the minimum and maximum utility values among the 36 candidate products. This normalization preserved relative preference ordering while placing different objective components on a comparable scale.

The cultural-pathway fit component $F_C$ measured the alignment between product strategies and cultural-resource translation potential. PathFit was assigned according to the prespecified PC category: PC01 used the mean of form and motif extractability, PC02 used functional-analogy potential, PC03 used interaction potential, and PC04 used narrative potential. PathFit was subsequently transformed into Culture_norm:

$$F_C(X) = \frac{1}{K}\sum_{i=1}^{36} x_i \, C_i^{norm}, \qquad C_i^{norm} = \frac{PathFit_i - 1}{4}$$

where $C_i^{norm}$ denotes the normalized cultural-pathway fit score for product $i$.

The cultural-safety component $F_S$ was derived from cultural-misinterpretation risk. Risk scores were reversed so that lower misinterpretation risk contributed higher safety values:

$$F_S(X) = \frac{1}{K}\sum_{i=1}^{36} x_i \, S_i^{norm}, \qquad S_i^{norm} = \frac{5 - MisreadRisk_i}{4}$$

where $S_i^{norm}$ denotes the normalized cultural-safety score for product $i$. Cultural-misinterpretation risk was therefore treated as a portfolio-level trade-off rather than an exclusion criterion, allowing culturally valuable but interpretation-sensitive concepts to remain eligible.

The portfolio diversity component evaluated balance across PC categories and TP pathways, using Shannon entropy as a quantitative measure of portfolio diversity [26]. Let $p_c$ and $q_t$ represent the proportions of selected products belonging to PC category $c$ and TP pathway $t$, respectively. Normalized Shannon entropy was defined as:

$$F_D(X) = \frac{1}{2}\frac{H_{PC}}{\ln 4} + \frac{1}{2}\frac{H_{TP}}{\ln 5}$$

where:

$$H_{PC} = -\sum_{c=1}^{4} p_c \ln p_c\,, \qquad H_{TP} = -\sum_{t=1}^{5} q_t \ln q_t$$

Here, $H_{PC}$ and $H_{TP}$ denote the Shannon entropy of the PC and TP distributions, respectively. This formulation encouraged balanced representation across design strategies rather than simply aggregating individual product values. PC and TP therefore served two distinct roles: entropy promoted portfolio-level balance, whereas coverage constraints ensured minimum representation.

The Main-Balanced scenario assigned equal weights:

$$w_U = w_C = w_S = w_D = 0.25$$

Equal weighting was not interpreted as evidence that the four objectives were equally important in practice. Instead, it served as a neutral baseline specification because no independent managerial preference data were available to empirically identify objective weights [27]. Alternative market-oriented and culture-oriented weighting scenarios were evaluated in sensitivity analyses.

All portfolios were required to satisfy prespecified hard constraints:

$$\sum_{i=1}^{36} x_i = K$$

The feasibility constraints were:

1. only ENGINEERING_SCREENED products were eligible;

2. no more than one product could be selected from the same Western Zhou cultural prototype;

3. all four PC categories had to be represented;

4. all five TP pathways had to be represented;

5. for $K = 8$, each PC category or TP pathway could contribute no more than three products;

6. for $K = 6$ and $K = 10$, the corresponding upper limits were adjusted to two and four products.

These requirements jointly defined the feasible portfolio space. Engineering eligibility, portfolio size, prototype uniqueness, and PC/TP coverage were implemented as hard constraints rather than approximated through penalty functions.

2.6.2. Constrained B-SBOA Implementation

To solve the discrete portfolio problem defined in Section 2.6.1, a constrained Binary Secretary Bird Optimization Algorithm (B-SBOA) was developed from the population-based search framework of the Secretary Bird Optimization Algorithm (SBOA) [28]. The original SBOA balances global exploration and local exploitation through hunting and escape update mechanisms in a continuous search space. Recent work has demonstrated that SBOA can be adapted to binary combinatorial optimization through explicit binarization mechanisms [29]. The present implementation retained these core continuous position-update mechanisms while adapting them to the binary portfolio-decision space defined by the 36 candidate products. All 36 formal products entered the search space directly, without Top-10% pre-screening based on individual utility or cultural indicators, and no portfolio-size penalty was used. Fixed portfolio size, engineering eligibility, cultural-prototype uniqueness, and PC/TP coverage requirements were instead handled through explicit feasibility constraints, consistent with the broader principle that constraint handling in population-based optimization need not rely exclusively on penalty-function formulations [30].

A population size of $M = 40$ was used. Initial population positions were first generated in the 36-dimensional continuous search space and mapped to binary candidate portfolios using a fixed threshold of 0.5: for each dimension, the corresponding product was assigned a value of 1 when the continuous position exceeded 0.5 and 0 otherwise. The resulting binary portfolios were then subjected to candidate-aware feasibility repair to restore feasibility under the corresponding scenario. Following Repair, the associated continuous positions were synchronized to the repaired 0/1 feasible states and portfolio objective values were recalculated, ensuring consistency among position, portfolio state, and fitness before formal iterations began. During subsequent iterations, the original SBOA hunting and escape update mechanisms were retained. After each continuous position update, the same 0.5-threshold binary-mapping rule was applied, followed by candidate-aware feasibility repair and reevaluation using the objective function defined in Section 2.6.1. Lévy-flight perturbation was retained in the attack stage, with $\eta = 1.5$ and a scaling factor of 0.01 in the formal implementation. The formal implementation therefore retained the original SBOA continuous position-update mechanisms, followed by binary mapping, candidate-aware feasibility repair, and objective reevaluation after each position update.

Candidate-aware feasibility repair was restricted to restoring the current candidate solution to the formal feasible space. Repair checked and restored engineering eligibility, cultural-prototype uniqueness, PC/TP coverage and category bounds, and the exact portfolio size $K$. The procedure

operated directly on the current candidate and its constraint state and did not rely on a pregenerated feasible-solution library, anchor pool, preoptimized portfolio, or information from the Exact optimizer. Repair therefore served as a feasibility-restoration procedure rather than as an independent portfolio optimizer.

To distinguish improvements attributable to B-SBOA search from those produced by an additional local-improvement procedure, the 1-swap local-search module available in the underlying implementation was treated as optional and disabled in all formal runs. The formal optimization path therefore consisted of continuous SBOA search, binary mapping, and candidate-aware feasibility repair, with Repair itself not performing repeated 1-swap local optimization.

Each optimization scenario was run for 300 iterations using 30 fixed random seeds for independent repetitions. A Repair-only baseline was additionally constructed to evaluate the incremental contribution of the B-SBOA search stage under a common feasibility mechanism. This baseline used the same random seeds, initialization procedure, and candidate-aware feasibility repair as B-SBOA but omitted all subsequent B-SBOA iterations. The Exact optimizer was kept strictly separate from the stochastic search process and was not used during initialization, Repair, position updates, or local search. It served only as an independent post-hoc optimality benchmark, as described in Section 2.6.3.

2.6.3. Exact Benchmark and Sensitivity Analyses

To evaluate the solution quality of B-SBOA for the current 36-product portfolio problem, an Exact benchmark was constructed independently of the stochastic search process, consistent with the distinction between metaheuristic search and exact methods that can establish solution optimality [31]. The formal Exact solver used prototype-grouped dynamic programming, enumerating either no selection or the selection of one product within each cultural prototype and retaining the optimal additive objective contribution for states defined by portfolio size and PC/TP category counts. The corresponding portfolio-diversity component was incorporated once category counts were determined. Because this procedure retained the objective function and hard constraints specified in Section 2.6.1, it provided the global optimum benchmark for the present 36-product problem. The Exact solver was executed only after the Repair-only and B-SBOA stochastic runs had been completed and was not used during B-SBOA initialization, Repair, position updates, or local search.

Each B-SBOA run was compared with the Exact optimum under the corresponding scenario. Let $J_r$ denote the final objective value from stochastic run $r$ and $J^*$ the corresponding Exact optimum. The optimality gap was defined as:

$$Gap_r = \frac{J^* - J_r}{J^*} \times 100\%$$

Solution quality across the 30 independent runs was summarized using the best objective value, mean objective value, standard deviation, number of Exact hits, mean optimality gap, and maximum optimality gap. For runs attaining the Exact optimum, the first iteration at which the optimum was reached was additionally recorded to characterize stochastic convergence. Product-level selection frequencies across the 30 runs were also calculated to assess portfolio-membership stability.

To distinguish the contribution of B-SBOA search from feasibility restoration alone, Repair-only and B-SBOA used the same 30 random seeds, initialization procedure, and candidate-aware feasibility repair and were paired by seed. This component-level comparison was designed to isolate the incremental contribution of the B-SBOA search stage from the common feasibility-restoration mechanism, consistent with recommendations that metaheuristic proposals separately examine component contributions and statistically validate performance differences [32]. Within each scenario, final objective values from the two procedures were compared using a one-sided paired Wilcoxon signed-rank test to assess whether the addition of B-SBOA search produced a positive improvement under a common initialization and feasibility mechanism [33]. The Exact benchmark was not treated as a third stochastic algorithm in this statistical comparison and was used only to establish the global optimum, calculate optimality gaps, and quantify Exact-hit performance. The formal algorithmic comparison therefore concerned B-SBOA with feasibility repair versus Repair-only and excluded the disabled 1-swap local-search module.

Sensitivity analysis provides a means of examining the robustness of multi-criteria decision outcomes to alternative decision specifications [34]. Scenario sensitivity was evaluated along two dimensions: portfolio size and objective weighting. The Main-Balanced specification used $K = 8$ with $w_U = w_C = w_S = w_D = 0.25$. Portfolio-size sensitivity retained equal objective weights while varying portfolio size to $K = 6$ and $K = 10$. Weight sensitivity retained $K = 8$ and introduced two alternative decision-priority scenarios. The Market-Oriented scenario used $w_U/w_C/w_S/w_D = 0.45/0.20/0.15/0.20$, whereas the Culture-Oriented scenario used $0.20/0.40/0.25/0.15$. All scenarios used the same formal candidate set and objective-construction principles, together with their corresponding portfolio-size and category constraints, and were evaluated using Repair-only, 30 independent B-SBOA runs, and the Exact solver.

Scenario robustness focused primarily on optimal portfolio membership rather than direct comparisons of absolute objective values across differently specified scenarios. For any two portfolios $A$ and $B$, membership overlap was quantified using the Jaccard index [35]:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Common, entered, and exited products and cross-scenario product-retention patterns were additionally examined to identify products retained across changes in portfolio size or decision

priorities. Because different values of $K$ directly alter portfolio size, Jaccard overlap across size scenarios was interpreted together with the numbers of common, entered, and exited products rather than in isolation. More generally, changes in portfolio size or objective weights alter the feasible space or evaluation function; absolute objective values $J$ from different scenarios were therefore not interpreted as indicating that one scenario was superior to another. Scenario comparisons instead focused on portfolio composition and its sensitivity to alternative decision settings.

As the portfolio-decision extension of the label-sensitivity analysis described in Section 2.5, the prespecified PC/TP labels in the optimization input were replaced with the independently validated PC/TP labels. Except for the PC/TP labels and the objective components and constraints directly determined by them, Design-Balanced Utilities, underlying cultural-resource indicators, normalization principles, scenario weights, engineering status, non-PC/TP constraints, Repair logic, B-SBOA parameters, and random seeds were held fixed. The Exact-solver procedure was likewise held fixed, but the corresponding Exact optimum was recomputed under the validated labels. The Original-label baseline was first reproduced under the frozen specification, after which the validated-label branch was evaluated under otherwise identical analytical conditions. This procedure followed the label-sensitivity principle of keeping the data, models, and algorithms fixed while changing only the PC/TP labels.

Label robustness was assessed separately in terms of objective-value and portfolio-membership changes within the same scenario, including relative changes in Exact objective values, Jaccard overlap, common/entered/exited products, and 30-run product-selection frequencies. Objective-value robustness and portfolio-composition robustness were therefore treated as distinct properties, avoiding the interpretation of similar optimal objective values as evidence of equally stable portfolio membership.

Finally, consumer-heterogeneity parameters were used only for post-hoc characterization of the selected portfolios. For the prespecified consumer dimensions, portfolio-level mean slopes, the numbers of products with positive and negative slopes, and products with the most extreme slopes were summarized to assess whether selected portfolios exhibited systematic consumer-profile tendencies; these measures did not enter the optimization objective or alter portfolio selection.

### 2.7. Statistical Reporting

Unless otherwise specified, statistical inference used a two-sided alpha level of 0.05. Likelihood-ratio tests were used for nested-model comparisons, respondent-clustered robust Wald tests for coefficient-based inference in the conditional-logit models, joint F-tests with HC3 heteroskedasticity-robust standard errors for prototype-controlled product-level models, and Spearman rank correlations for the exploratory prototype-level associations. The paired comparison between B-SBOA and Repair-only was evaluated using a one-sided Wilcoxon signed-rank test because the prespecified

directional question was whether the addition of the B-SBOA search stage improved objective values under matched initialization and feasibility-repair conditions. Benjamini–Hochberg false discovery rate correction was applied to the predefined multiple-testing families described above. Sample sizes are reported for the corresponding analyses, and exact $P$values are reported where available; values below the numerical reporting threshold are reported as $P < 0.001$.

Code availability

The custom code used to implement the constrained B-SBOA, candidate-aware feasibility-repair procedure, Repair-only baseline, Exact benchmark, and portfolio-sensitivity analyses is available from the corresponding author upon reasonable request.

Use of generative artificial intelligence

ChatGPT (OpenAI) was used during manuscript preparation solely to assist with English-language editing, wording refinement, structural organization, and textual consistency checking. All AI-assisted text was reviewed, revised, and verified by the authors. Generative AI was not used to formulate the research questions, design the study, generate research data, conduct statistical analyses, interpret the results, or draw the scientific conclusions. The authors take full responsibility for the final content of the manuscript.

## 3. Results

### 3.1. Sample Characteristics and Descriptive Preferences

Among the 407 participants who completed all BWS tasks, three were excluded according to the prespecified data-quality criteria, yielding a final analytical sample of 404 participants across all four experimental cells. Sample characteristics, museum cultural-product purchase experience, cultural familiarity, professional background, and consumer-orientation measures are summarized in Table 1.

Each of the 36 products received 808 exposures. Standardized Best–Worst scores ranged from −0.2624 to 0.3837. C033 received the highest score (0.3837), followed by C029 (0.2859) and C014 (0.2005), whereas C004 (−0.2624), C008 (−0.1906), and C020 (−0.1881) received the lowest scores. These scores provide the descriptive preference ordering; design-adjusted MaxDiff utilities and Design-Balanced Utilities are reported in Section 3.2.

Table 1. Characteristics of the Final Analytical Sample

| Characteristic | Category | *n* (%) or Mean (SD) |
|---|---|---|
| Gender | Male | 140 (34.7%) |
| | Female | 264 (65.3%) |
| Age | 18–24 years | 106 (26.2%) |

| | | |
|---|---|---|
| | 25–34 years | 284 (70.3%) |
| | ≥35 years | 14 (3.5%) |
| Museum cultural-product purchases, previous 12 months | 0 | 128 (31.7%) |
| | 1 | 98 (24.3%) |
| | 2–3 | 148 (36.6%) |
| | 4–6 | 23 (5.7%) |
| | ≥7 | 7 (1.7%) |
| Relevant professional background | Yes | 228 (56.4%) |
| | No | 163 (40.3%) |
| | Uncertain | 13 (3.2%) |
| Familiarity with Chinese bronze culture | — | 2.84 (1.21) |
| Collection orientation | — | 3.99 (1.16) |
| Importance of practical functionality | — | 3.48 (1.38) |
| Importance of interactive experience | — | 3.75 (1.32) |

Note. Values are presented as n (%) for categorical variables and mean (SD) for continuous variables. Percentages may not total 100% because of rounding. Familiarity with Chinese bronze culture, collection orientation, importance of practical functionality, and importance of interactive experience were measured on five-point scales.

### 3.2. MaxDiff Utilities, Design Effects, and Robustness

The pooled MaxDiff model showed significant differentiation in relative utilities across the 36 products ($\chi^2(35) = 1067.931, p < 0.001$). After experimental-design factors were incorporated, Product × Sequence, Product × Version, and Product × Version × Sequence effects were all significant (Supplementary Table S1). No individual Sequence effect remained significant after FDR correction within either V1 or V2. Version differences remained significant for five products: C004, C005, and C019 under S1, and C016 and C020 under S2; C004 and C011 remained significant for the Version × Sequence difference-in-differences. Relative product utilities therefore showed some sensitivity to experimental condition.

Design-Balanced Utilities were consequently used as the primary preference measure. Utilities ranged from −0.473 to 0.830, with C033 ranked highest, followed by C029, C028, C022, and C014;

C004 ranked lowest (Figure 2). The design-balanced estimates were highly consistent with pooled MaxDiff utilities (Pearson $r = 0.9929$; Spearman $\rho = 0.9864$). The five highest-ranked products were identical, and eight of the top 10 overlapped, indicating that design balancing produced mainly local ranking adjustments rather than changing the overall preference structure.

After an additional 16 participants meeting the stricter response-quality criteria were excluded, the sensitivity sample decreased from $N = 404$ to $N = 388$. Re-estimated Design-Balanced Utilities remained highly consistent with the primary estimates (Pearson $r = 0.9989$; Spearman $\rho = 0.9985$); the Top 5 and Top 10 sets were identical, and the maximum rank change was one position. The primary $N = 404$ sample was therefore retained for subsequent analyses.

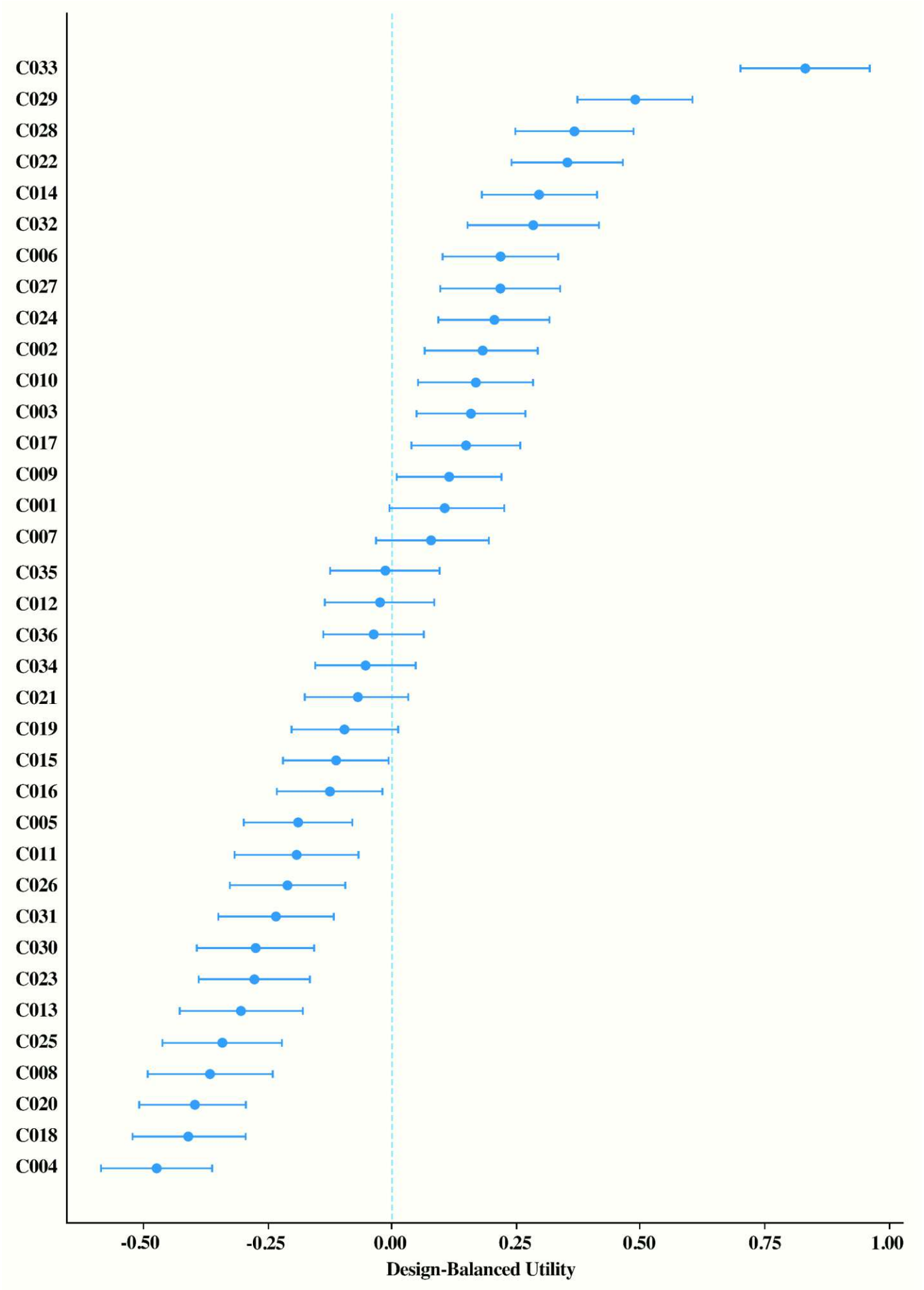

C033
C029
C028
C022
C014
C032
C006
C027
C024
C002
C010
C003
C017
C009
C001
C007
C035
C012
C036
C034
C021
C019
C015
C016
C005
C011
C026
C031
C030
C023
C013
C025
C008
C020
C018
C004
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
Design-Balanced Utility

Figure 2. Design-Balanced Utilities for the 36 product concepts. Points indicate Design-Balanced Utility estimates, and horizontal intervals indicate 95% confidence intervals. Products are ordered from highest to lowest utility; the vertical reference line indicates zero.

### 3.3. Consumer Preference Heterogeneity

After jointly controlling for the remaining consumer characteristics and experimental-design terms, five of the six prespecified consumer variables showed significant global heterogeneity in product preferences (Table 2). Museum cultural-product purchase experience, collection orientation, importance of practical functionality, importance of interactive experience, and relevant professional background remained significant after BH-FDR correction, whereas familiarity with Chinese bronze culture did not show an independent significant effect. Professional background was evaluated in the subsample with unambiguous yes/no responses; the remaining variables were evaluated in the full analytical sample.

Product-level effects differed across consumer characteristics. Purchase experience showed significant effects for six products, with positive slopes for C033 (0.287), C022 (0.182), and C010 (0.180), and negative slopes for C013 (−0.240), C006 (−0.234), and C012 (−0.166). Professional background was associated with higher relative utility for C019 (0.511) and lower relative utility for C020 (−0.390). Although collection orientation, practical-function importance, and interaction importance showed significant global heterogeneity, no individual product remained significant after FDR correction within their respective 36 product-level tests, suggesting that these effects were distributed across multiple products rather than concentrated in a small number of alternatives.

Table 2. Global Tests and Product-Level FDR Summary of Consumer Preference Heterogeneity

| Consumer characteristic | $N$ | Conditional LR $\chi^2(35)$ | BH-FDR $q$ | FDR-significant products（$n$） |
|---|---|---|---|---|
| Museum cultural-product purchase experience | 404 | 168.161 | <0.001 | 6 |
| Familiarity with Chinese bronze culture | 404 | 45.282 | 0.114 | 0 |
| Collection orientation | 404 | 56.290 | 0.015 | 0 |
| Importance of practical functionality | 404 | 123.764 | <0.001 | 0 |
| Importance of interactive experience | 404 | 82.742 | <0.001 | 0 |
| Relevant professional background | 391 | 105.401 | <0.001 | 2 |

Note. Conditional likelihood-ratio tests evaluated each consumer characteristic while retaining the remaining consumer variables and experimental-design controls. The six global heterogeneity tests constituted one predefined testing family and were adjusted using the Benjamini–Hochberg false

discovery rate procedure. Product-level significance was assessed separately within the 36 product-specific slopes for each consumer characteristic. For the five standardized continuous variables, slopes represent changes in relative product utility associated with a one-standard-deviation increase in the corresponding characteristic. Professional background was evaluated among participants providing an unambiguous yes/no response and compares participants with and without a relevant professional background.

### 3.4. Design Strategies, Cultural Resources, and Validation Robustness

Under the prespecified PC/TP labels, PC04 and PC03 showed descriptively higher mean utilities than PC02 and PC01, while TP05 showed the highest descriptive mean utility. After controlling for the 15 cultural prototypes, however, neither the overall PC nor TP effect was significant. These category-level differences were therefore treated as descriptive rather than as evidence that PC or TP classifications systematically explained consumer utility.

At the level of the 15 cultural prototypes, none of the six cultural-resource dimensions—form extractability, motif extractability, narrative potential, functional-analogy potential, interaction potential, or cultural-misinterpretation risk—showed a significant association with mean Design-Balanced Utility after BH-FDR correction. Thus, within this exploratory prototype-level analysis, no consistent monotonic relationship was detected between cultural-resource potential and consumer-derived utility.

Consumer × PC analyses showed that product-level heterogeneity could only partly be summarized by broader Product Concept strategies (Supplementary Table S2 and Fig. 3). Purchase experience, collection orientation, and importance of practical functionality remained significant after BH-FDR correction, whereas cultural familiarity, importance of interactive experience, and professional background did not. Higher purchase experience was associated with relatively positive slopes for PC03 and PC04 and negative slopes for PC01 and PC02, while collection orientation and practical-function importance were most positively associated with PC02. These results indicate that some consumer heterogeneity occurred primarily among specific products rather than being fully captured by the four PC categories.

Independent post-hoc validation showed substantial agreement for both classifications, with Cohen's κ values of 0.779 for PC and 0.825 for TP. Quadratic-weighted κ values were 0.645 for Prototype Identifiability, 0.723 for Cultural Anchor Preservation, and 0.823 for Boundary Compliance, and no cultural-validity red-line products were identified. Validated labels matched 88.9% of the prespecified PC labels and 91.7% of the prespecified TP labels, with six products changing PC and/or TP classification. Despite these changes, the prototype-controlled PC and TP effects remained nonsignificant, while the BH-FDR significance status and slope-sign patterns of all six Consumer × PC tests remained unchanged.

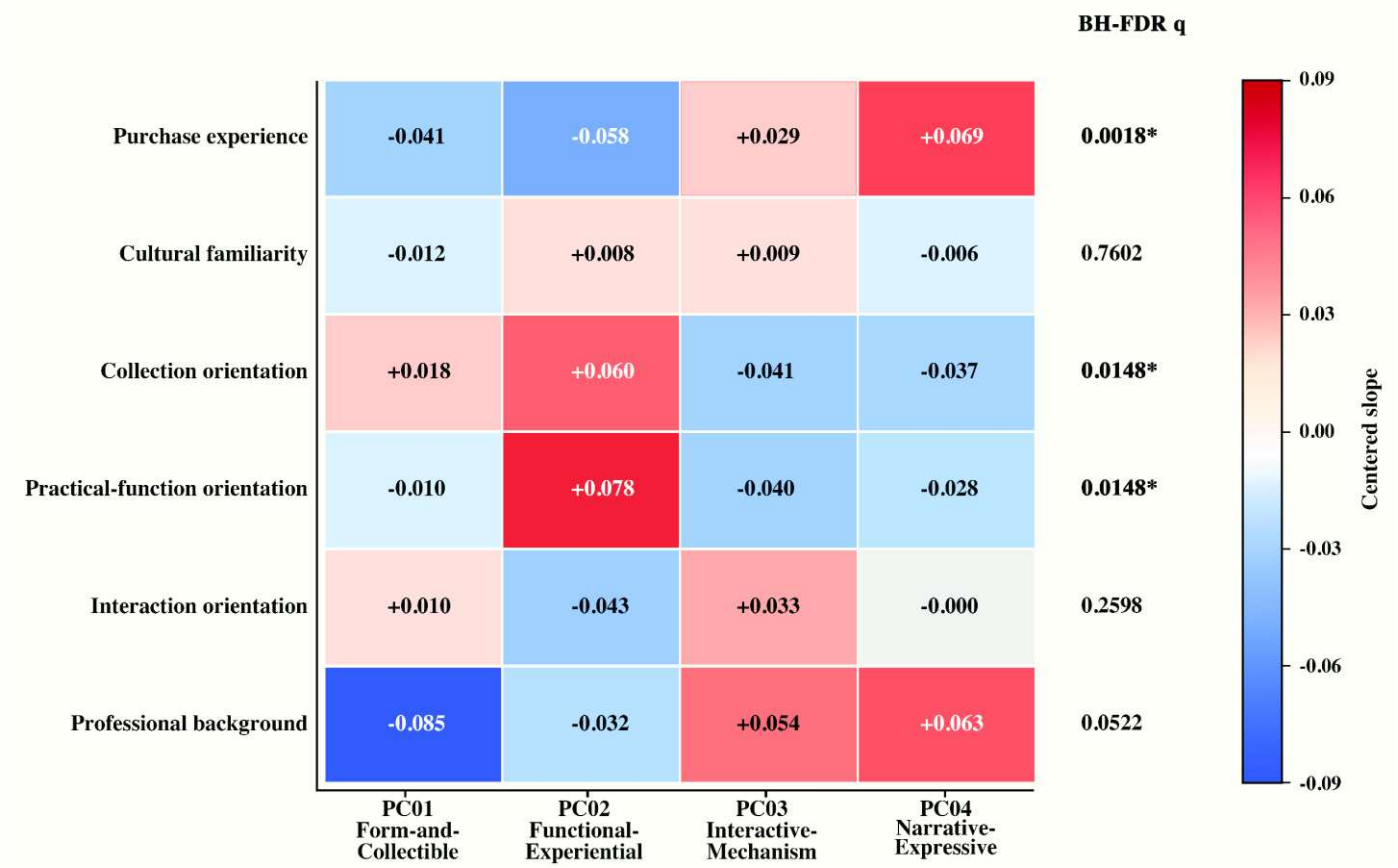



Figure 3. Consumer heterogeneity across Product Concept strategies. Cells show centered Consumer × PC slopes. BH-FDR-adjusted q values are shown on the right; asterisks indicate q < 0.05.

3.5. Behavioral Evidence-Informed Portfolio Optimization

Under the prespecified-label Main-Balanced specification $(K = 8)$, the Exact benchmark identified a global optimum of $J = 0.690506$. The optimal portfolio comprised C006, C012, C022, C024, C026, C028, C033, and C034, representing eight distinct cultural prototypes (Table 3). PC01–PC04 were each represented by two products, while TP01–TP05 were represented by 2, 1, 3, 1, and 1 products, respectively, satisfying all prototype-uniqueness and PC/TP coverage constraints.

The optimal portfolio differed from simply selecting the eight products with the highest Design-Balanced Utilities. Five selected products ranked within the consumer-utility Top 10, whereas the remaining three ranked 18th, 20th, and 27th. Portfolio selection therefore reflected consumer preference together with cultural-pathway fit, cultural safety, design-strategy diversity, and the prespecified feasibility constraints rather than individual-product preference rankings alone.

Consumer-heterogeneity parameters were not included in the optimization objective. Post-hoc profiling showed both positive- and negative-slope products across purchase experience, cultural familiarity, collection orientation, practical-function importance, interaction orientation, and professional background, with no uniformly directional consumer profile across the selected portfolio.

Table 3. Exact Main-Balanced Optimal Portfolio under the Prespecified PC/TP Labels

| Product ID | Cultural Prototype | PC | TP | Design-Balanced Utility Rank |
|---|---|---|---|---|
| C006 | WZ03 | PC01 | TP01 | 7 |
| C012 | WZ05 | PC02 | TP03 | 18 |
| C022 | WZ09 | PC04 | TP04 | 4 |
| C024 | WZ10 | PC04 | TP05 | 9 |
| C026 | WZ11 | PC02 | TP03 | 27 |
| C028 | WZ12 | PC03 | TP03 | 3 |
| C033 | WZ14 | PC03 | TP02 | 1 |
| C034 | WZ15 | PC01 | TP01 | 20 |

Note. The portfolio is the Exact global optimum under the Main-Balanced specification (K=8) using the prespecified PC/TP labels, with $J^*$=0.690506. All eight selected products originated from distinct cultural prototypes. PC01–PC04 were each represented by two products, while TP01–TP05 were represented by 2, 1, 3, 1, and 1 products, respectively. Design-Balanced Utility ranks refer to the full set of 36 candidate products.

3.6. Algorithmic, Scenario, and Label Robustness

Across all five prespecified optimization scenarios, the best B-SBOA run attained the corresponding Exact global optimum, although Exact-hit stability differed across scenarios (Table 4). Exact-hit rates were 10/30 for Main-Balanced, 30/30 for Size-Low, 9/30 for Size-High, and 30/30 for both Market-Oriented and Culture-Oriented scenarios. Under matched random seeds, initialization, and feasibility repair, B-SBOA significantly improved objective values over Repair-only in all five scenarios, indicating a consistent contribution of the search stage beyond feasibility restoration alone.

Table 4. Algorithmic Performance and Exact-Benchmark Validation across Prespecified Portfolio Scenarios

| Scenario | $K$ | Repair-only Mean $J$ | B-SBOA Mean $J$ ± SD | Exact $J^*$ | Exact Hits $n/30$ | Mean Gap (%) | B-SBOA > Repair-only $n/30$ |
|---|---|---|---|---|---|---|---|
| Main-Balanced | 8 | 0.681586 | 0.689542 ± 0.000819 | 0.690506 | 10/30 | 0.140 | 29/30 |
| Size-Low | 6 | 0.695315 | 0.701756 ± 0.000000 | 0.701756 | 30/30 | 0.000 | 30/30 |
| Size-High | 10 | 0.667902 | 0.680775 ± 0.001802 | 0.682617 | 9/30 | 0.270 | 30/30 |
| Market-Oriented | 8 | 0.666053 | 0.670508 ± 0.000000 | 0.670508 | 30/30 | 0.000 | 24/30 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Culture-Oriented | 8 | 0.688679 | 0.705374 ± 0.000000 | 0.705374 | 30/30 | 0.000 | 30/30 |

Note. B-SBOA and Repair-only were paired using the same 30 random seeds, initialization procedure, and candidate-aware feasibility-repair mechanism. Exact denotes the independently computed global optimum and was used only as a post-hoc benchmark. Exact Hits indicates the number of B-SBOA runs attaining the Exact optimum; Mean Gap is the mean relative optimality gap across the 30 runs. B-SBOA > Repair-only indicates the number of paired seeds for which B-SBOA produced a higher final objective value; the remaining runs were ties. One-sided paired Wilcoxon signed-rank tests indicated significant improvement of B-SBOA over Repair-only in all five scenarios ($P < 0.05$).

Portfolio composition was sensitive to both portfolio size and objective weighting (Figure 4). Relative to the Main-Balanced Exact portfolio, Jaccard indices were 0.556 for Size-Low, 0.636 for Size-High, 0.455 for Market-Oriented, and 0.333 for Culture-Oriented. C006, C024, and C028 were retained across all five Exact portfolios, while C022 and C033 were retained in four. Thus, the optimal portfolios contained both stable members and products whose inclusion depended more strongly on the decision specification. Because portfolio size and objective weights altered the feasible space or evaluation function, absolute objective values were not compared as measures of cross-scenario superiority.

| Product ID | Optimization scenarios | | | | | Selected in n/5 scenarios |
|---|---|---|---|---|---|---|
| | Main-Balanced (K=8) | Size-Low (K=6) | Size-High (K=10) | Market-Oriented (K=8) | Culture-Oriented (K=8) | |
| C006 | ● | ● | ● | ● | ● | 5/5 |
| C024 | ● | ● | ● | ● | ● | 5/5 |
| C028 | ● | ● | ● | ● | ● | 5/5 |
| C022 | ● | ● | ● | ● | ○ | 4/5 |
| C033 | ● | ● | ● | ● | ○ | 4/5 |
| C026 | ● | ○ | ● | ○ | ● | 3/5 |
| C003 | ○ | ○ | ○ | ● | ● | 2/5 |
| C012 | ● | ○ | ● | ○ | ○ | 2/5 |
| C035 | ○ | ● | ○ | ● | ○ | 2/5 |
| C036 | ○ | ○ | ● | ○ | ● | 2/5 |
| C010 | ○ | ○ | ● | ○ | ○ | 1/5 |
| C014 | ○ | ○ | ○ | ● | ○ | 1/5 |
| C015 | ○ | ○ | ● | ○ | ○ | 1/5 |
| C016 | ○ | ○ | ○ | ○ | ● | 1/5 |
| C021 | ○ | ○ | ○ | ○ | ● | 1/5 |
| C034 | ● | ○ | ○ | ○ | ○ | 1/5 |

● Included in exact optimal portfolio ○ Not included

Note: Only products selected in at least one of the five exact optimal portfolios are shown. The remaining 20 products were never selected under any scenario.

Figure 4. Product membership across the five Exact optimal portfolio scenarios. Filled circles indicate products included in the corresponding Exact optimal portfolio; open circles indicate products not selected. The rightmost column shows the number of scenarios in which each displayed product was selected.

Label-sensitivity analysis produced a similar distinction between objective-value and portfolio-membership robustness. Replacing the prespecified PC/TP labels with independently validated labels changed the Exact optimal portfolio in all five scenarios, while the absolute relative change in the Exact objective remained below 0.4%. Jaccard overlap between Original- and Validated-label portfolios ranged from 0.455 to 0.714; under Main-Balanced, five of eight products were retained ($Jaccard = 0.455$). Classification-boundary changes therefore had limited effects on optimal objective values but more substantial effects on the specific products selected.

## 4. Discussion

This study combined product-level consumer choice, preference heterogeneity, design and cultural-resource characteristics, and constrained portfolio decision making to examine museum cultural and creative product development. The findings reveal three related patterns. Consumer preferences were clearly differentiated across specific product concepts and varied systematically across several consumer characteristics, yet only part of this heterogeneity could be represented by broader Product Concept strategies. Broader Product Concept and Translation Pathway classifications, as well as cultural-resource potential, did not consistently account for product-level consumer utility. At the portfolio level, consumer preference remained an important input but did not by itself determine the optimal product set once cultural and design requirements were considered. Together, these findings highlight the analytical distinction between product-level consumer evidence, broader design classifications, cultural-resource assessment, and portfolio-level decision requirements.

The findings first emphasize the importance of distinguishing specific product configurations from broader design classifications. Previous research has shown that functionality, cultural expression, visual design, and other product attributes can influence consumer evaluations and purchase judgments [36,37]. The present results extend this perspective by showing that consumer sensitivity to concrete product characteristics does not necessarily translate into stable preference advantages for broad design categories. Although some descriptive differences appeared across PC and TP categories, these effects were not robust after cultural-prototype differences were controlled, and the same overall inference was retained under independently validated labels. Consumer preferences therefore appeared to be more strongly differentiated among specific product configurations than between broad PC or TP categories. For museum product development, such

classifications may remain useful for organizing design strategies, but they should not substitute for direct consumer-choice evidence when specific alternatives are evaluated or prioritized.

Consumer heterogeneity further supports the need to retain product-level evidence. Purchase experience, collection orientation, practical-function importance, interaction importance, and professional background were associated with differences in relative product preferences, but these differences did not operate uniformly across analytical levels. Some heterogeneity could be summarized by broader PC strategies, particularly that associated with purchase experience, collection orientation, and practical-function importance, whereas the effects associated with interaction importance and professional background were more evident at the level of specific products. This cross-level divergence indicates that product-level and strategy-level heterogeneity are not interchangeable. Consumer segmentation based only on broad design categories may therefore obscure meaningful differences among products within the same category, particularly when museums must choose among concrete development alternatives.

The nonsignificant effect of cultural familiarity should be interpreted cautiously. Familiarity with Chinese bronze culture did not independently explain relative product preferences after the remaining consumer characteristics were considered, but this does not imply that cultural factors are irrelevant to cultural-product consumption. Previous research has linked cultural motivation, product identification, and perceived authenticity to purchase-related responses [38]. The present study measured general familiarity rather than constructs such as cultural identity, emotional attachment, or perceived authenticity. Cultural familiarity should therefore not be treated as a sufficient proxy for broader cultural-psychological engagement, and future research should distinguish these constructs when examining consumer responses to heritage-based products.

A similar distinction applies to cultural-resource characteristics. None of the six resource dimensions showed a consistent association with mean consumer utility across the 15 cultural prototypes after multiple-testing correction. Given the exploratory nature and limited prototype-level sample, this finding should not be interpreted as evidence that cultural-resource characteristics are unimportant. Rather, higher translation potential did not consistently correspond to higher consumer preference within the present prototype set. Cultural-resource assessment and consumer-choice evidence therefore served different functions: the former identified translatable cultural characteristics, preservation requirements, and interpretive risks, whereas the latter captured consumers' relative preferences among specific product concepts. Cultural-resource potential should consequently complement, rather than substitute for, product-level consumer-choice evidence.

The distinction between product-level preference and portfolio-level decision making became particularly clear in the optimization results. The Main-Balanced solution was not obtained by simply

selecting the highest-utility products because portfolio construction also incorporated cultural-pathway fit, cultural safety, design-strategy diversity, and explicit feasibility constraints. This demonstrates that individual-product preference and portfolio-level optimality are related but not equivalent: consumer-choice evidence can inform portfolio selection, but it represents only one component of a broader decision structure when museums seek to develop a coordinated set of products with different cultural, functional, and design roles.

Scenario and label sensitivity further indicate that portfolio optimization should be interpreted as conditional decision support rather than as a universally invariant ranking of "best" products [39]. Changes in portfolio size, objective weighting, and PC/TP classification boundaries altered portfolio membership, while some products remained stable across specifications. Importantly, independently validated labels produced only small changes in optimal objective values but more noticeable changes in the specific products selected. Objective-value robustness and membership robustness should therefore be treated as distinct properties. In practical applications, optimized portfolios are better interpreted as transparent, specification-dependent alternatives that make trade-offs explicit rather than as unique and irreplaceable solutions.

From a computational perspective, B-SBOA consistently improved upon the Repair-only baseline under matched initialization and feasibility-repair conditions, while the best run reached the Exact benchmark in all five scenarios. Because the present 36-product problem was also solvable to global optimality using the Exact method, B-SBOA should not be interpreted as a necessary replacement for exact optimization. Its contribution in the present study is better viewed as an Exact-validated constrained search procedure whose potential advantages may become more relevant for larger candidate spaces or more complex constraint structures. Future work should therefore evaluate its scalability under decision problems for which exhaustive or exact solution methods become computationally restrictive.

From a practical perspective, the findings support a staged approach to museum cultural and creative product development in which cultural evidence, consumer-choice evidence, and portfolio decision making remain interconnected but non-substitutable. Product development should remain grounded in traceable cultural prototypes, cultural anchors, and translation boundaries, while consumer evaluation should retain product-level choice evidence rather than relying solely on broad design categories. When museums seek to develop a coordinated product set rather than a single "winning" concept, trade-offs among consumer preference, cultural safety, design coverage, portfolio size, and feasibility should be made explicit. This approach provides a transparent basis for integrating cultural-resource information with consumer evidence in portfolio-level product planning.

Several limitations should be acknowledged. First, the study used on-site convenience sampling, with participants concentrated in younger age groups, and focused on Western Zhou bronze cultural

resources associated with the Baoji Bronze Ware Museum; generalizability to other populations, museums, heritage categories, and cultural contexts therefore requires further examination. Second, relative purchase preferences were elicited using standardized visual stimuli under an assumption of comparable and acceptable prices, which cannot fully substitute for physical products, actual prices, or revealed purchasing behavior [40]. Third, some consumer orientations were measured using single items, while the cultural-resource analysis involved only 15 prototypes and the PC/TP analyses 36 product concepts, limiting broader category-level inference. Finally, the portfolio model used prespecified objective weights, evaluation rules, and constraints. Future studies should examine larger candidate spaces, incorporate stakeholder-derived decision weights, and evaluate whether the optimization framework remains effective under more complex decision settings [41].

## 5. Conclusions

This study shows that consumer preferences for museum cultural and creative products are differentiated primarily at the level of specific product concepts and cannot be fully represented by broad design classifications or cultural-resource characteristics. Preference heterogeneity was evident across several consumer characteristics, but only part of it could be summarized by broader Product Concept strategies. Extending consumer-derived utility into constrained portfolio selection further showed that individual-product preference rankings are not equivalent to portfolio-level optimality when cultural and design requirements are considered simultaneously. Together, these findings support an analytical framework in which cultural-resource assessment, product-level consumer choice, preference heterogeneity, and portfolio decision support remain distinct but complementary components of museum cultural and creative product development.

### Data Availability

The participant-level datasets generated and analysed during the current study are not publicly available due to privacy and ethical restrictions associated with informed consent and institutional ethical approval. De-identified data from the visual-stimulus pretest, Best–Worst Scaling experiment, and consumer-characteristic measures are available from the corresponding author upon reasonable request, provided that the proposed use is consistent with the conditions of informed consent and applicable ethical requirements. Requests for data access should be directed to Ping Ji at jiping@sust.edu.cn.

### Acknowledgements

The authors would like to thank all participants who took part in the visual-stimulus pretest and the formal Best–Worst Scaling (BWS) study, as well as those who provided assistance during the implementation of the research.

### Author Contributions

Conceptualization, B.Z., P.J. and Q.J.; methodology, B.Z. and Q.J.; software, Q.J.; validation, B.Z., Y.G. and Q.J.; formal analysis, B.Z. and Q.J.; investigation, B.Z. and Y.G.; resources, P.J. and Q.J.; data curation, B.Z. and Y.G.; writing—original

draft preparation, B.Z.; writing—review and editing, P.J., Y.G. and Q.J.; visualization, B.Z. and Q.J.; supervision, P.J. and Q.J.; project administration, P.J. and Q.J.

Funding

This research received no external funding.

Competing interests

The authors declare no competing interests.

## References

1 Qiu, L., Rahman, A. R. A. & Dolah, M. S. b. The Role of Souvenirs in Enhancing Local Cultural Sustainability: A Systematic Literature Review. *Sustainability* **16**, 3893 (2024).

2 Xia, J., Li, Y., He, Y. & Ma, K. How museum-based creative-product experiences shape cultural confidence via cultural cognition and cultural identity. *Frontiers in Psychology* **Volume 16 - 2025**, doi:10.3389/fpsyg.2025.1742322 (2026).

3 Liu, L. & Zhao, H. Research on consumers' purchase intention of cultural and creative products—Metaphor design based on traditional cultural symbols. *PLOS ONE* **19**, e0301678, doi:10.1371/journal.pone.0301678 (2024).

4 Chen, C. *et al.* A hybrid method for assessing the perceived cultural value of cultural and creative products in e-commerce visual-textual presentation. *Scientific Reports*, doi:10.1038/s41598-026-58779-2 (2026).

5 Huang, H., Chen, H. & Zhan, Y. A Study on Consumers' Perceptions of Museum Cultural and Creative Products through Online Textual Reviews: An Example from Palace Museum's Cultural and Creative Flagship Store. *Behavioral Sciences* **13**, 318 (2023).

6 Cheng, H. *et al.* Constructing and validating the museum product creativity measurement (MPCM): dimensions for creativity assessment of souvenir products in Chinese urban historical museums. *Humanities and Social Sciences Communications* **11**, 280, doi:10.1057/s41599-024-02780-5 (2024).

7 Tu, J.-C., Liu, L.-X. & Cui, Y. A Study on Consumers' Preferences for the Palace Museum's Cultural and Creative Products from the Perspective of Cultural Sustainability. *Sustainability* **11**, 3502 (2019).

8 Li, Z., Shu, S., Shao, J., Booth, E. & Morrison, A. M. Innovative or Not? The Effects of Consumer Perceived Value on Purchase Intentions for the Palace Museum's Cultural and Creative Products. *Sustainability* **13**, 2412 (2021).

9 Gu, M. & Zhao, T. Research on the purchase intention of museum digital cultural and creative products based on value adoption model. *Scientific Reports* **15**, 18184, doi:10.1038/s41598-025-02140-6 (2025).

10 Louviere, J., Lings, I., Islam, T., Gudergan, S. & Flynn, T. An introduction to the application of (case 1) best–worst scaling in marketing research. *International Journal of Research in Marketing* **30**, 292–303, doi:https://doi.org/10.1016/j.ijresmar.2012.10.002 (2013).

11 Karim, S., Craig, B. M., Vass, C. & Groothuis-Oudshoorn, C. G. M. Current Practices for Accounting for Preference Heterogeneity in Health-Related Discrete Choice Experiments: A Systematic Review. *PharmacoEconomics* **40**, 943–956, doi:10.1007/s40273-022-01178-y (2022).

12 Dong, P. & Li, X. Cultural Identity and Value Perception as Drivers of Purchase Intention: A Structural Equation Model Analysis of Cultural Products in Luoyang City. *Sustainability* **17**, 1317 (2025).

13 Hou, L. *et al*. Sustainable design of museum bronze cultural-creative products through an integrated Kano-AHP-TOPSIS approach. *Discover Sustainability* **6**, 1212, doi:10.1007/s43621-025-02130-8 (2025).

14 Wang, Z., Zhou, J., Zhou, Z. & Li, F. An Integrated KANO–AHP–DEMATEL–VIKOR Framework for Sustainable Design Decision Evaluation of Museum Cultural and Creative Products. *Sustainability* **17**, 10328 (2025).

15 Salo, A., Doumpos, M., Liesiö, J. & Zopounidis, C. Fifty years of portfolio optimization. *European Journal of Operational Research* **318**, 1–18, doi:https://doi.org/10.1016/j.ejor.2023.12.031 (2024).

16 Xiang, Z.-R. *et al*. Cultural product design: an approach for identifying cultural carriers, transforming cultural elements and applying cultural features. *Design Science* **11**, e39, doi:10.1017/dsj.2025.10031 (2025).

17 Boxebeld, S. Ordering effects in discrete choice experiments: A systematic literature review across domains. *Journal of Choice Modelling* **51**, 100489, doi:https://doi.org/10.1016/j.jocm.2024.100489 (2024).

18 Jaeger, S. R., Andersen, G. B. H., Chheang, S. L. & Llobell, F. The ErrVarNorm (EVN) index in Case 1 Best–Worst Scaling: Further methodological and data quality investigations. *Food Quality and Preference* **141**, 105892, doi:https://doi.org/10.1016/j.foodqual.2026.105892 (2026).

19 Marley, A. A. J. & Louviere, J. J. Some probabilistic models of best, worst, and best–worst choices. *Journal of Mathematical Psychology* **49**, 464–480, doi:https://doi.org/10.1016/j.jmp.2005.05.003 (2005).

20 Colin Cameron, A. & Miller, D. L. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources* **50**, 317–372, doi:10.3368/jhr.50.2.317 (2015).

21 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x (1995).

22 Craig, B. M., de Bekker-Grob, E. W., González Sepúlveda, J. M. & Greene, W. H. A Guide to Observable Differences in Stated Preference Evidence. *The Patient - Patient-Centered Outcomes Research* **15**, 329–339, doi:10.1007/s40271-021-00551-x (2022).

23 MacKinnon, J. G. & White, H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**, 305–325, doi:https://doi.org/10.1016/0304-4076(85)90158-7 (1985).

24 Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* **70**, 213–220, doi:10.1037/h0026256 (1968).

25 de Raadt, A., Warrens, M. J., Bosker, R. J. & Kiers, H. A. L. A Comparison of Reliability Coefficients for Ordinal Rating Scales. *Journal of Classification* **38**, 519–543, doi:10.1007/s00357-021-09386-5 (2021).

26 Aksaraylı, M. & Pala, O. A polynomial goal programming model for portfolio optimization based on entropy and higher moments. *Expert Systems with Applications* **94**, 185–192, doi:https://doi.org/10.1016/j.eswa.2017.10.056 (2018).

27 Więckowski, J. & Sałabun, W. Supporting multi-criteria decision-making processes with unknown criteria weights. *Engineering Applications of Artificial Intelligence* **140**, 109699, doi:https://doi.org/10.1016/j.engappai.2024.109699 (2025).

28 Fu, Y., Liu, D., Chen, J. & He, L. Secretary bird optimization algorithm: a new metaheuristic for solving global optimization problems. *Artificial Intelligence Review* **57**, 123, doi:10.1007/s10462-024-10729-y (2024).

29 Crawford, B. *et al.* Binary Secretary Bird Optimization Algorithm for the Set Covering Problem. *Mathematics* **13**, 2482 (2025).

30 Deb, K. An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering* **186**, 311–338, doi:https://doi.org/10.1016/S0045-7825(99)00389-8 (2000).

31 Jourdan, L., Basseur, M. & Talbi, E. G. Hybridizing exact methods and metaheuristics: A taxonomy. *European Journal of Operational Research* **199**, 620–629, doi:https://doi.org/10.1016/j.ejor.2007.07.035 (2009).

32 LaTorre, A. *et al.* A prescription of methodological guidelines for comparing bio-inspired optimization algorithms. *Swarm and Evolutionary Computation* **67**, 100973, doi:10.1016/j.swevo.2021.100973 (2021).

33 Amin, A. A. H., Aladdin, A. M., Hasan, D. O., Mohammed-Taha, S. R. & Rashid, T. A. Enhancing Algorithm Selection through Comprehensive Performance Evaluation: Statistical Analysis of Stochastic Algorithms. *Computation* **11**, 231 (2023).

34 Demir, G., Chatterjee, P. & Pamucar, D. Sensitivity analysis in multi-criteria decision making: A state-of-the-art research perspective using bibliometric analysis. *Expert Systems with Applications* **237**, 121660, doi:https://doi.org/10.1016/j.eswa.2023.121660 (2024).

35 Darre, M. E., Domisch, S., Hermoso, V., Ørsted, M. & Arsanjani, J. J. Optimizing river restoration: A multi-scenario approach to barrier removal prioritization. *Journal of Environmental Management* **389**, 126176, doi:https://doi.org/10.1016/j.jenvman.2025.126176 (2025).

36 Liu, C., Samsudin, M. R. & Zou, Y. The Impact of Visual Elements of Packaging Design on Purchase Intention: Brand Experience as a Mediator in the Tea Bag Product Category. *Behavioral Sciences* **15**, 181 (2025).

37 Lin, F. & Ryu, K. How product design affects repurchase intention, eWOM, and museum visit intention: museum mystery boxes in China. *Journal of Travel & Tourism Marketing* **40**, 434–451, doi:10.1080/10548408.2023.2258154 (2023).

38 Li, X., Niu, Y. & Xu, J. Factors Affecting Purchase Intention of Hanfu: Considering Product Identification, Cultural Motivation, and Perceived Authenticity. *Behavioral Sciences* **13**, 689 (2023).

39 Gašperov, B., Begušić, S., Bauman, T. & Kostanjčar, Z. Quality-diversity and Novelty Search for Portfolio Optimization and Beyond. *Computational Economics* **67**, 2961–2984, doi:10.1007/s10614-025-10985-2 (2026).

40 Zhang, Y. *et al.* Prediction accuracy of discrete choice experiments in health-related research: a systematic review and meta-analysis. *EClinicalMedicine* **79**, 102965, doi:10.1016/j.eclinm.2024.102965 (2025).

41 Lienert, J. Real-world relevance of multicriteria decision analysis applied to environmental problems: A critical review of operational research literature. *EURO Journal on Decision Processes* **14**, 100069, doi:https://doi.org/10.1016/j.ejdp.2026.100069 (2026).

## Supplementary Files

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

- SupplementaryInformation.pdf