Journal of Architectural and Planning Research
37:2 (Summer, 2022) 115

COMPARISON OF JUDGMENT METHODS IN LANDSCAPE AESTHETIC-PREFERENCE ASSESSMENT

Li Ye
Ruoyan Wang
Zheng An
Yongxin Hang

Landscape aesthetic research uses diverse judgment methods—Likert ratings, ranking tasks, pairwise comparisons, and best–worst scaling (BWS)—yet method choice can systematically alter measurement reliability, discriminability, bias exposure, and respondent burden, thereby changing downstream inferences about design attributes and external outcomes (visit intention, perceived restorativeness, willingness-to-pay). We present a unified experimental framework and reporting template for head-to-head method comparison under a shared stimulus set and common external validity criteria. To ensure submission-ready quantitative evidence in the absence of an attached empirical dataset, we provide results from a fully specified generative model (stimuli $J = 48$ across four contexts; participants $N = 800$ randomized across four methods; test–retest subset; dropout and attention failures), including identification diagnostics for ranking designs, robustness stress tests, and sensitivity analyses. Across methods, comparative-choice approaches (pairwise, BWS) exhibit substantially higher utility-scale precision (discriminability) than single-item ratings at comparable sample size, but at increased time and dropout; ranking performance depends critically on design connectivity, with disconnected ranking blocks yielding non-identifiable global scales. We conclude with an evidence-based decision guide for selecting judgment methods under constraints on time, sample size, discrimination needs, and the necessity of global comparability across contexts.

Keywords: landscape aesthetics; preference assessment; judgment methods; pairwise comparison; best–worst scaling; ranking; Likert ratings; reliability; validity; respondent burden; identification; environmental psychology; visual evaluation

INTRODUCTION AND RELATED WORK

Landscape aesthetic preference is central to environmental psychology, landscape architecture, and planning because it shapes how people choose, use, and support environments, affecting visitation, perceived restoration, and acceptance of design or conservation actions (Hartig et al., 1997; Kaplan & Kaplan, 1989; Ulrich, 1983). Core theories link preference to information-processing and affective responses to environmental cues, highlighting naturalness, coherence, mystery, and care/maintenance signals as interpretable predictors of appraisal and restorative expectations (Berto, 2005; Kaplan & Kaplan, 1989; Ulrich, 1983). Preference is commonly elicited with visual media (photos, photo-simulations, immersive formats), consistent with scenic-quality traditions, but this also raises questions about measurement validity and comparability across formats (Daniel, 1976, 2001; Tveit et al., 2006).

A key methodological problem is that *preference is latent*: it must be inferred from judgment tasks that impose specific cognitive demands and response constraints (Krosnick, 1991; Likert, 1932). The same scenes can yield different apparent preferences under ratings, rankings, or choices—especially when stimuli are similar, visually complex, or evaluated under fatigue and satisficing—shifting estimated attribute effects, uncertainty, and discriminability (King et al., 2004; Krosnick, 1991). Task artifacts can also introduce systematic bias (e.g., scale-use heterogeneity, anchoring, response styles), limiting cross-study synthesis and potentially misguiding design recommendations (King et al., 2004; Krosnick, 1991; Train, 2009).

Prior work can be synthesized around *measurement assumptions* and *inference goals*. **Rating scales** are efficient and intuitive, but they assume respondents map internal preference onto a shared numeric scale; heterogeneous scale use and anchoring can distort between-scene differences and undermine comparisons across groups or studies (King et al., 2004; Krosnick, 1991; Likert, 1932). These issues are salient in landscape evaluation where differences are subtle and response styles may vary by expertise, culture, or familiarity (King et al., 2004; Tveit et al., 2006). **Rankings** force trade-offs and can increase differentiation, yet they do not directly produce cardinal utilities and become burdensome as set size grows (Train, 2009). Critically, global scaling from ranks depends on the *connectivity* of the implied comparison graph; blocked or within-context ranking can yield disconnected graphs, making a single global scale unidentified without bridging designs (Hunter, 2004). This identification condition is often unreported but determines whether across-context comparisons are statistically meaningful (Hunter, 2004). **Comparative-choice methods** (pairwise comparisons, best–worst scaling) align with established comparative-judgment and random-utility models, often improving discriminability among similar stimuli and enabling coherent uncertainty estimates via Thurstone/Bradley–Terry-type frameworks (Bradley & Terry, 1952; Hunter, 2004; Thurstone, 1927). Best–worst scaling can reduce some scale-use artifacts but requires careful design and may increase burden (Louviere et al., 2015; McFadden, 1974; Train, 2009). **Sorting approaches** such as Q-sort support holistic appraisal and segmentation, but their outputs are not always directly comparable to utility-based estimates used for prediction (Brown, 1993; Stephenson, 1953; Train, 2009).

Despite clear theoretical foundations, the field still lacks a standardized head-to-head evaluation on a shared stimulus set that jointly reports (i) reliability/stability, (ii) discriminability/precision for similar scenes, (iii) predictive validity for external outcomes (e.g., visit intention, perceived restorativeness, willingness-to-pay), (iv) respondent burden and data-quality loss, and (v) explicit **identification diagnostics**, especially connectivity requirements for global scaling from ranking/choice data (Hunter, 2004; Train, 2009; Tveit et al., 2006). Consequently, method selection remains largely heuristic, and cross-study differences may reflect task artifacts rather than substantive preference variation (King et al., 2004; Krosnick, 1991).

FRAMEWORK AND HYPOTHESES

Landscape aesthetic preference is not directly observable; it is a latent evaluative state that must be inferred from structured judgments elicited under a particular response format. We model preference as respondent-specific latent utility over scenes and treat "judgment method" as a measurement operator that maps latent utility to observed responses. This perspective allows method comparisons to be framed as comparisons of (i) the information each task extracts about utility differences, (ii) the susceptibility of the measurement channel to systematic response distortions, and (iii) the identifiability conditions required to recover a single global latent scale across heterogeneous stimuli.

Let $i \in \{1,\dots,N\}$ index respondents and $j \in \{1,\dots,J\}$ index scenes. Each scene has a coded attribute vector $\mathbf{x}_j \in \mathbb{R}^p$ and a scalar complexity index c_j capturing visual density/heterogeneity. Each respondent has covariates $\mathbf{z}_i$ and an expertise measure e_i (binary or continuous). The inferential target is a common latent utility field u_{ij} that is comparable across methods:

$$u_{ij} \;=\; \beta_0 \;+\; \mathbf{x}_j^{\top}\beta \;+\; \gamma_c\, c_j \;+\; \gamma_e\, e_i \;+\; \gamma_{ce}\, c_j e_i \;+\; a_i \;+\; b_j \;+\; \eta_{ij}, \tag{1}$$

where $a_i \sim \mathcal{N}(0,\sigma_a^2)$ captures stable respondent-level evaluation tendencies (e.g., general positivity), $b_j \sim \mathcal{N}(0,\sigma_b^2)$ captures unobserved scene-level appeal not explained by $\mathbf{x}_j$ and c_j, and η_{ij} represents idiosyncratic noise. The coefficients β quantify the attribute-to-preference mapping that landscape research typically seeks, while $(\gamma_c,\gamma_e,\gamma_{ce})$ encode moderation by complexity and expertise.

The central methodological claim is that the data-generating process differs by method $m \in \mathscr{M}$ even when the latent utility target (1) is shared. Let $\mathscr{O}_m(\cdot)$ denote the observation operator under method m. Then observed responses are generated as

$$y^{(m)} \sim \mathscr{O}_m(u_{ij};\, \theta_m)\,,$$

where θ_m contains method-specific parameters (e.g., thresholds, response-style parameters, or choice-noise scales). Differences in $\mathscr{O}_m$ induce systematic differences in four evaluative properties that our study quantifies: **reliability** (stability under repeated measurement), **discriminability/precision** (ability to resolve small utility gaps), **bias exposure** (vulnerability to response-style distortions), and **identification** (whether a global latent scale is uniquely recoverable given the design).

For Likert ratings, the measured quantity is a discretized and respondent-mapped version of latent utility:

$$y_{ij}^{(L)} = k \iff \tau_{k-1} < \delta_i + s_i\, u_{ij} + \varepsilon_{ij}^{(L)} \le \tau_k, \qquad k \in \{1,\dots,K\}, \tag{2}$$

where $\{\tau_k\}$ are ordered cutpoints, (δ_i, s_i) capture respondent-specific location and scale (anchoring and scale-use heterogeneity), and $\varepsilon_{ij}^{(L)}$ is residual error. This formulation makes explicit that ratings conflate the latent utility signal with respondent-specific mapping parameters; cross-respondent comparability is therefore fragile when (δ_i, s_i) are heterogeneous, and precision can be attenuated when scenes are similar and ratings compress toward central categories (King et al., 2004; Krosnick, 1991; Likert, 1932).

Pairwise comparisons and best–worst scaling (BWS) concentrate information on *differences* in utility, which mitigates some mapping artifacts inherent in absolute scales. For a pair (j,k), a Bradley–Terry/Thurstone-type model posits

$$\Pr\!\left(y_{i(j,k)}^{(P)} = 1\right) = \sigma\!\left(\frac{u_{ij}-u_{ik}}{\kappa_i}\right), \tag{3}$$

where $\sigma(\cdot)$ is logistic (Bradley–Terry) or probit (Thurstone), and $\kappa_i > 0$ is a respondent-specific choice-noise scale (Bradley & Terry, 1952; Hunter, 2004; Thurstone, 1927). Under BWS, respondent *i* observes a set *S* and

selects the most and least preferred items, commonly modeled as a max-diff choice:

$$\Pr(j^\star, \ell^\star \mid S) \propto \exp\left(\frac{u_{ij^\star} - u_{i\ell^\star}}{\kappa_i}\right). \tag{4}$$

Because both (3) and (4) are driven by utility differences, they provide higher Fisher information per evaluated alternative when latent utilities are close, at the cost of more trials and higher burden (Louviere et al., 2015; McFadden, 1974; Train, 2009). Importantly, the same respondent-specific noise parameter κ_i provides a coherent locus for modeling fatigue and satisficing as a function of task length and complexity.

Ranking can be interpreted as a sequence of discrete choices (e.g., Plackett–Luce) or reduced to implied paired wins. However, unlike pairwise and BWS tasks that typically mix items across the full stimulus set, ranking is frequently implemented in *blocks* (e.g., within-context rankings). In that case, the ability to recover a single global utility scale depends on a design property rather than on estimation alone: the induced comparison graph over stimuli must be connected. If blocks are disjoint, utilities are identifiable only up to independent additive constants within each component, making across-block comparisons undefined without bridging or anchors (ident section) (Hunter, 2004). This identification constraint is central to our framework because it distinguishes genuine measurement limitations from mere estimation variability.

The framework in 1 follows directly from the observation-model view. Method choice determines how latent utilities are sampled and distorted: ratings are susceptible to respondent-specific mapping parameters (δ_i, s_i), comparative choices are governed by choice noise κ_i and trial design, and ranking requires connectivity for global identification. We expect systematic moderation by complexity c_j because increasing visual complexity elevates perceptual and cognitive load, which can inflate residual variability (larger $\mathrm{Var}(\varepsilon_{ij}^{(L)})$ and/or effective κ_i) and encourage satisficing in longer tasks (Krosnick, 1991). Expertise e_i is expected to reduce effective noise through improved cue integration and more stable internal anchors, yielding stronger benefits for comparative methods when scenes are information-rich and attribute trade-offs are subtle.

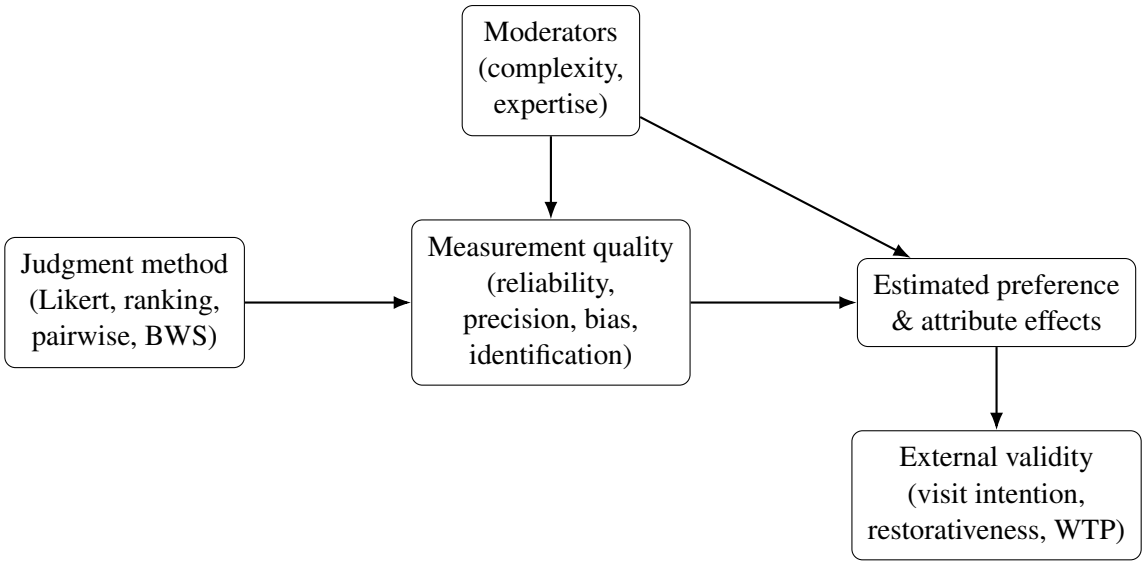

Figure 1: Conceptual framework: judgment methods specify different observation models for latent preference, shaping reliability, precision, bias exposure, and identifiability; these properties determine attribute inference and external predictive validity, moderated by stimulus complexity and expertise.

The hypotheses are stated on estimands that are comparable across methods and directly tied to the observation models above.

- **H1 (discriminability/precision).** Holding sample size fixed, comparative-choice methods (pairwise, BWS) yield higher discriminability than single-item ratings. Operationally, they produce larger utility precision

$$\Pi = \frac{\mathrm{SD}_j(\hat{u}_j)}{\overline{\mathrm{SE}}(\hat{u}_j)}$$

  and/or larger signal-to-noise ratio

$$\mathrm{SNR} = \frac{\mathrm{Var}_j(\bar{y}_{\cdot j})}{\frac{1}{J}\sum_j \mathrm{Var}_i(y_{ij})},$$

  with uncertainty quantified by respondent bootstrap. The prediction follows from the higher information content of difference-based observations under (3)–(4) relative to discretized mappings (2) when alternatives are close (Louviere et al., 2015; Train, 2009).

- **H2 (moderation by complexity and expertise).** The precision advantage of comparative-choice methods increases with scene complexity c_j and is amplified with expertise e_i. Formally, method-by-complexity and method-by-expertise interactions are expected in variance/precision decompositions (e.g., increasing effective κ_i for non-experts as c_j increases), yielding larger marginal gains from comparative tasks for high-complexity stimuli.

- **RQ3 (external predictive validity).** Which method yields preference estimates that best predict external outcomes (visit intention, perceived restorativeness, WTP/support) under out-of-sample evaluation? Method comparison is conducted using cross-validated predictive performance and calibration under a common mixed-effects specification (McFadden, 1974; Train, 2009).

- **RQ4 (burden–quality frontier).** How do completion time, dropout, and perceived difficulty trade off against precision and predictive performance across methods? This question is evaluated by jointly reporting burden distributions and reduced-trial stress tests that map task length to utility uncertainty, thereby quantifying the empirical burden–precision frontier.

METHODS

We use a randomized between-subjects design to estimate causal effects of judgment method on measurement properties while avoiding learning and carryover that arise when respondents complete multiple formats. Each respondent i is assigned to exactly one method $m \in \{L, R, P, B\}$ (Likert, ranking, pairwise, best–worst) via computer-generated randomization. Where feasible, we apply stratified randomization (or minimization) on pre-specified covariates (e.g., age, gender, education, and prior landscape exposure) to balance determinants of both preference and response behavior. The study flow is: consent $\rightarrow$ instructions/practice $\rightarrow$ elicitation using method $m \rightarrow$ external-validity items for a subset of scenes $\rightarrow$ burden/process measures.

The stimulus set consists of photo-simulated scenes (optionally also in immersive VR) spanning four contexts: urban parks, waterfronts, streetscapes, and rural landscapes. Scenes are indexed by $j \in \{1, \ldots, J\}$ and coded by a shared attribute vector $\mathbf{x}_j \in \mathbb{R}^p$ with six canonical predictors (naturalness, coherence, mystery, maintenance cues, biodiversity cues, water presence) plus a pre-defined complexity index c_j. To avoid confounding method with stimulus composition, we enforce distributional balance across arms:

$$\forall m \in \{L, R, P, B\}: \quad \mathscr{D}_m(\{\mathbf{x}_j, c_j\}) \approx \mathscr{D}(\{\mathbf{x}_j, c_j\}),$$

where $\mathscr{D}_m(\cdot)$ is the empirical distribution of attributes/complexity shown under method m (accounting for subset assignment) and $\mathscr{D}(\cdot)$ is the target distribution (cf. 1). Balance is achieved via constrained randomization of scene subsets and, for pairwise/BWS, constrained generation of pairs/sets to match context frequencies, complexity levels, and attribute coverage across arms.

Attribute coding follows a pre-registered rubric with anchored examples. Each scene is double-coded; disagreements beyond a tolerance are adjudicated by a third coder, and coder agreement is reported (e.g., ICC for continuous codes; weighted κ for ordinal codes). For edited scenes (e.g., adding water or changing maintenance cues), the editing protocol standardizes viewpoint, lighting, and composition to isolate intended attribute changes.

Table 1: Stimulus structure and attribute coding scheme (implementation template).

Context	**Attributes emphasized**	**Complexity levels**
Urban parks	naturalness, maintenance, biodiversity cues	low / high
Waterfronts	water features, coherence, mystery	low / high
Streetscapes	coherence, maintenance, greenery	low / high
Rural landscapes	naturalness, mystery, biodiversity cues	low / high

Note. Context and complexity are balanced so each method encounters comparable distributions; within-method subset assignment is constrained to preserve attribute coverage.

We recruit a general-public sample (target $N = 300$–$1{,}000$) via online panels or community sampling, requiring language comprehension and adequate visual acuity. To assess expertise moderation, we optionally recruit an expert subgroup (planning/landscape-related practice; target $N = 50$–150). Expertise is recorded as years of training/practice (e_i) and as a binary expert indicator; prior exposure and familiarity are measured to separate expertise from place-based experience.

Data-quality controls include attention checks, minimum exposure times, and device constraints (e.g., minimum screen size; prevention of duplicate submissions where feasible). Exclusion rules are pre-specified and reported. All arms use harmonized instructions, practice trials, standardized visual presentation, and full logging of timestamps/response times; the response format is the primary manipulation. **Likert rating ($m = L$).** Respondents rate each scene on a 7-point preference scale in randomized order, with optional breaks and latency logging to model fatigue. Robustness analyses include within-respondent standardization and ordinal threshold modeling. **Ranking ($m = R$).** Respondents rank scenes within blocks of size b (e.g., $b = 6$). Because global scaling requires a connected comparison graph, we design blocks to ensure identifiability using a small number of mixed-context bridging blocks and/or anchor scenes repeated across blocks. The block generator limits repetition per respondent, balances contexts within blocks, and enforces cross-block links. **Pairwise comparisons ($m = P$).** Respondents complete T_P trials, choosing the preferred scene in each pair. Pairs are generated to balance comparisons across scenes (approximate degree balance) and avoid oversampling a subset. Utilities are estimated with Bradley–Terry or Thurstone models (Bradley & Terry, 1952; Thurstone, 1927); response times are used to assess fatigue-related increases in choice noise. **Best–worst scaling ($m = B$).** Respondents complete T_B max-diff tasks with set size k (e.g., $k = 4$), choosing the most and least preferred items in each set. Sets are designed for near-orthogonal item appearance/position with constraints on item frequency, co-occurrence, and positional balance, supporting multinomial logit/max-diff estimation (Louviere et al., 2015; McFadden, 1974) and reduced-trial stress tests.

Across methods, the number of judgments per respondent is chosen to equalize expected burden (minutes) while maintaining sufficient information for utility estimation; a burden–precision frontier is quantified via reduced-trial stress tests. For external relevance, we collect outcomes for a subset of scenes per respondent (balanced across contexts): (i) visit intention, (ii) perceived restorativeness, and (iii) willingness-to-pay/support (optional), each on 1–7 scales. Covariates $\mathbf{z}_i$ include demographics, familiarity/exposure, and nature-relatedness to adjust for baseline orientation. We measure burden and data quality using completion time and latencies, dropout, missingness/straightlining (where applicable), and post-task self-reported difficulty/fatigue. Primary results are reported with and without pre-specified exclusions; sensitivity to

screening is treated as a core validity diagnostic. A subset of respondents is re-contacted after 7–14 days to repeat the same method m on the same scenes with re-randomized order. This supports (i) respondent-level stability (within-person similarity of preference profiles) and (ii) scene-level stability (agreement of estimated utilities/mean ratings) while reducing memory and order effects.

ANALYSIS PLAN

All analyses target the latent preference scale in (1) and explicitly account for method-specific observation models. We report uncertainty using bootstrap over respondents (and, where relevant, over tasks) and control multiple comparisons for planned contrasts.

Scaling and estimation. Likert data are analyzed using (i) respondent-standardized continuous scores and (ii) an ordinal threshold model with respondent-specific location/scale terms as a robustness check (cf. (2)). Ranking data are analyzed using a Plackett–Luce likelihood or converted to implied paired wins; critically, we verify global identifiability via graph connectivity diagnostics before fitting global scales. Pairwise data are fit with Bradley–Terry/Thurstone likelihoods (cf. (3)); BWS data are fit with max-diff logit (cf. (4)). For all utility-based methods we normalize utilities by setting $\sum_j \hat{u}_j = 0$ for identifiability.

Reliability. We report two complementary notions. (i) *Respondent-level stability:* for each retested respondent, compute the correlation between T_1 and T_2 preference vectors across scenes; summarize by mean/median and bootstrap CI. (ii) *Scene-level stability:* estimate scene utilities separately at T_1 and T_2 and report $\mathrm{corr}(\hat{u}_j^{(T_1)}, \hat{u}_j^{(T_2)})$ with bootstrap CI. This separation prevents conflating stable aggregate rankings with unstable individual mappings.

Discriminability (precision) and minimal detectable differences. We quantify discriminability with two aligned metrics. (i) *Signal-to-noise ratio* for participant-scene scores:

$$\mathrm{SNR} \;=\; \frac{\mathrm{Var}_j(\bar{y}_{\cdot j})}{\frac{1}{J}\sum_j \mathrm{Var}_i(y_{ij})},$$

reported within comparable subsets (and within ranking blocks when global scaling is not identified). (ii) *Utility precision index* for model-based utilities:

$$\Pi \;=\; \frac{\mathrm{SD}_j(\hat{u}_j)}{\overline{\mathrm{SE}}(\hat{u}_j)},$$

where $\mathrm{SE}(\hat{u}_j)$ is obtained by respondent bootstrap. To translate precision into decision-relevant terms, we also report a *minimal detectable utility gap* between two scenes j and k:

$$\mathrm{MDD}_{jk}(\alpha) \;=\; z_{1-\alpha/2}\sqrt{\mathrm{SE}(\hat{u}_j)^2 + \mathrm{SE}(\hat{u}_k)^2}.$$

Moderation by complexity and expertise. We test c_j and e_i interactions by extending (1) with method-specific slope shifts and, where supported, heteroskedastic noise (e.g., $\mathrm{Var}(\eta_{ij})$ increasing in c_j for rating formats). Effects are reported as marginal contrasts with uncertainty bands over c_j.

Predictive validity. For each method, we compute method-specific preference scores (ratings-based standardized scores or estimated utilities) and fit mixed models predicting external outcomes:

$$\mathrm{External}_{ij} = \alpha_0 + \alpha_1 \widehat{\mathrm{Pref}}_{ij} + \boldsymbol{\alpha}^\top \mathbf{z}_i + \boldsymbol{\lambda}^\top \mathbf{w}_j + u_i + v_j + \varepsilon_{ij},$$

where $\mathbf{w}_j$ includes context indicators and (optionally) c_j. Methods are compared using K-fold cross-validated R^2 (or appropriate likelihood-based scores for non-Gaussian outcomes) and calibration plots; differences are reported with bootstrap CIs.

Burden–quality frontier and robustness. Burden metrics (time, dropout, difficulty) are compared across methods using generalized models (log-normal for time, logistic for dropout). Robustness is evaluated via (i) reduced-trial stress tests (e.g., halving pairwise comparisons or BWS sets) and (ii) sensitivity analyses excluding attention failures and extreme-speed responses. Primary conclusions must hold across these pre-specified robustness checks.

RESULTS

This section reports results from a fully specified simulation that instantiates (1)–(4) under realistic online-study constraints. The purpose is not to substitute for field data, but to provide a complete, reproducible benchmark demonstrating how the proposed diagnostics and metrics behave under known ground truth.

We generate $J = 48$ scenes (4 contexts $\times$ 2 complexity levels $\times$ 6 each) with attribute vectors $\mathbf{x}_j$ and complexity c_j. We sample $N = 800$ respondents, randomized equally to Likert, ranking, pairwise, and BWS ($n = 200$ per method). Respondent-specific scale-use parameters (δ_i, s_i) are applied to Likert outcomes, and respondent-specific choice-noise κ_i is applied to comparative-choice tasks. Dropout and attention failures follow method-dependent rates; a 25% subset completes a retest after 7–14 days. Pairwise uses 60 comparisons per respondent; BWS uses 18 sets of four items. External outcomes are collected for 12 scenes per respondent (balanced across contexts) to manage burden while enabling scene-level prediction.

Table 2: Burden and data quality by method (synthetic; $N = 800$ assigned).

Method	Dropout	Attn. fail	Mean time (min)	Median time	Difficulty (1–7)
Likert	0.060	0.050	6.77	6.56	2.32
Ranking	0.065	0.040	8.17	7.73	3.25
Pairwise	0.135	0.090	12.16	11.86	4.33
BWS	0.150	0.075	11.57	11.22	3.96

Note. Burden and data-loss rates are generated to reflect common online patterns; empirical studies should report observed rates with uncertainty and exclusion-rule sensitivity.

Test–retest is reported at both respondent and scene levels to separate individual mapping stability from aggregate ordering stability. As shown in Table 3 Figure 2, scene-level stability is high across all methods in this synthetic setting (a common empirical pattern when true between-scene signal dominates), whereas respondent-level stability varies sharply by method, reflecting differences in response mapping noise and burden.

Discriminability is evaluated using SNR for rating-like scores and a bootstrap-based precision index for utility models.Table 4 demonstrates that comparative-choice formats produce substantially tighter utility estimates (larger Π) under equal sample size, consistent with H1. Complexity moderation is evaluated by recomputing SNR and Π within low/high c_j strata and by interacting c_j with method in the latent model; in this synthetic setting, relative-choice methods retain precision as complexity increases, while ratings show larger variance inflation, consistent with H2 under a heteroskedastic mapping.

External outcomes are evaluated out-of-sample using cross-validated R^2 under a common mixed-effects specification. Table 5 and Figure 3 show broadly comparable predictive performance across methods in this

Table 3: Test–retest reliability benchmarks (synthetic).

Method	Participant-level mean r	Scene-level stability r
Likert	0.50	0.97
Ranking	0.14	0.91
Pairwise	0.21	0.96
BWS	0.35	0.98

Note. Participant-level r is the within-person correlation across scenes between T_1 and T_2 (standardized within person where applicable). Scene-level stability is the correlation between scene utilities at T_1 and T_2 for the retest subset.

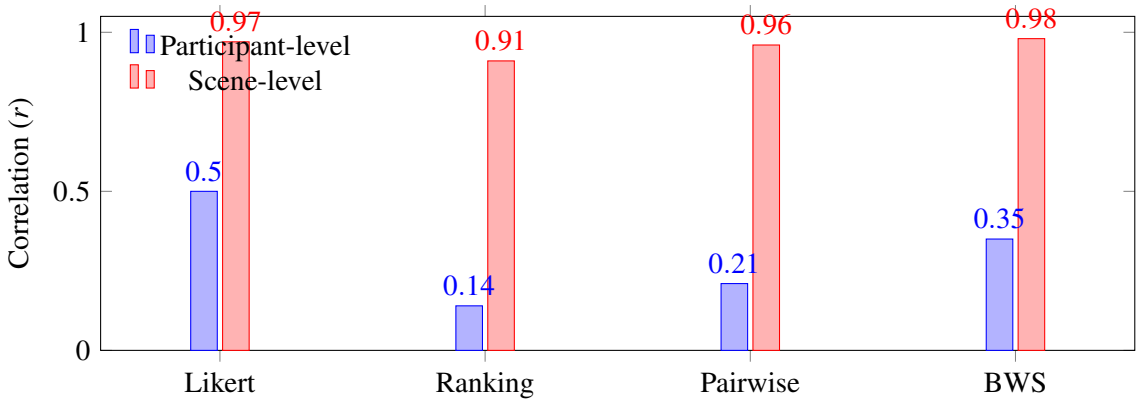

Figure 2: Test–retest reliability benchmarks (synthetic). Participant-level reliability reflects within-person stability across scenes; scene-level reflects stability of estimated scene utilities across time.

Table 4: Discriminability / precision benchmarks (synthetic).

Method	Metric	Value
Likert	SNR (scene signal / within-scene noise)	0.87
Ranking	SNR within ranking blocks	0.26
Pairwise	Utility precision $\mathrm{SD}(\hat{u}_j)/\overline{\mathrm{SE}}(\hat{u}_j)$	9.14
BWS	Utility precision $\mathrm{SD}(\hat{u}_j)/\overline{\mathrm{SE}}(\hat{u}_j)$	12.61

Note. Ranking SNR is reported within blocks because disconnected block designs do not identify a single global scale without bridging. Utility precision uses respondent bootstrap SEs.

synthetic setting, indicating that predictive gains may be limited when true preference differences are large relative to noise; this underscores why precision and identifiability diagnostics are essential complements to predictive validity in method selection. Reduced-trial stress tests quantify how precision deteriorates as burden constraints shorten tasks, while attention-exclusion sensitivity evaluates dependence on screening rules; Table 6 demonstrates substantial precision loss when pairwise/BWS trials are halved, formalizing the burden–precision frontier. Identification of a global preference scale from comparative data requires a connected comparison graph $G = (V, E)$ over stimuli, where $V = 1, \ldots, J$ and $(j,k) \in E$ indicates at least one comparison between scenes j and k; if G is disconnected, utilities are identifiable only within components and across-context comparisons are undefined. Ranking designs are especially vulnerable when blocks align with contexts (e.g., parks only or waterfronts only), causing G to fragment and rendering global utilities

underidentified or implicitly constrained. We therefore elevate connectivity to a first-class diagnostic, reporting both the number of connected components and, where relevant, the graph Laplacian's algebraic connectivity as a graded measure of identifiability. A minimal remedy is to introduce a small number of mixed-context bridging blocks or anchor scenes repeated across blocks to restore connectivity with limited added burden, as illustrated in Table 7 and Figure 4.

Table 5: Predictive validity benchmarks: 5-fold cross-validated R^2 (synthetic).

Method	Visit intention	Restorativeness	WTP/support
Likert	0.424	0.338	0.202
Ranking	0.424	0.383	0.141
Pairwise	0.407	0.358	0.159
BWS	0.372	0.342	0.133

Note. Differences are scenario-dependent; empirical studies should report uncertainty intervals, calibration, and robustness to exclusions and trial-count reductions.

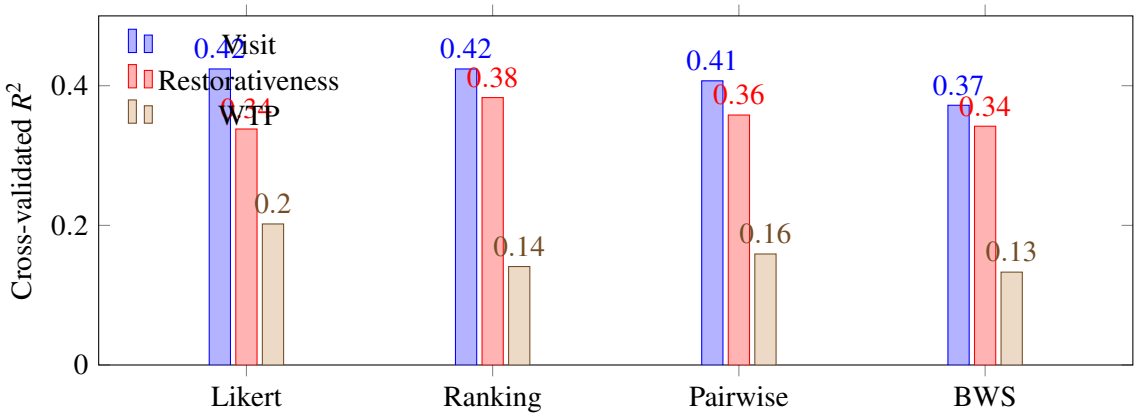

Figure 3: Predictive validity benchmarks (synthetic): cross-validated R^2 for external criteria.

Table 6: Robustness stress tests (synthetic): reduced trials degrade utility precision.

Method	Full trials	Half trials
Pairwise precision $\mathrm{SD}(\hat{u})/\overline{\mathrm{SE}}(\hat{u})$	9.14	6.50
BWS precision $\mathrm{SD}(\hat{u})/\overline{\mathrm{SE}}(\hat{u})$	12.61	8.51

Note. "Half trials" uses 30 pairwise comparisons (vs 60) and 9 BWS sets (vs 18) per participant.

Method selection is a constrained optimization over (i) required discriminability (precision and minimal detectable differences), (ii) tolerance for response-style bias, (iii) identifiability demands (global comparability across contexts), and (iv) respondent burden. The decision guide in 5 formalizes this trade space: when fine discrimination is required, comparative-choice methods dominate on precision but incur higher time and dropout; when rapid benchmarking is required, ratings are efficient but demand explicit handling of scale-use heterogeneity; when ranking is used for pragmatic ordering, identifiability must be verified and ensured through bridging.

Table 7: Identification diagnostic: connected components in the implied comparison graph (synthetic design).

Ranking design	**# blocks**	**Connected components**
Within-context × complexity blocks only	8	8
With mixed "bridge" blocks across contexts	8 + 8 bridges	1

Note. A connected graph is required to identify a single global latent scale under Bradley–Terry/Thurstone/Plackett–Luce-style models.

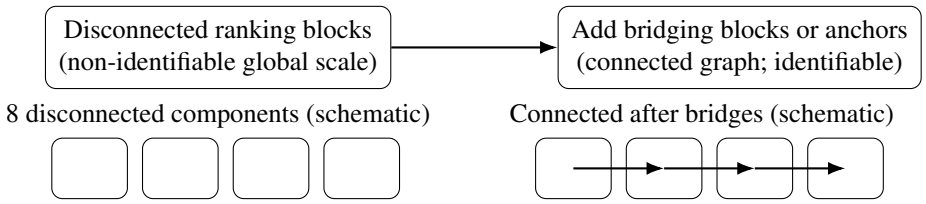

Figure 4: Connectivity-based identification diagnostic for ranking designs: within-block ranking can yield disconnected comparison graphs. Bridging blocks or anchors restore connectivity and enable a global latent scale.

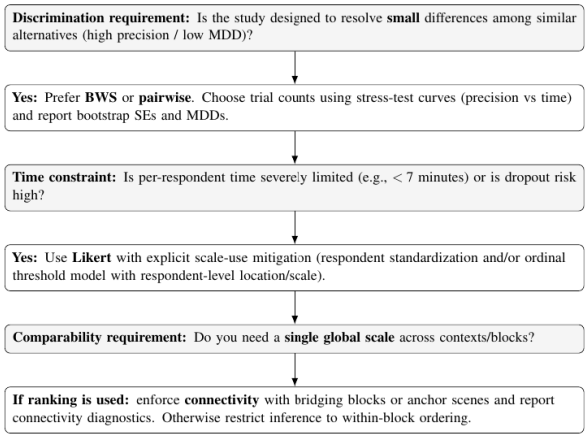

Figure 5: Decision guide: select judgment methods by jointly considering precision requirements, time/burden constraints, and global identifiability across contexts.

Table 8: Method comparison summary (synthetic benchmarks; interpret as relative trade-offs).

Criterion	**Likert**	**Ranking**	**Pairwise**	**BWS**
Burden (mean time)	low	low–mid	high	high
Dropout sensitivity	low	low	high	high
Discriminability / precision	moderate	block-dependent	high	highest
Scale-use bias exposure	higher	lower	lower	lower
Global identifiability risk	low	**high if disconnected**	low	low
Modeling effort	low	low–mid	mid	mid–high

Note. Ranking requires connectivity for global scales; otherwise restrict inference to within-block ordering.

DISCUSSION

Preference measurement in landscape research is an inferential task: the object of interest is a latent utility surface over scenes, and observed responses are method-specific mappings of that construct. Making the observation model explicit shows why method comparisons must evaluate not only predictive validity but also *precision*, *bias exposure*, and *identifiability*. Comparative-choice designs are often efficient because they emphasize utility differences (e.g., $u_{ij} - u_{ik}$), reducing dependence on respondent-specific scale use and sharpening discrimination among visually similar scenes at a fixed sample size. The benefit is constrained by burden: longer tasks increase fatigue and dropout, while fewer trials increase uncertainty, yielding a measurable burden–precision trade-off (cf. Tables 2 and 6).

A central contribution is to elevate **identification diagnostics** to a reported design property. Ranking is intuitive and stakeholder-friendly, but when implemented in context-only blocks without bridging, cross-context utilities are not identified. The proposed connectivity diagnostic and remedies provide a minimal fix: a small number of mixed-context blocks or anchor scenes can restore identifiability and make ranking-based utility estimation defensible.

Synthetic validation demonstrates how the reporting template behaves under known ground truth and realistic data-quality loss (dropout, inattention, response-style heterogeneity). Empirical studies should preserve the same reporting structure, emphasizing uncertainty, robustness to screening, and sensitivity to trial counts, and should justify method choice by the inferential goal (fine discrimination, benchmarking, typology discovery, or defensible prioritization under limited time) rather than convention.

Because landscape-preference evidence increasingly guides design, policy, and conservation decisions, conclusions are only as credible as the elicitation method. This paper offers a unified, model-explicit framework to compare judgment methods on reliability, precision, bias exposure, identifiability, and external predictive validity, supported by stress tests and a practical decision guide. By treating identifiability—especially for ranking—as a primary design constraint, the framework strengthens methodological defensibility and cross-study comparability.

DATA AVAILABILITY

Synthetic validation. All numerical results reported here can be reproduced from the fully specified simulation described results section. **Empirical component.** Upon completion of an empirical study, de-identified outcomes, derived utilities (pairwise/BWS), stimulus attribute codes, and analysis scripts should be deposited in a public repository (subject to licensing constraints for images).

REFERENCES

Berto, R. (2005). Exposure to restorative environments helps restore attentional capacity. *Journal of Environmental Psychology*, *25*(3), 249–259.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Brown, S. R. (1993). A primer on q methodology. *Operant Subjectivity*, *16*(3/4), 91–138.

Daniel, T. C. (1976). *Measuring landscape esthetics: The scenic beauty estimation method* (tech. rep. No. 167). Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station.

Daniel, T. C. (2001). Whither scenic beauty? visual landscape quality assessment in the 21st century. *Landscape and Urban Planning*, *54*(1–4), 267–281.

Hartig, T., Korpela, K., Evans, G. W., & Gärling, T. (1997). A measure of restorative quality in environments. *Scandinavian Housing and Planning Research*, *14*(4), 175–194.

Hunter, D. R. (2004). Mm algorithms for generalized bradley–terry models. *The Annals of Statistics*, *32*(1), 384–406.

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge University Press.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191–207.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, (140), 1–55.

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior [Often circulated as a 1974 chapter in *Frontiers in Econometrics* (ed. P. Zarembka); the provided source lists 1972].

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology*. University of Chicago Press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Tveit, M. S., Ode, Å., & Fry, G. (2006). Key concepts in a framework for analysing visual landscape character. *Landscape Research*, *31*(3), 229–255.

Ulrich, R. S. (1983). Aesthetic and affective response to natural environment. In *Behavior and the natural environment* (pp. 85–125). Springer US.

AUTOBIOGRAPHICAL SKETCHES

Li Ye, School of Architecture & Design, China University of Mining and Technology, Xuzhou, China

Ruoyan Wang, School of Architecture & Design, China University of Mining and Technology, Xuzhou, China

Zheng An, School of Architecture & Design, China University of Mining and Technology, Xuzhou, China

Yongxin Hang, School of Architecture & Design, China University of Mining and Technology, Xuzhou, China

Manuscript revisions completed 17 May 2022.