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.

Material–finish coupling in real-market products and evidence- labeled recommendation of surface attributes

Yuwei Zhang
University of Malaya
Chou Yong Tan
University of Malaya
Beichen Wang
University of Malaya
Yew Hoong Wong
yhwong@um.edu.my

University of Malaya

Article

Keywords: Kansei engineering, Color–material–finish (CMF), Surface finish, Semantic differential, Quantification theory type I, Small kitchen appliances

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11496252/v1

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

Additional Declarations: No competing interests reported.

Material–finish coupling in real-market products and evidence-labeled recommendation of surface attributes

Yuwei Zhang[1], Chou Yong Tan[1], Beichen Wang[2] & Yew Hoong Wong[1*]

[1]Department of Mechanical Engineering, Faculty of Engineering, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[2]Department of Urban & Regional Planning, Faculty of Built Environment, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[*]Corresponding author: Yew Hoong Wong (yhwong@um.edu.my)

Abstract

The appearance of a product depends on its base material and on how its surface is finished, and real products pair these two in particular ways. Kansei engineering studies of color, material and finish have not modeled this pairing. This study collected 180 small kitchen appliances from online retail and brand channels, and 167 consumers each rated 30 of them on 18 pairs of opposite adjectives, such as cheap–premium and light–heavy. Material and finish were closely associated (Cramér's $V = 0.586$), and 12 of the 28 combinations did not occur, so each observed material–finish pairing was modeled as one unit. Different surface attributes were associated with different impressions. Finish and decoration were associated with perceived quality, and a color print with lower quality, whereas transparency and base material were associated with perceived weight and solidity. In cross-validation these attributes predicted perceived quality and solidity above chance, but they did not predict how vivid or striking a product looked, which may depend on color. Based on these estimates, the study ranks combinations of surface attributes for a design goal and marks which parts of each top-ranked combination have statistical support, so that designers can see which choices the data back.

Keywords: Kansei engineering; Color–material–finish (CMF); Surface finish; Semantic differential; Quantification theory type I; Small kitchen appliances

Introduction

Kansei engineering translates consumers' feelings about products into design elements [1,2]. In a typical study, consumers rate products on Semantic Differential scales, factor analysis reduces the ratings to a few imagery dimensions, and a statistical model links design elements

to these dimensions [3]. The dimensions usually follow Osgood's Evaluation, Potency and Activity [4,5]. When the design elements are categorical, as surface attributes are, the usual model is Quantification Theory Type I (QTTI), which gives each category a score on each dimension [6,7].

A Kansei model can be used in two directions [8]. It can predict ratings from appearance attributes, and it can search for the combination of design elements that best matches a target feeling. The second use has been developed for two decades by pairing a mapping model with a search over the design space, in studies of personal digital assistants, electric bicycles and industrial equipment, among other products [9–13]. Neural networks have been used for product form since the early 2000s [14]. Later studies enlarged the search space and used more flexible prediction models [11–13], and QTTI is often fitted together with a back-propagation neural network [15,16]. The approach has been applied across many product domains [17]. Lin et al. compared QTTI, grey prediction and neural networks on this task and found QTTI good enough to determine the optimal form combination [10]. In most of this work the predictors describe form geometry.

Surface finish affects product judgments outside Kansei engineering. Gloss raises preference for both new and second-hand products [18]. Among problem-solving products, glossy finishes raised the perceived effectiveness of problem-avoiding products and matte finishes that of problem-approaching products [19]. Within Kansei engineering, Color–Material–Finish (CMF) has more recently been treated as a design dimension separate from form [20], with quantitative studies of automotive seats [21], cameras [22] and furniture [20]. Sun, for example, rendered 24 armchairs, tightened their classification in advance to limit category overload and multicollinearity, and had them rated by design professionals [20]. Rendered samples give experimental control, and an orthogonal design can balance the combinations of attributes. However, they cannot show how base materials and surface finishes co-occur in real products.

In the sampled products, some pairings of material and finish are rare or absent. None of the 22 glass or ceramic products had a brushed finish. Coupling between appearance attributes has also been recorded outside Kansei engineering. Kim et al. coded 770 chairs from a design-history collection by 14 colors and 11 materials and excluded 15 of the 154 possible color–material combinations as irrelevant, such as transparent wood [23]. If a study observes real products, orthogonal design can no longer remove the coupling. A model that treats material and finish as separate design items can still estimate their separate contributions. However, an exhaustive search over its combinations can return pairings that do not exist on the market.

Small samples also limit what such studies can estimate. With 24 rendered armchairs [20] or eight commercial cameras [22], a material-by-finish cross-tabulation is too sparse to estimate the association or to identify empty cells. This study uses 180 small kitchen appliances from open retail channels, rated on 18 Semantic Differential pairs by 167 ordinary consumers, the end users whose inclusion earlier work has called for [20]. This gives 90,180 ratings and about 28 raters per product. With this sample the cross-tabulation is dense enough to measure the coupling. The observed material–finish pairing can then serve as the unit of the design matrix, and sample support can limit the candidate space of a recommender.

This study reports three results. First, substrate (base material) and finish are strongly coupled in the sampled products, and a design matrix built from observed material–finish pairings excludes pairings absent from the sample and fits about as well as the element-wise matrix. Second, categorical surface attributes contribute to different imagery dimensions, here named Quality, Potency and Activity after Osgood's Evaluation, Potency and Activity. Decoration and primary finish carry Quality, and transparency and primary substrate carry Potency. The same split between finish and decoration on one side and transparency on the other appears in models of single word pairs, where ten of the eleven significant models fall into three groups by their dominant design item. Under repeated cross-validation the categorical coding predicts Quality and Potency above chance but not Activity. Third, a goal-driven recommender built on QTTI category scores labels every component of its output with its evidence status and reports no support for the dimension that the categorical coding did not predict. Figure 1a outlines the workflow.

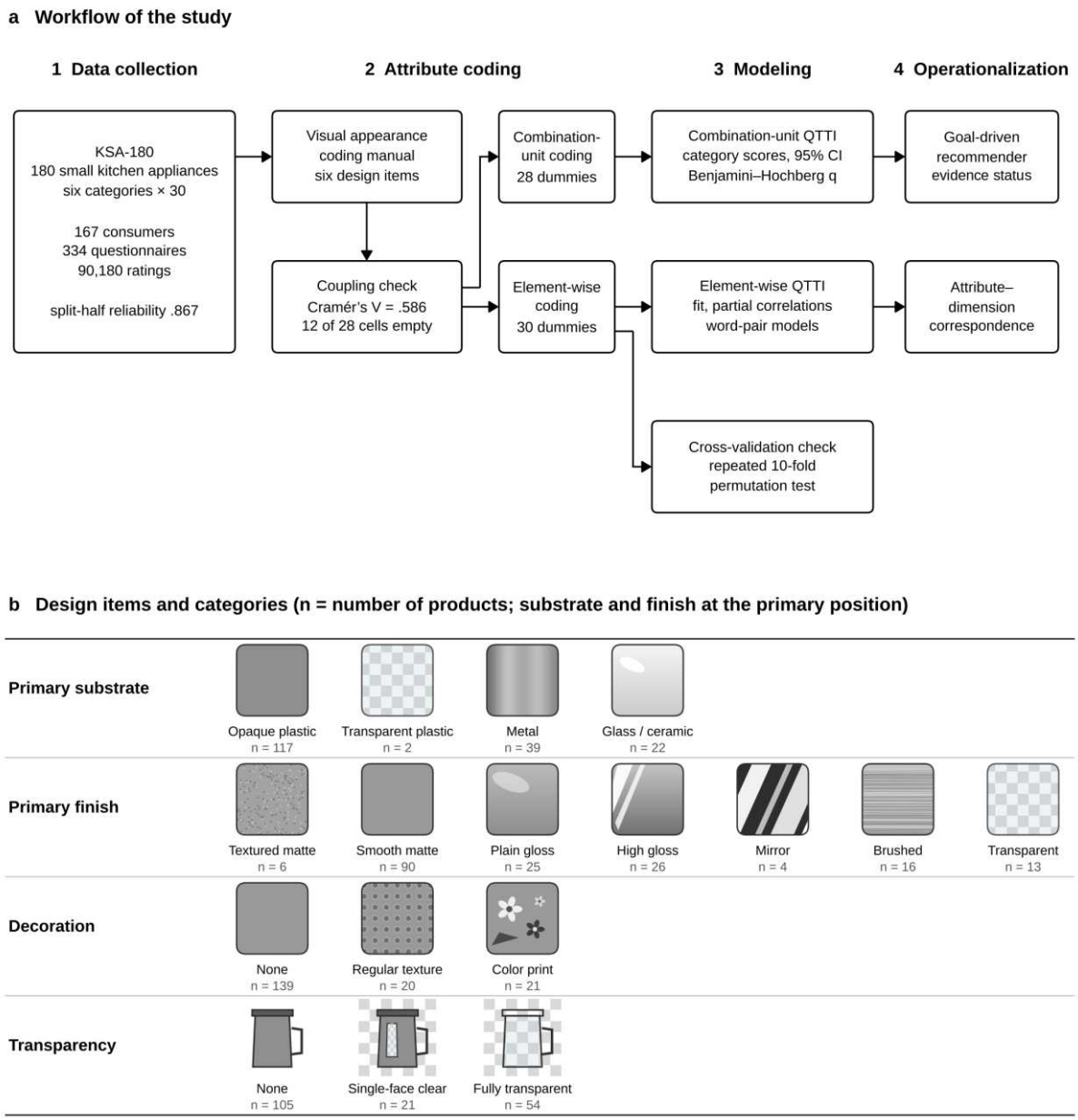


Figure 1. Workflow and coding scheme of the study. (a) Workflow. Ratings and appearance codes of the KSA-180 dataset feed two design matrices. The element-wise matrix gives the relative contribution of each design item and is used for the cross-validation check; the combination-unit matrix gives the category scores used by the recommender. (b) Design items and categories of the visual appearance coding manual, with the number of products in each category; substrate and finish counts refer to the primary position. The swatches are schematic.

Methods

Products and images

The unit of analysis is 180 small kitchen appliances from six categories of 30 products each (food preparation, portable aids, beverage making, baking, cooking and cleaning). The products were collected from JD.com, Taobao, Shopee and brand channels across price ranges and brand tiers so that appearance attributes vary. All images received the same preprocessing (background removal, logo masking, a standardized viewpoint, cropping and scaling to 800 × 800 pixels and, for some images, removal of stray reflections), so that attributes are comparable across images and brand information is removed. The products were not sampled to be representative of the market, and their appearance attributes were not manipulated. The dataset is denoted KSA-180 (examples in Fig. 2).



Figure 2. Examples of the product images used as stimuli (28 of the 180 products), shown after preprocessing. Images were collected from JD.com, Taobao, Shopee and brand channels, and brand logos were masked.

Participants and ethics

The analyzed sample comprises 167 consumers. Across the 334 questionnaires, respondents were aged 18–24 (21.9%), 25–34 (33.2%), 35–44 (33.5%) and 45 or older (11.4%), and 52.1% were men. Small appliances were used almost daily in 29.9% of questionnaires, several times a week in 31.4%, rarely in 26.3% and never in 12.3%. The study was approved by Universiti Malaya Research Ethics Committee (Reference Number: UM.TNC(P&I)/UMREC/2/5586). All methods were performed in accordance with the relevant guidelines and regulations, and informed consent was obtained from all participants.

Affective ratings

Ratings were collected with an online questionnaire on Wenjuanxin (www.wjx.cn) using the Semantic Differential method on a 7-point scale (−3 to +3). The 18 adjective pairs were drawn from adjectives used in earlier Kansei engineering studies and in consumer reviews of small kitchen appliances on JD.com, Taobao and Shopee, and were chosen so that evaluation, potency and activity were each represented [4,5]. Grouped by intended dimension (Quality, Potency and Activity), the pairs were cheap–premium, coarse–refined, ordinary–exquisite, cluttered–clean, abrupt–harmonious and harsh–comfortable (Quality); frail–robust, delicate–sturdy, flimsy–solid, soft–firm, light–heavy and rounded–angular (Potency); and plain–vivid, dull–intense, simple–colorful, muted–bright, understated–flamboyant and restrained–striking (Activity). The 180 products were divided into 12 questionnaires of 15 products, two for each category, holding the first and the second half of that category. Each questionnaire contained 273 items, namely 270 ratings (15 products × 18 pairs) and three items on age group, gender and frequency of small-appliance use. Each participant completed two questionnaires, in most cases the two halves of one category, and thus rated 30 products.

Two exclusion criteria were applied. Questionnaires completed in less than 3 s per item (819 s in total) were excluded, as were questionnaires that selected the same option across consecutive items (straightlining). After cleaning, every retained questionnaire had a within-response standard deviation of at least 0.76. The remaining 334 questionnaires from 167 participants gave 90,180 ratings, aggregated into 5010 product-rating records. Each product was rated by 24 to 32 participants (mean 27.8).

Reliability

Because the analyses use product mean ratings, the stability of these means was assessed by the intraclass correlation coefficient ICC(2,k) [24] and by the Spearman–Brown corrected split-half reliability [25,26], for each word pair and for each dimension score. ICC(2,k) was computed within each questionnaire, where the same raters rated all 15 products, and averaged over the 12 questionnaires. For the split-half reliability, the raters of each product were divided at random into two halves 500 times, and the Spearman–Brown corrected correlations between the half means were averaged. The split-half reliability of a dimension score is taken as the ceiling of its explainable variance and is used for attenuation correction.

Dimension structure

The 18 pairs were chosen to represent evaluation, potency and activity, and exploratory factor analysis of the 180 × 18 matrix of product mean ratings was used to check whether the ratings follow this intended structure. Suitability for factoring was checked by the Kaiser–Meyer–Olkin measure [27] and Bartlett's test [28], and the number of factors was set by parallel analysis against 1000 random normal matrices of the same size [29]. Factors were extracted by minimum residual estimation with promax rotation. The check covered whether each pair loaded highest on its intended factor, how many cross-loadings exceeded 0.25, and whether the assignment held under maximum likelihood and principal component extraction and under varimax, promax and oblimin rotation. Each dimension score is the mean of its six word pairs. Cronbach's α was computed on the product means and on the 5010 individual rating records. The structure was also examined by hierarchical clustering of the rating matrix (correlation distance, average linkage) and by an unrotated principal component analysis.

Appearance coding

Categorical coding followed a purpose-built visual appearance coding manual. For each category the manual gives an operational definition, visual criteria, rules for ambiguous cases and a hierarchical coding order, and it sets consistency checks between fields that prevent logically incompatible combinations. Six design items were coded (primary substrate, primary finish, secondary substrate, secondary finish, decoration and transparency; Fig. 1b), where substrate denotes the base material of a part. The manual defines 5 substrate categories and 9 finish categories. Four substrates appeared in the sample, transparent plastic in only two products, and seven finishes appeared at the primary position, where code A ('not applicable') cannot occur by definition. At the secondary position the same four substrates and seven finishes appeared, and products without a secondary part were coded 'none or not applicable' for substrate and A for finish. The 4 × 7 grid underlies the cross-tabulation of primary substrate and primary finish. Decoration was coded at three levels (none, regular texture and color print), and transparency likewise (none, single-face clear and fully transparent). Color was not coded, because hue, saturation and brightness are continuous quantities.

Coupling test and design matrices

The association between substrate and finish determines which design matrices are feasible. Cramér's V was therefore computed from their cross-tabulation, together with the number of empty cells and the rank of the full interaction matrix. Because many expected counts are small,

the p value of the χ^2 test was obtained by Monte Carlo simulation (20,000 permutations). If the association is strong and many cells are empty, that matrix is rank-deficient and main and interaction effects cannot be identified separately. Pearson residuals, $(O - E)/\sqrt{E}$, from the independence model show which cells are over- or under-represented. Secondary substrate and secondary finish were tested in the same way, and Cramér's V was also computed for every pair of the six design items. Two design matrices were then built (Fig. 1a).

In the element-wise matrix, the six design items were one-hot expanded into dummy variables, as in earlier QTTI studies. This matrix is used to report the relative contribution of each design item, which requires substrate and finish to enter as separate items. The six items have 4, 7, 5, 8, 3 and 3 observed levels, which gives 30 dummy variables and, with the intercept, 31 columns. The dummies of each item sum to the intercept column, which removes six dimensions, and the 'none or not applicable' level of secondary substrate coincides with code A of secondary finish ($n = 57$), which removes a seventh. The effective rank is therefore $31 - 7 = 24$, which leaves $k = 23$ free coefficients besides the intercept and $180 - 24 = 156$ residual degrees of freedom.

In the combination-unit matrix, primary substrate and primary finish were merged into one categorical variable whose values are their observed pairings, and secondary substrate and secondary finish were merged in the same way. Pairings that occurred fewer than three times were merged into a residual category. Decoration and transparency remained element-wise. This matrix estimates one effect per observed pairing and treats substrate and finish as one unit. Because the recommender enumerates pairings as units, it cannot add the coefficients of glass/ceramic and brushed to produce a scheme that is absent from the sample. The matrix has 28 dummy variables, namely 10 primary and 12 secondary pairing categories (each count includes the residual category) and three levels each for decoration and transparency. With the intercept it has 29 columns, and the dummies of each of the four items sum to the intercept column, so the effective rank is $29 - 4 = 25$, which leaves $k = 24$ free coefficients besides the intercept and 155 residual degrees of freedom. All 180 products entered the models. Categories that remain small after merging were retained, and their estimates are read through their confidence intervals.

Category scores are reported as deviations from the weighted mean of their design item. Each centered score is a linear combination of the regression coefficients, and its standard error and 95% confidence interval follow from their covariance matrix.

QTTI models and effect sizes

QTTI was estimated by least squares with category dummies as predictors and dimension scores as dependent variables [6]. Reported values are R, R^2, adjusted R^2, F and p, the partial correlation of each design item and the range of its category scores. The partial correlation of a design item is derived from the F test between the full model and the model without that item, and the range is the difference between its highest and lowest category scores. QTTI was fitted to both design matrices, and the element-wise model was also fitted to each of the 18 word pairs separately. The element-wise models give the relative contribution of each item and can be compared with the literature, and the combination-unit models drive the recommender. Earlier studies reduced the number of categories in advance [20]. Here the ratio of products to categories is 6.0 for the element-wise matrix (180 products, 30 categories) and 6.4 for the combination-unit matrix (28 categories), and the predictive validity of the element-wise models was checked by repeated cross-validation.

Effect size is computed as

$$f^2 = R^2 / (1 - R^2) \tag{1}$$

following Cohen [30]. Because the dependent variable is a product mean, explanatory power was also corrected for attenuation [31]. The corrected value is R^2/ρ, where ρ is the split-half reliability of that dimension's mean ratings, and it is reported next to the raw value. The correction concerns the unreliability of the ratings; the categorical predictors are treated as measured without error. Adjusted R^2 is reported as the conservative estimate of explained variance.

All tests were two-sided at $\alpha = .05$. A category score is judged significant when its 95% confidence interval excludes zero. Because the three dimensions together involve 84 contrasts (28 categories in the combination-unit matrix × 3 dimensions), Benjamini–Hochberg false-discovery-rate q values are also reported [32], and the same correction is applied across the 18 word-pair models.

One-way analyses of variance

Whether the material–finish pairing itself is associated with the ratings was tested by a one-way analysis of variance (ANOVA) over the pairings with at least five products, because the full interaction matrix cannot separate main and interaction effects. One-way ANOVAs with η^2 also compared the categories of single design items on each dimension without a linear

model. Primary finish was compared within opaque plastic so that the substrate is held constant, and primary substrate was compared across the substrates with at least 20 products.

Cross-validation

Out-of-sample predictive validity was assessed by 10-fold cross-validation repeated ten times, with folds stratified by product category (three products per category in each fold). Standardization and any penalty selection were fitted within the training folds. In each repetition, the out-of-fold predictions for all 180 products were pooled to compute the cross-validated R^2 and the correlation r between predicted and observed means, and these values were averaged over repetitions. The additive category model was fitted with and without ridge shrinkage, with the ridge penalty chosen by leave-one-out cross-validation within each training fold. Four further models were fitted to the same element-wise design matrix: a back-propagation neural network (one hidden layer of 10 ReLU units, L-BFGS solver, L2 penalty chosen between 10^{-4} and 30 by inner 5-fold cross-validation), support vector regression (RBF kernel, C = 10, ε = 0.05), random forest (500 trees) and gradient boosting (300 rounds, maximum depth 3, learning rate 0.05). Whether the cross-validated R^2 exceeded chance was tested by permuting the ratings across products 100 times and repeating the cross-validation on one fixed partition, with p = (1 + number of permuted values ≥ observed value)/101. The random seed was 42.

Goal-driven recommender

The recommender uses the combination-unit category scores as its scoring function. QTTI gives each design category a score whose sign and size show its contribution, so candidate combinations can be ranked by adding scores. A design goal is expressed as weights on the imagery dimensions (for example Quality +1 and Potency −1), and each candidate is scored as

$$\text{Target score} = \Sigma_{dim}\, w_{dim} \times \Sigma_{item}\, \text{CS}_{dim}(\text{item, category}) \qquad (2)$$

where CS is the category score on that dimension and w is the target weight. Candidates are ranked by target score.

The recommender enumerates only attribute levels observed in this sample. Primary and secondary material–finish pairings are limited to those found in at least five products, and the residual category is excluded. The design matrix gives a separate category to every pairing with at least three products so that each receives an estimate. The recommender, like the pairing ANOVA, uses a limit of five products because the estimates of pairings with three or four

products have wide intervals. Each component of a scheme (primary pairing, secondary pairing, decoration and transparency) is labeled with its evidence status, that is, whether its centered score excludes zero at the 95% level on any dimension in the target. Readers can then see which parts of a scheme have statistical support and which rest on model inference. Apart from 'no secondary material', the secondary pairings have small samples, so their status is reported but not counted as support even when the interval excludes zero. Stability was assessed by resampling the products 1000 times with replacement within each category, re-estimating the category scores and recording the top choice for each component.

Four design goals demonstrate the output: maximize Quality (Quality +1); high quality, light (Quality +1, Potency −1); high quality, calm (Quality +1, Activity −1); and expressive and light (Activity +1, Potency −1). The fourth goal includes Activity, and its evidence status is reported with the others.

Software

All analyses were run in Python 3.13 with NumPy 2.1, SciPy 1.16, pandas 2.2, scikit-learn 1.6 and factor_analyzer 0.5.

Results

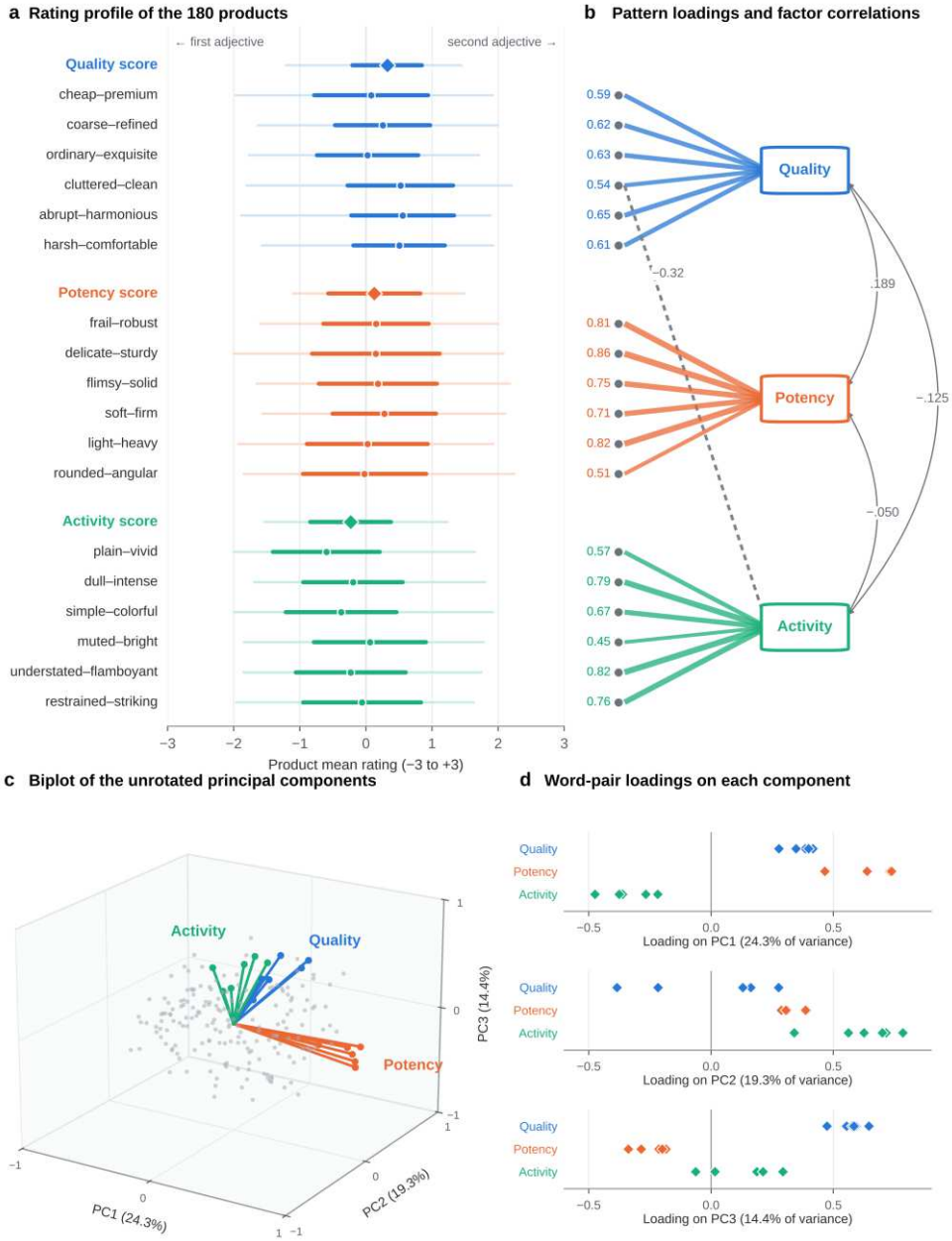
Rating reliability and dimensional structure

At the word-pair level, ICC(2,k) ranged from 0.809 to 0.900 (Supplementary Table S1), and the Spearman–Brown split-half reliability gave nearly the same values (mean 0.867), so about 87% of the variance of a typical word-pair mean is reproducible signal. The dimension scores reach split-half reliabilities of 0.834 (Quality), 0.906 (Potency) and 0.880 (Activity). Figure 3a shows the distribution of the product means for the 18 word pairs and the three dimension scores (descriptive statistics in Supplementary Table S2).

The matrix of product means was suitable for factoring (Kaiser–Meyer–Olkin measure 0.822, item values 0.737 to 0.873; Bartlett's test $\chi^2(153) = 1427.7$, $p < .001$). Parallel analysis retained three factors, whose eigenvalues (4.38, 3.47 and 2.59) exceeded the 95th percentiles of the random eigenvalues (1.71, 1.55 and 1.46); the fourth eigenvalue (1.10) fell below its random counterpart (1.37). The three factors accounted for 49.9% of the variance. Every pair loaded highest on its intended factor (primary loadings 0.45 to 0.86), so each factor collected the six pairs chosen for it. One cross-loading exceeded 0.25 (cluttered–clean, −0.32 on Activity; Fig.

3b; Supplementary Table S3), and each pair's share of its squared loadings on its own factor averaged 0.93 (minimum 0.72, cluttered–clean). The same assignment held under the other extraction and rotation methods. Cronbach's α was .772, .880 and .831 on the product means and .807, .827 and .807 on the individual rating records, and the factor correlations were .189 (Quality–Potency), −.125 (Quality–Activity) and −.050 (Potency–Activity).

Two further analyses give a consistent structure. Hierarchical clustering of the 180 × 18 rating matrix (correlation distance, average linkage) cut into three groups reproduced the factor assignment exactly, with six word pairs in each group. In an unrotated principal component analysis of the same matrix, the first three components explained 58.0% of the variance (24.3%, 19.3% and 14.4%; Fig. 3c,d). The Potency pairs loaded on average 0.67 on the first component, the Activity pairs 0.62 on the second and the Quality pairs 0.57 on the third, and 17 of the 18 pairs loaded most strongly on the component of their own factor (the exception is muted–bright). The Quality pairs also loaded on the first component (mean 0.37), and the Activity pairs loaded negatively on it (mean −0.34). The three dimension scores were used as dependent variables.



316

317 **Figure 3.** Structure of the rating matrix. (a) Product mean ratings of the 18 word pairs and the
318 three dimension scores (n = 180). Points are means, thick bars ±1 SD and thin lines the range;
319 diamonds mark the dimension scores, and positive values lie toward the second adjective. (b)
320 Pattern loadings of the three-factor solution (minimum residual extraction, promax rotation).
321 Line width is proportional to the loading, loadings below 0.25 are omitted, and the dashed line
322 marks the one cross-loading above 0.25. Arcs give the factor correlations. (c) Biplot from an
323 unrotated principal component analysis of the standardized 180 × 18 rating matrix. Gray points
324 are products, with scores rescaled to the range of the axes, and lines are word-pair loadings
325 colored by factor. (d) Loadings of the 18 word pairs on each of the first three components,

which together explain 58.0% of the variance.

Coupling between substrate and finish

The primary substrate and the primary finish are strongly associated (Fig. 4; Cramér's V = 0.586, $\chi^2(18, N = 180) = 185.1$, Monte Carlo $p < .001$). Twelve of the 28 possible combinations are empty, so only 16 of the 28 columns of the full interaction matrix are estimable and 12 parameters cannot be identified. Of the 16 pairings that occur, six have at least five products. The largest, opaque plastic with smooth matte, accounts for 66 products, or 37% of the sample. Co-occurrence therefore concentrates in a few combinations.

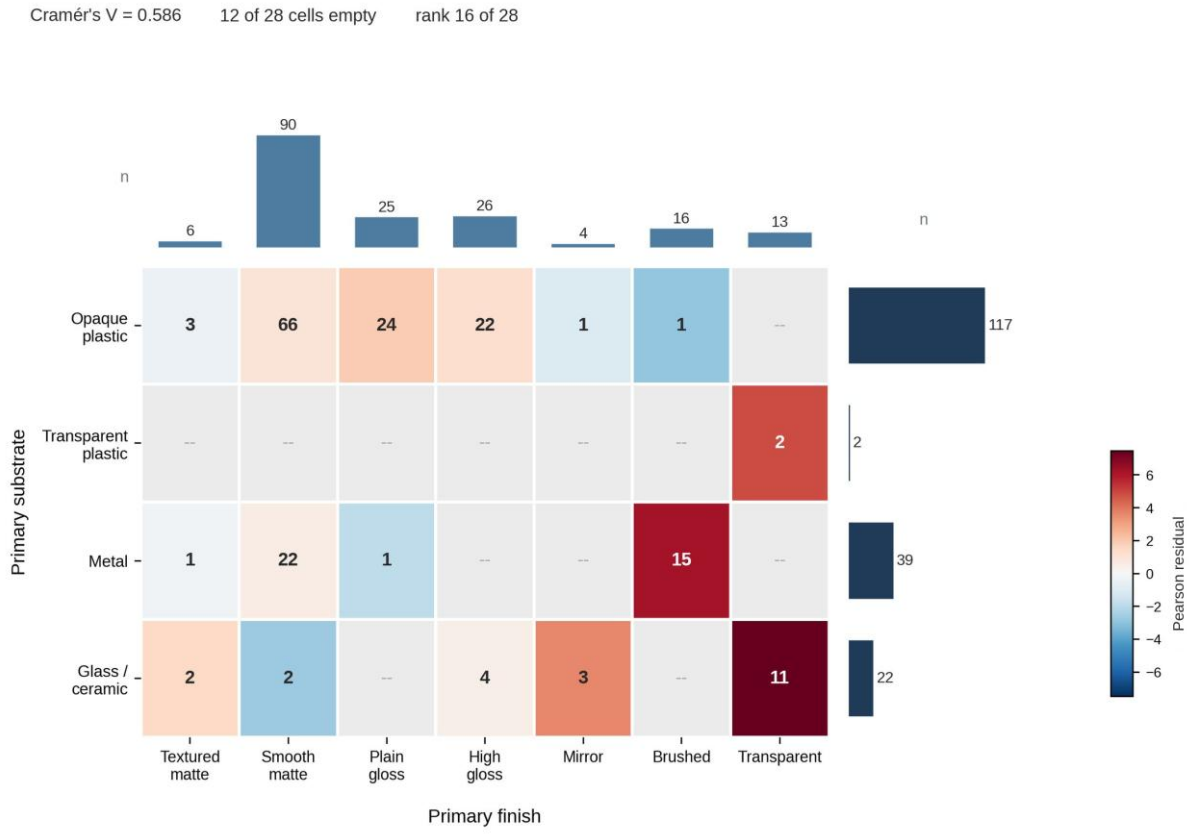


Figure 4. Co-occurrence of primary substrate and primary finish. Cell values are product counts; color shows the Pearson residual, $(O - E)/\sqrt{E}$, from the independence model, and gray cells mark combinations absent from the sample. Marginal bars give totals. Finish code A (not applicable) takes no value at the primary position. $\chi^2(18, N = 180) = 185.1$, Monte Carlo $p < .001$; Cramér's $V = 0.586$; the full interaction matrix has rank 16 of 28.

Pearson residuals from the independence model show that the coupling also takes the form of absence. Glass/ceramic with transparent finish (n = 11) reaches +7.5 and metal with brushed finish (n = 15) reaches +6.2. Opaque plastic received a brushed finish in one product, against

10.4 expected under independence (residual −2.9).

Among all pairs of the six design items, the strongest association is between secondary substrate and secondary finish ($V = 0.727$). Transparency is moderately to strongly associated with secondary finish ($V = 0.566$) and with secondary substrate ($V = 0.523$). The 'none or not applicable' level of secondary substrate and the 'not applicable' level of secondary finish coincide exactly (both $n = 57$). Decoration is the least associated item ($V \leq 0.260$), so its estimates are the least confounded.

An element-wise design matrix allows the coefficients of any substrate and any finish to be added, and with this coupling many of the resulting pairings never occur. For the secondary material it also admits 'no secondary material with a secondary finish'. The combination-unit matrix (Methods) excludes such combinations.

Because main and interaction effects cannot be separated in the full interaction matrix, whether the pairing itself is associated with the ratings was tested by a one-way ANOVA over the six pairings with at least five products (160 products). The largest effect is on cheap–premium ($F(5, 154) = 8.68$, $p < .001$, $\eta^2 = 0.220$). At the composite level the effects are smaller (Potency $\eta^2 = 0.118$, $p = .002$; Quality $\eta^2 = 0.081$, $p = .021$), and the effect on Activity is not significant ($p = .883$). Among the six Quality pairs, η^2 ranges from 0.016 (harsh–comfortable) to 0.220 (cheap–premium), and only cheap–premium exceeds 0.10.

Fit of the categorical mapping

Supplementary Table S4 reports the element-wise QTTI models. The Quality and Potency models are significant ($F(23, 156) = 2.69$ and 2.89, $p < .001$), with $f^2 = 0.397$ and 0.426. Their adjusted R^2 values are 0.179 and 0.195. After correction for rating reliability, R^2 is 0.341 for Quality and 0.330 for Potency ($f^2 = 0.517$ and 0.492). The Activity model is not significant ($F(23, 156) = 0.89$, $p = .616$; $f^2 = 0.131$). The combination-unit models, which drive the recommender, fit about as closely. Their multiple correlations are 0.507, 0.550 and 0.321 for Quality, Potency and Activity (adjusted $R^2 = 0.142$, 0.194 and −0.036), against 0.533, 0.547 and 0.340 for the element-wise models.

Under repeated 10-fold cross-validation, the cross-validated R^2 of the same additive model exceeded all 100 permuted values for Quality and Potency ($p = .010$, the smallest value attainable with 100 permutations; Supplementary Table S5), with cross-validated $r = 0.30$ and 0.33. With ridge shrinkage of the category coefficients, r was 0.30 and 0.35 ($p = .010$ for both). For Activity, the cross-validated R^2 did not exceed chance ($p = .57$ without shrinkage and p

= .33 with shrinkage). Four higher-capacity models fitted to the same predictors (a back-propagation neural network, support vector regression, random forest and gradient boosting) had lower mean cross-validated R^2 than the ridge-regularized additive model on Quality and Potency in this sample (Supplementary Fig. S1; Supplementary Table S6).

Design items and imagery dimensions

The partial correlations (Supplementary Table S7, Fig. 5a) rank decoration (0.360) and primary finish (0.324) highest for Quality and primary substrate lowest (0.065). For Potency, transparency (0.338) and primary substrate (0.287) rank highest, followed by secondary finish, primary finish and secondary substrate, and decoration ranks lowest (0.043). The category-score range of primary substrate is 0.251 on Quality (4.2% of the scale span) and 1.324 on Potency (22.1%). Categorical surface attributes therefore contribute to specific dimensions.

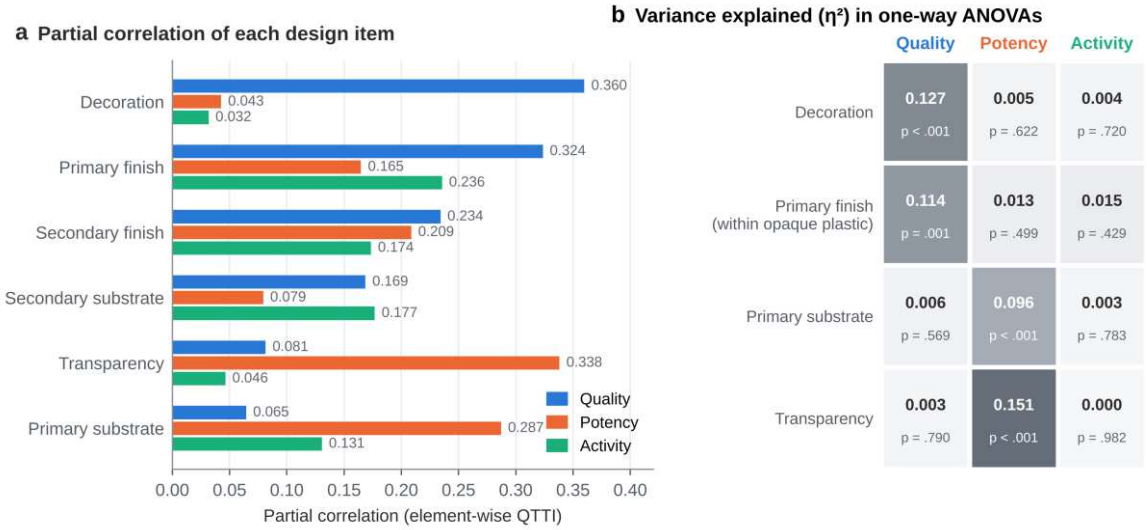


Figure 5. Contribution of the design items to the three imagery dimensions. (a) Partial correlation of each design item in the element-wise QTTI models. Decoration and primary finish rank highest for Quality, and transparency and primary substrate rank highest for Potency. (b) Proportion of variance explained (η^2) by four design items in one-way ANOVAs, with p values (Supplementary Table S8); darker cells mark larger η^2. Primary finish is compared within opaque plastic.

The same pattern appears without a linear model. Within opaque plastic, which holds the substrate constant and so removes the coupling, the three main finishes (smooth matte, plain gloss and high gloss; 112 products) differ in Quality ($\eta^2 = 0.114$, $p = .001$; spread of group means 0.461) but not in Potency ($\eta^2 = 0.013$, $p = .50$) or Activity ($\eta^2 = 0.015$, $p = .43$). Supplementary Table S8 and Fig. 5b extend this comparison to four design items. Each item

explains more than seven times as much variance on its dominant dimension as on either of the other two, and no value in the Activity column exceeds 0.015.

Eleven word-pair models are significant, including five of the six Quality pairs, all six Potency pairs and none of the Activity pairs. After Benjamini–Hochberg correction across the 18 models, ten remain significant, and soft–firm (p = .040) does not. Cheap–premium has the highest f^2 (0.706). Across all 18 pairs, multiple correlations range from 0.309 to 0.643 (median 0.456).

Supplementary Table S9 and Supplementary Fig. S2 give the dominant design item of each significant model, that is, the item with the highest partial correlation. Ten of the eleven significant models fall into three groups by dominant item. Primary finish dominates cheap–premium, coarse–refined and ordinary–exquisite, which concern worth and workmanship. Decoration dominates cluttered–clean and abrupt–harmonious, which concern visual order. Transparency dominates frail–robust, delicate–sturdy, flimsy–solid, soft–firm and light–heavy, which concern physical substance and are five of the six Potency pairs. The remaining significant model, rounded–angular, is dominated by secondary finish. Within Quality, the pairs on worth and workmanship and the pairs on visual order have different dominant items.

For cheap–premium (R = 0.643, R^2 = 0.414, adjusted R^2 = 0.327, $p < .001$), the partial correlation of primary finish is 0.487 and its category scores span 1.787, or 29.8% of the scale span. For rounded–angular, the partial correlation of secondary finish is 0.286 and its range 1.292. The range comes mainly from the eight products with a mirror-finish secondary part (+1.094); the other secondary finishes with more than two products lie between −0.198 and +0.283.

Category scores

Supplementary Table S10 and Fig. 6 give the centered category scores with 95% confidence intervals. Of the 84 contrasts, 10 have intervals that exclude zero (3 for Quality, 6 for Potency and 1 for Activity), and 3 have $q < .05$ after Benjamini–Hochberg correction. Color print contributes negatively to Quality (−0.476, q = .001). No decoration contributes slightly positively to Quality (+0.075, q = .018); the effect is small, but with 139 products in this category its standard error is 0.021. Fully transparent parts contribute negatively to Potency (−0.479, q = .003).

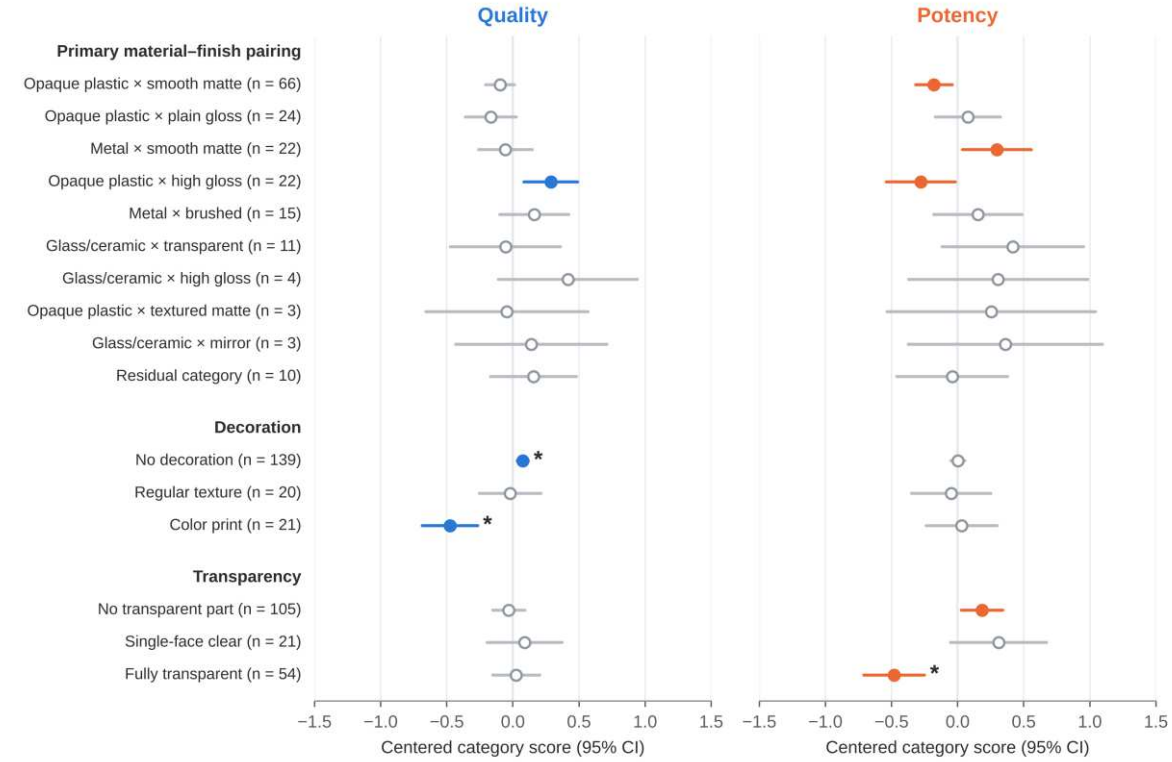


Figure 6. Centered category scores for Quality and Potency with 95% confidence intervals (combination-unit QTTI). Filled markers indicate intervals that exclude zero, and asterisks indicate $q < .05$ after Benjamini–Hochberg correction across 84 contrasts. Secondary pairings are not plotted; the one secondary pairing whose interval excludes zero on Quality or Potency is listed in Supplementary Table S10. No Activity contrast survives correction.

Decoration accounts for two of these three results, and its effect on Quality has two levels. A regular texture scored close to the item mean on Quality (−0.021, $p = .861$), whereas color print contributed negatively (−0.476). Opaque plastic with high gloss contributes positively to Quality (+0.287, $p = .006$, $q = .128$) and is the only plastic pairing whose interval lies above zero on Quality. Together with the color-print effect, this is consistent with Quality depending more on finish and decoration than on substrate.

Activity

For Activity, the composite models are not significant (element-wise $p = .616$, combination-unit $p = .803$), and the cross-validated predictions are not better than chance. None of the six word-pair models reaches significance ($p = .125$ to $.825$). The single Activity interval that excludes zero (secondary opaque plastic × smooth matte, $p = .036$) has $q = .316$ after correction, so no centered contrast survives correction. The categorical coding therefore did not predict Activity.

Recommender output

The recommender ranked all feasible combinations in the candidate space for four design goals. Table 1 and Fig. 7 give the top-ranked scheme for each goal with the evidence status of every component. For goals that include Quality, it selects no decoration, and for goals that lower Potency, it selects fully transparent parts. These choices follow the corrected effects in Supplementary Table S10.

Table 1. Recommender output for the four design goals.

Goal	Target weights	Primary pairing	Secondary pairing	Decoration	Transparency	Target score	Evidence
Maximize Quality	Quality +1	Opaque plastic × high gloss	Metal × smooth matte	None	Single-face clear	0.629	Y / n / Y / n
High quality, light	Quality +1, Potency −1	Opaque plastic × high gloss	No secondary material	None	Fully transparent	1.284	Y / n / Y / Y
High quality, calm	Quality +1, Activity −1	Metal × brushed	Metal × brushed	None	None	0.883	n / n / Y / n
Expressive and light	Activity +1, Potency −1	Opaque plastic × high gloss	Opaque plastic × smooth matte	Regular texture	Fully transparent	1.547	Y / n* / n / Y

The four evidence symbols refer to primary pairing, secondary pairing, decoration and transparency. Y, the 95% interval excludes zero on at least one dimension in the target; n, the interval includes zero and the point estimate has the sign the goal favors; n*, a secondary pairing whose interval excludes zero, which is not counted as supported because secondary pairings have small samples (Methods). Target scores are weighted sums of category scores and are not comparable across goals.

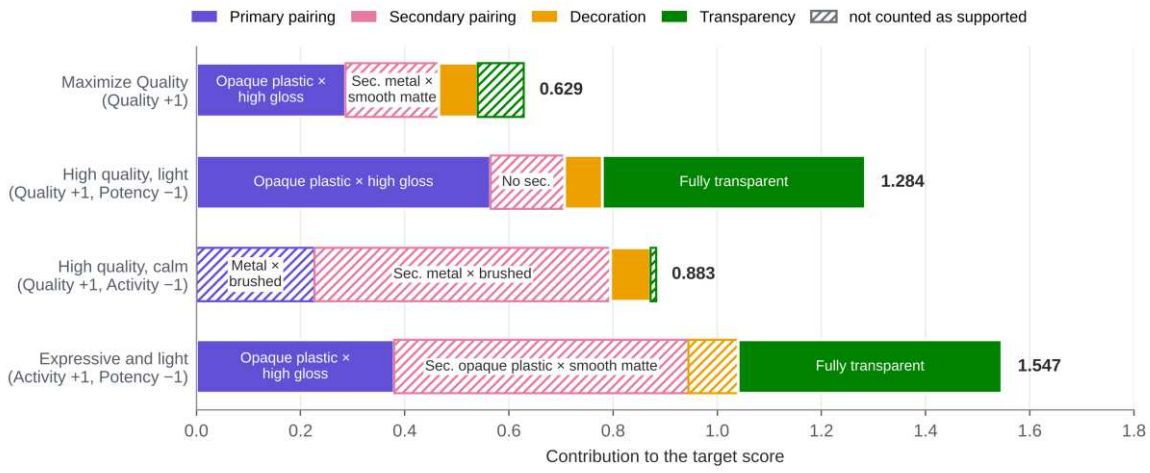


Figure 7. Contribution of each design item to the target score of the recommended scheme for each of the four design goals. Solid segments mark components counted as supported (Table 1); hatched segments are not counted as supported. Target scores are not comparable across goals because the weights differ.

The four schemes have 16 component positions. Among the 12 positions of the three main design items, 8 have support. Five of these rest on corrected effects (no decoration on Quality,

q = .018, in three schemes; fully transparent parts on Potency, q = .003, in two schemes). The other three rest on opaque plastic with high gloss, with q = .128 on Quality in the first two schemes and q = .316 on Potency in the fourth. The four secondary-pairing positions are not counted (Methods). Across 1000 bootstrap resamples, supported components remained the top choice in 65% to 100% of cases (median 80%) and unsupported components in 27% to 78% (median 51%). The complete four-component scheme recurred in 7% to 32% of resamples.

The fourth goal, expressive and light, combines Activity with low Potency. Its two supported components (opaque plastic with high gloss and fully transparent parts) take their support from Potency. Along Activity, the only interval that excludes zero before correction has q = .316. The scheme therefore has no supported component in the direction of Activity, which matches the result that the categorical coding did not predict this dimension.

Discussion

In the sampled products, substrate and finish were strongly coupled, and 12 of the 28 combinations were absent. Sun avoided the coupling by reducing the categories in advance and using rendered samples [20]. For observational samples, the combination unit uses only observed material–finish pairings, with a fit similar to that of the element-wise matrix, while combinations across design items are still built additively. Combinations of appearance attributes have served as descriptive units before, as in the color–material combinations of chairs [23]. In the Kansei studies we reviewed, we did not find the material–finish pairing used as the unit of a design matrix for effect estimation.

Earlier work has shown that surface finish affects product evaluation [18,19]. With six design items in one model, this study estimates which attributes are associated with which imagery dimensions. In this sample, perceived value was more strongly associated with finish and decoration than with substrate, and perceived weight and solidity were more strongly associated with transparency and substrate. The contribution of the substrate to Quality is small both with finish and decoration controlled (partial correlation 0.065) and without control (η^2 = 0.006; Supplementary Table S8). The two levels of decoration differ on Quality, possibly because a texture treats the surface itself, whereas a print adds content on top of it.

Under cross-validation, the higher-capacity models did not predict better than the additive model with ridge shrinkage. This agrees with the finding of Lin et al. that QTTI is good enough for optimal form design [10], and it means that the additive structure behind the readable

category scores does not give up predictive accuracy relative to the more flexible models in this sample and with these settings. The cross-validated correlations of 0.30 to 0.35 indicate that the categorical coding captures part of the variation in Quality and Potency ratings. The categorical coding did not predict Activity. Three of its word pairs name chromatic properties, which suggests continuous color variables as a candidate source.

The optimal-combination studies cited in the Introduction report one solution per goal without labeling its parts [9–13]. The recommender reports such a solution together with the evidence status of each component, and supported components were also the more stable choices under resampling. Apart from ‘no secondary material’, the secondary pairings have small samples, so the data leave the choice of secondary material open.

The sample consists of small kitchen appliances shown as static front views, and because the meaning of a material depends on context [33], the correspondences reported here apply to this product domain. Logos were masked, so brand effects are outside the data, and differences in image production between retail sources may remain in how surfaces are presented. Product form was not among the coded design items, so finish effects on form-related pairs such as rounded–angular may share variance with form. Most participants rated products from a single category, so rater and category effects overlap. Pairings with few products cannot ground a recommendation. Glass/ceramic with high gloss ($n = 4$) has a Quality estimate of +0.416, but its 95% interval of −0.081 to 0.912 spans about one sixth of the scale. The combination unit handles coupling within the primary and the secondary materials but keeps additivity across design items, so the best secondary configuration is the same for every primary pairing. The recommender was checked against the same sample from which its scores were estimated.

With real-market products and ordinary-consumer ratings, this study estimated which surface attributes are associated with which imagery dimensions and turned these estimates into recommendations that label each component with its evidence status.

Data availability

The datasets analyzed in this study (individual ratings, product-level mean ratings, appearance codes and analysis code) are not publicly available because they are being used in ongoing doctoral research. They are available from the corresponding author on reasonable request. Summary statistics supporting the findings are provided in the Supplementary Information.

References

1. Nagamachi, M. Kansei engineering: a new ergonomic consumer-oriented technology for product development. *Int. J. Ind. Ergon.* **15,** 3–11. https://doi.org/10.1016/0169-8141(94)00052-5 (1995).

2. Nagamachi, M. (ed.) *Kansei/Affective Engineering* (CRC Press, 2011). https://doi.org/10.1201/ebk1439821336.

3. Schütte, S. T. W., Eklund, J., Axelsson, J. R. C. & Nagamachi, M. Concepts, methods and tools in Kansei engineering. *Theor. Issues Ergon. Sci.* **5,** 214–231. https://doi.org/10.1080/1463922021000049980 (2004).

4. Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. *The Measurement of Meaning* (University of Illinois Press, 1957).

5. Osgood, C. E., May, W. H. & Miron, M. S. *Cross-Cultural Universals of Affective Meaning* (University of Illinois Press, 1975).

6. Hayashi, C. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Ann. Inst. Stat. Math.* **3,** 69–98. https://doi.org/10.1007/bf02949778 (1951).

7. Jindo, T. & Hirasago, K. Application studies to car interior of Kansei engineering. *Int. J. Ind. Ergon.* **19,** 105–114. https://doi.org/10.1016/s0169-8141(96)00007-8 (1997).

8. Matsubara, Y. & Nagamachi, M. Hybrid Kansei engineering system and design support. *Int. J. Ind. Ergon.* **19,** 81–92. https://doi.org/10.1016/s0169-8141(96)00005-4 (1997).

9. Lai, H.-H., Lin, Y.-C., Yeh, C.-H. & Wei, C.-H. User-oriented design for the optimal combination on product design. *Int. J. Prod. Econ.* **100,** 253–267. https://doi.org/10.1016/j.ijpe.2004.11.005 (2006).

10. Lin, Y.-C., Yeh, C.-H., Wang, C.-C. & Wei, C.-C. Is the linear modeling technique good enough for optimal form design? A comparison of quantitative analysis models. *Sci. World J.* **2012,** 689842. https://doi.org/10.1100/2012/689842 (2012).

11. Hsiao, S.-W., Chiu, F.-Y. & Lu, S.-H. Product-form design model based on genetic algorithms. *Int. J. Ind. Ergon.* **40,** 237–246. https://doi.org/10.1016/j.ergon.2010.01.009 (2010).

12. Wang, T., Xu, M., Yang, L., Zhou, M. & Sun, X. Constructing a MOEA approach for product form Kansei design based on text mining and BPNN. *J. Intell. Fuzzy Syst.* **46,** 8865–8885. https://doi.org/10.3233/jifs-230668 (2024).

13. Song, S., Yue, J. & Yang, X. Kansei design optimization of torque tool inspection cabinets using XGBoost prediction models. *Appl. Sci.* **16,** 3884. https://doi.org/10.3390/app16083884 (2026).

14. Hsiao, S.-W. & Huang, H. C. A neural network based approach for product form design. *Des. Stud.* **23,** 67–84. https://doi.org/10.1016/s0142-694x(01)00015-1 (2002).

15. Lin, Z.-H., Woo, J.-C., Luo, F. & Chen, Y.-T. Research on sound imagery of electric shavers based on Kansei engineering and multiple artificial neural networks. *Appl. Sci.* **12,** 10329. https://doi.org/10.3390/app122010329 (2022).

16. Lin, Z.-H., Woo, J.-C., Luo, F. & Pan, G.-Q. Multisensory design of electric shavers based on Kansei engineering and artificial neural networks. *Math. Probl. Eng.* **2023,** 1188537.

https://doi.org/10.1155/2023/1188537 (2023).

17. López, Ó., Murillo, C. & González, A. Systematic literature reviews in Kansei engineering for product design: a comparative study from 1995 to 2020. *Sensors* **21,** 6532. https://doi.org/10.3390/s21196532 (2021).

18. Song, J., Fan, L. & Jiang, Y. The glossy premium: effects of product glossiness on consumer judgments of firsthand and secondhand products. *J. Mark.* **90,** 157–175. https://doi.org/10.1177/00222429251383926 (2026).

19. Huang, Y., Song, W. & Yang, L. Avoiding through glossiness and approaching through matte: the effect of visual finish on perceived product effectiveness. *Psychol. Mark.* **40,** 262–273. https://doi.org/10.1002/mar.21764 (2023).

20. Sun, B. The mapping between color-material-finish (CMF) and style imagery: a case study of Neo-Chinese armchairs. *BioResources* **21,** 4340–4356. https://doi.org/10.15376/biores.21.2.4340-4356 (2026).

21. He, C., Sun, X. & Xiao, W. Research on evaluation of CMF based on vision in automobile seat design. In *Advances in Interdisciplinary Practice in Industrial Design* (ed. Shin, C. S.), *Advances in Intelligent Systems and Computing* Vol. 968, 146–158 (Springer, 2020). https://doi.org/10.1007/978-3-030-20470-9_17.

22. Xu, G. & Fang, T. Research on camera CMF design based on Kansei engineering. In *HCI International 2022 Posters*, Communications in Computer and Information Science Vol. 1580, 169–176 (Springer, 2022). https://doi.org/10.1007/978-3-031-06417-3_23.

23. Kim, B., Lee, G. & Suk, H.-J. A color-material network of chairs through materials and colors. *Arch. Des. Res.* **36,** 7–19. https://doi.org/10.15187/adr.2023.02.36.1.7 (2023).

24. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15,** 155–163. https://doi.org/10.1016/j.jcm.2016.02.012 (2016).

25. Spearman, C. Correlation calculated from faulty data. *Br. J. Psychol.* **3,** 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x (1910).

26. Brown, W. Some experimental results in the correlation of mental abilities. *Br. J. Psychol.* **3,** 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x (1910).

27. Kaiser, H. F. An index of factorial simplicity. *Psychometrika* **39,** 31–36. https://doi.org/10.1007/bf02291575 (1974).

28. Bartlett, M. S. Tests of significance in factor analysis. *Br. J. Psychol. Stat. Sect.* **3,** 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x (1950).

29. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30,** 179–185. https://doi.org/10.1007/bf02289447 (1965).

30. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Lawrence Erlbaum Associates, 1988).

31. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15,** 72–101. https://doi.org/10.2307/1412159 (1904).

32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful

approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).

33. Karana, E. & Hekkert, P. User-material-product interrelationships in attributing meanings. *Int. J. Des.* **4,** 43–52 (2010).

Funding

This research received no external funding.

Author contributions

Yuwei Zhang conceived the study, designed the methodology, collected the data, wrote the analysis code, carried out and validated the statistical analyses, and drafted the manuscript. Chou Yong Tan supervised the work, coordinated the project and reviewed the manuscript. Beichen Wang wrote analysis code and prepared the figures. Yew Hoong Wong supervised the work and administered the project. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Files

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

- Supplementary.docx