Mapping Cycling Experience in Valence–Arousal Space: A Geometric Approach to Appraisal Heterogeneity

Ayda Grisiute[1]
Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

Johanna Wörle
Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

Panagiotis Mavros
i3 (CNRS), Department of Social and Economic Sciences, Télécom Paris, Institut Polytechnique de Paris, France

Martin Raubal
Institute of Cartography and Geoinformation, ETH Zurich, Switzerland

Abstract

In this paper, we argue that heterogeneity in experiential appraisals can function as an informative signal for subjective cycling experience. We introduce a geometric framework to characterize the "shape" of experiential appraisal within a two-dimensional affect space. Using data from a crowd-sourced survey, in which participants ($N = 239$) evaluated 30 first-person cycling video scenarios, we describe appraisal heterogeneity not only by its overall extent (dispersion), but also by its directional structure (anisotropy), bimodality (polarization), and the diversity of affective categories and environmental cues (entropy). Our findings reveal a structural asymmetry. Negative environments result in a rigid, more linear stress response, whereas positive environments allow for diffuse, multi-polar appraisals. We further find that observer experience modulates the magnitude of dispersion, whereas the direction of appraisal remains relatively stable across demographic subgroups. We propose that this approach can quantitatively distinguish between those environments that are consistently perceived as non-bikeable and those that evoke conflicting interpretations. Our methodology complements user-centric models, such as cycling archetypes, by describing the underlying structure of divergent experiences.

2012 ACM Subject Classification Applied computing → Transportation; Applied computing → Psychology

Keywords and phrases bike infrastructure, affect grid, inter-rater variability, online survey

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.8

Supplementary Material *Software*: `https://github.com/mie-lab/cycling_experience`
Audiovisual: `https://www.youtube.com/@cycling-survey/playlists`

Declaration on GenAI/LLM use Generative AI tools were used for language editing and formatting. All scientific content, analyses, and conclusions were developed and validated by the authors.

1 Introduction

Cycling experience is widely recognized as a multidimensional construct, shaped not only by physical infrastructure but also by perceptual, affective, and cognitive processes [29]. As an embodied mode of travel, cycling requires the continuous processing of sensory, emotional, and interactional cues [27]. Experiential appraisals also have practical importance: perceived

[1] Corresponding author

comfort and safety may directly influence cycling uptake, with downstream impacts on public health, urban livability, and transport-related emissions [11]. In turn, rather than relying on objective infrastructural descriptors alone, researchers are increasingly measuring and modeling cycling experience through affective responses to environmental stimuli [8, 17].

Many studies report large appraisal variation which they commonly attribute to observer characteristics (demographics, archetypes) [7, 3]. This focus overlooks the fact that environments with identical mean ratings may elicit distinct response patterns. For instance, when evaluating the same street, cyclists' emotional responses might vary, appearing as a scattered cloud of data points (i.e., diffused). Alternatively, riders might split into distinct opinion clusters, such as half finding the street exciting and the other half finding it highly stressful (i.e., polarized). We refer to these structural characteristics of how appraisals are distributed across riders as *structural response properties*. However, little research has focused on analyzing these properties.

Structural differences could reveal whether an environment sends clear or conflicting signals to its users, serving as a diagnostic signal of *shared environmental legibility*. Urban legibility traditionally refers to the ease with which a city's parts can be recognized and organized into a coherent mental pattern [19]. Here, we extend this notion beyond navigational clarity to how reliably environmental cues convey a shared affective meaning. To assess this empirically, we need a representation in which structural response properties become visible. When subjective experience is represented two-dimensionally, distinct geometric signatures can emerge. Responses may be broadly scattered (dispersion), elongated along a dominant affective dimension (anisotropy), clustered into distinct camps (polarization), or spread across qualitatively distinct affective states (entropy). These properties offer a novel way to audit cycling environments, as scenarios with similar mean ratings may require different interventions depending on whether appraisals are consistent or contested.

To investigate these ideas, we re-analyzed a crowd-sourced online survey in which participants evaluated short, first-person videos of cycling scenarios using the Affect Grid, a two-dimensional measure of affective valence (pleasantness) and arousal (activation) [25]. By combining controlled video stimuli with a repeated-measures design, we obtained dense two-dimensional response distributions for each cycling scenario. Both the location and the structure of affective responses in valence–arousal space were examined to assess the extent to which this variation is driven by environmental properties as opposed to observer covariates.

We hypothesize that subjective cycling experience is not a product of cycling infrastructure (used here as an umbrella term for dedicated facilities such as separated bike lanes, advisory lanes, and shared-use paths) or observer traits alone, but that specific environments systematically elicit structured appraisal heterogeneity signatures. The following three research questions will be addressed:

- **RQ1.** How do cycling scenarios differ in valence–arousal appraisal space?
- **RQ2.** What distinct response geometry signatures characterize cycling scenarios, and how do they relate to environmental cues?
- **RQ3.** To what extent is the structure of appraisal heterogeneity moderated by observer characteristics versus scenario properties?

To evaluate this approach, we introduce a set of geometric metrics that characterize the structural response patterns of subjective appraisals (Sections 2 and 3). We demonstrate that high- and low-stress environments generate distinctly different appraisal signatures, which our geometric framework captures reliably across diverse observer backgrounds (Sections 4 and 5).

2 Background

This section introduces existing work on subjective cycling experience and discusses how it is typically measured. We first discuss cycling experience as an emerging multidimensional (perceptual, cognitive, and sensory) construct, and then evaluate current modeling and measurement approaches and sources of inter-subjective variability.

2.1 Cycling Experience as a Multidimensional Construct

Cycling research has traditionally evaluated cycling environments using objective indicators such as facility type, traffic volume, or network connectivity [8]. However, these descriptors only partially predict subjective experience [4, 5]. This leaves a "perception gap" where similar cycling infrastructures may evoke highly variable appraisals or, conversely, distinct environments result in comparable appraisals. For instance, low-traffic streets without dedicated cycling facilities are often rated similarly to high-traffic streets equipped with such facilities [1, 17]. Subjective cycling experience is therefore increasingly conceptualized as a perceptual–cognitive appraisal. In this paper, we adopt the definition of an episodic affective and evaluative response to environmental stimuli encountered during a trip from Lim et al. [18]. This process is inherently embodied: unlike car drivers or transit users, cyclists are continuously exposed to their surroundings, requiring the real-time processing of visual, auditory, and interactional signals [6, 27].

Most empirical studies attribute variability in environmental appraisals primarily to observer-related factors, such as demographic characteristics or cycling archetypes, which reflect variation in confidence, risk tolerance, habits, and cycling purpose [7, 3]. While valuable for identifying which individuals experience the environment differently, these approaches often treat observer heterogeneity as the primary cause. They may overlook the possibility that certain environments inherently afford multiple interpretations [22].

2.2 Tools and Modalities for Cycling Experience Measurement

Research on subjective cycling experience employs a range of measurement modalities, from in-situ ratings and physiological sensors to using virtual environments or videos as stimuli [16, 8]. These differ in the richness and timing of information provided, potentially capturing distinct facets of the experience, where evaluations emerge through different processes (recall, immediate reflection, sensory responses) [13].

Stimulus. When cycling, relevant cues unfold dynamically: perceived safety and comfort depend on motion, closing speeds, and overtaking dynamics [17]. While static images require viewers to infer dynamic processes, video and virtual environments (VR) preserve visual continuity and interaction cues essential for spatial affordance processing [9, 24]. Although they lack the vestibular and haptic feedback of physical cycling, research on the behavioral validity of traffic and cycling simulators suggests that sensory realism beyond a task-relevant threshold may yield diminishing returns for specific judgment tasks [26, 21].

Sampling strategy. Examining observer-driven variance vs. stimulus-driven ambiguity typically requires large, diverse samples. Although naturalistic studies offer high realism, they often rely on small, homogeneous samples. Online crowd-sourced surveys trade experimental control for access to heterogeneous participant pools [12]. In the context of cycling experience, several studies emphasize that large and diverse samples are essential to better differentiate stimulus-driven variation from observer-related differences [14, 1].

Measurement scales. Objective measures to assess cycling experience typically use psychophysiological sensors (e.g., heart rate, skin conductance) to capture variations in physiological arousal, commonly an indicator of vigilance and stress [16, 28]. Subjective measures, on the other hand, rely on self-report methods such as interviews or rating scales to measure overall appraisal (e.g., comfort, safety, pleasantness), thus primarily reflecting affective valence [28]. However, both dimensions are essential for a holistic cycling evaluation. Affect research suggests that the valence–arousal relationship depends on stimulus clarity. In clear situations, pleasantness and stress covary in a direct, predictable manner. In contrast, when environments present ambiguous affective cues, pleasantness and stress do not predictably covary, resulting in scattered responses that vary widely across individuals [20, 15]. Yet, cycling research has rarely explored whether mixed reactions to a single environment reflect the cyclists' characteristics, or if certain spaces naturally elicit polarized responses.

Two instruments are most commonly used in the joint measurement of valence and arousal: the *Self-Assessment Manikin* (SAM) [2] and Russell's *Affect Grid* [25]. SAM uses three independent pictorial scales for valence, arousal, and dominance, sampled sequentially. The Affect Grid captures valence and arousal jointly as a single choice in a 2D coordinate space, omitting dominance. In this study, we adopt the Affect Grid for two reasons. First, its single-choice, two-dimensional mapping minimizes cognitive fatigue during the rapid evaluation of 30 scenarios, whereas SAM requires three sequential ratings per stimulus. Second, the Affect Grid is inherently spatial, forming a 2D coordinate system that directly supports the geometric metrics central to our analysis.

3 Methodology

We developed this geometric approach to subjective experience by examining response distributions in valence–arousal space. Our methodology combined (i) first-person video stimuli of diverse cycling scenarios and (ii) a repeated-measures online survey that yielded dense affective response distributions across participants. We quantified not only mean affect but also the magnitude and structure of appraisal. Finally, we tested whether observer covariates systematically moderated responses to high-heterogeneity cycling scenarios. Figure 1 provides an overview of the study methodology.

3.1 Study Design

Participants, Recruitment, and Ethics. Participants ($N = 313$) were recruited via institutional mailing lists, social media, physical leaflets, and newsletters using snowball sampling. Inclusion criteria required age ≥ 18, normal or corrected-to-normal vision, and sufficient English proficiency. Participation was voluntary and could be discontinued at any time. Data were stored, processed, and reported in accordance with ETH Zurich data protection practices. Ethical approval was granted by the institutional ethics committee (Application 24-ETHICS-386).

Cycling Scenario Generation and Selection. First-person perspective cycling footage was recorded in Zurich, Switzerland between September and October 2024. To minimize confounding effects from extreme congestion or weather, recording occurred during fair weather and non-peak hours. Footage was captured using a handlebar-mounted action camera (DJI Action 4) on an e-bike, maintaining a mean speed of 16.5 km/h, with small variations due to local topography. Raw footage was processed to remove intersections, stops, and maneuvering, keeping linear segments which were then cut into 30-second clips. From this

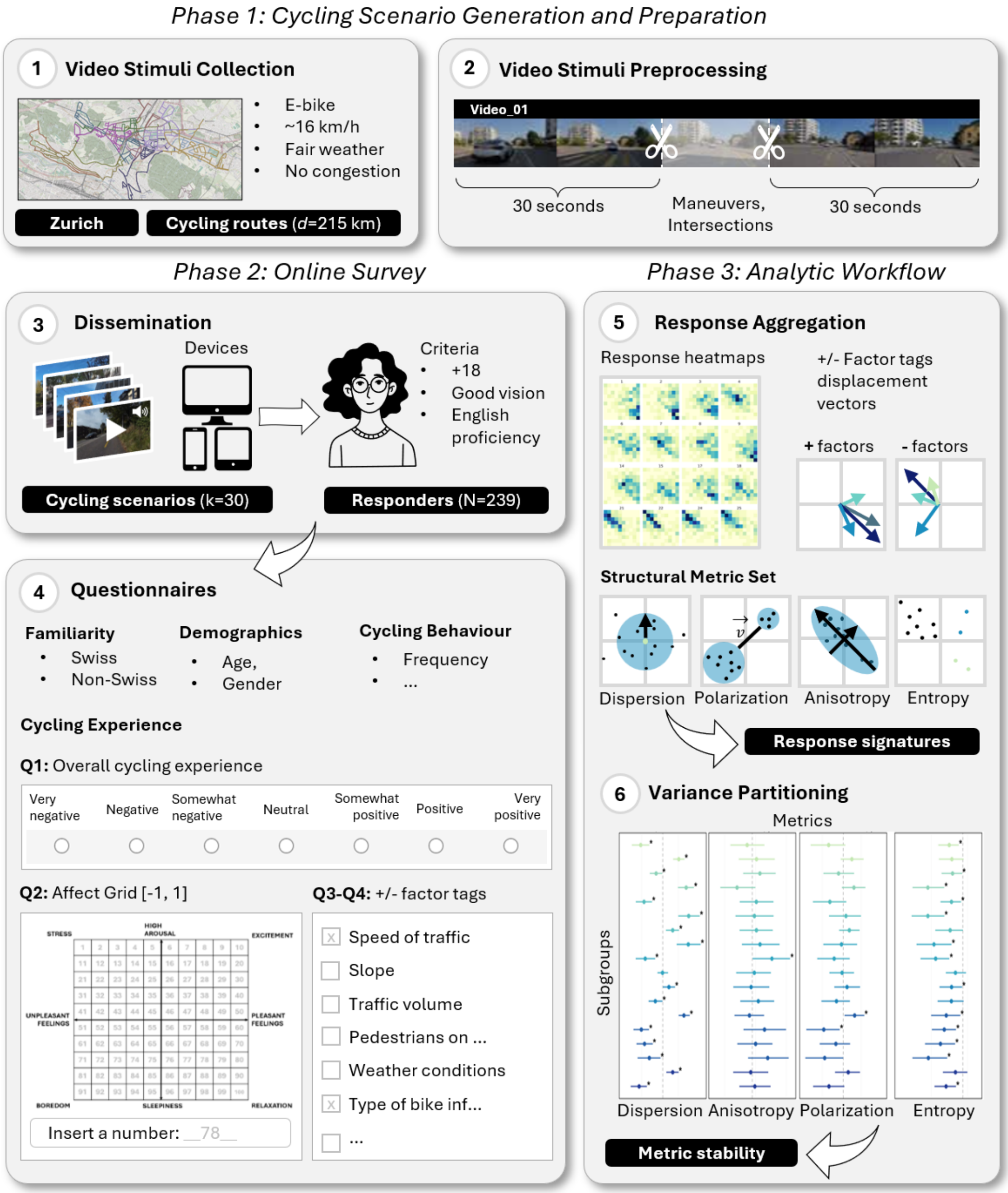

Figure 1 Procedural workflow of the study methodology.

pool, 30 scenarios (see Supplementary Material) were selected to sample diverse co-occurring environmental characteristics drawn from the cycling literature [1, 10]. Specifically, these factors include cycling infrastructure type, traffic speed and volume, car overtakes, slope, surface quality, pedestrian and cyclist presence, surrounding greenery and buildings, and weather conditions (see Figure 2). They correspond with the positive and negative factors used in our questionnaire (Section 3). Additional factors (e.g., public transit lanes, points of interest, side parking) were excluded to keep the questionnaire workload manageable and avoid respondent fatigue. Privacy was ensured by anonymizing faces and license plates using the EgoBlur model followed by manual verification [23].

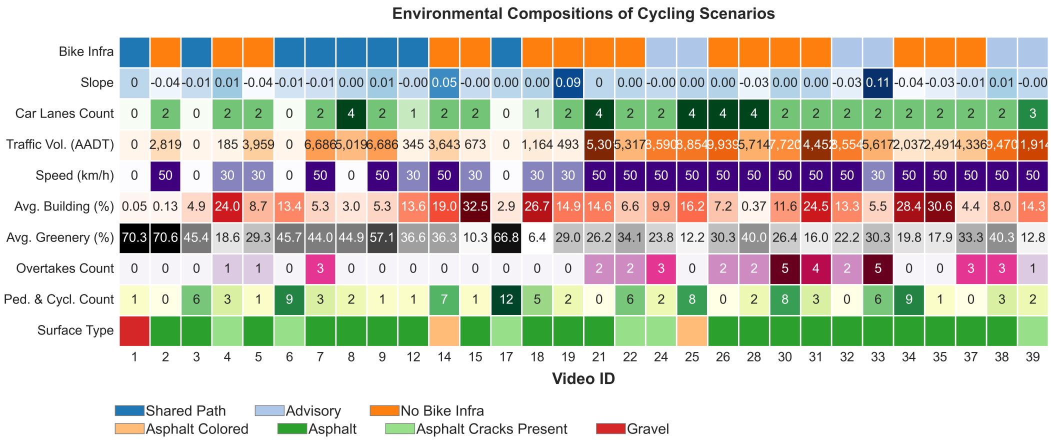

Figure 2 The co-occurrence and variance of specific infrastructural and environmental attributes across the selected video scenarios.

Procedure. The survey was implemented in Microsoft Forms and administered in English. After providing informed consent, participants completed a background questionnaire capturing demographics and cycling characteristics (purpose, environment, frequency, confidence). Participants then viewed the 30 video clips (with available audio) in randomized order. Following each scenario, they provided three evaluations: (i) an overall cycling experience rating on a 7-point rating scale to serve as a unidimensional benchmark; (ii) an affective appraisal using the Affect Grid [25]; and (iii) a selection of perceived positive and negative environmental factors introduced above. We used country of residence (Switzerland versus other) as a high-level proxy for environmental familiarity to test for moderators in affective responses. The Affect Grid required participants to choose a single cell from a valence–arousal coordinate system organized into four quadrants, which was then mapped onto a continuous $[-1, 1]$ scale for geometric analysis (see Figure 1). The cycling scenario viewing time was 15 minutes with an expected average completion time of 30 minutes.

3.2 Geometric Characterization of Appraisal

We treated the set of valence–arousal responses for each cycling scenario as a distinct geometric object in two-dimensional affect space. This step was motivated by the fact that it is often overlooked in the literature and by the observation from an initial visual screening that scenarios with similar mean ratings can arise from qualitatively different response configurations: some may cluster tightly, while others spread diffusely, align along a dominant affective axis, or split into competing clusters. To reveal these patterns, we quantified appraisal structure using five metrics (see Table 1). These metrics separated the magnitude of heterogeneity (Dispersion) from its structural properties: its directional orientation (Anisotropy Index) and bimodality (Polarization Index). We also utilized Quadrant Entropy and Cue Entropy to assess whether affective ambiguity aligned with a diversity of perceived environmental cues. Our approach is based on the premise that the valence–arousal relationship is conditional on stimulus clarity [20].

Table 1 Taxonomy of metrics for quantifying appraisal heterogeneity in the $[-1, 1]$ valence (V) and arousal (A) space. Indices j and i denote cycling scenario and participant; N_j is the sample size for scenario j.

Metric	Sym.	Definition, Formula & Diagnostic Interpretation
Location & Magnitude		
Affective Centroid	$\mathbf{C}_j$	*Definitions:* The arithmetic mean (center of mass) of the affective responses. $\mathbf{C}_j = (\bar{V}_j, \bar{A}_j)$ *Interpretation:* The aggregate affective "consensus" for scenario j.
Dispersion	D_j	*Definitions:* The average Euclidean distance of individual responses from the scenario's centroid. $D_j = \frac{1}{N_j} \sum_i \lVert \mathbf{r}_{i,j} - \mathbf{C}_j \rVert_2$ *Interpretation:* The magnitude of appraisal heterogeneity, with higher values (larger radius) indicating lower consensus. In the $[-1, 1]$ Affect Grid coordinate system, the maximum possible dispersion is $D \approx 2.83$ (extreme cases in opposite corners); a range of 0.59–0.77 represents moderate spread relative to the scale's bounds.
Geometric Structure		
Anisotropy Index	AI_j	*Definitions:* A measure of the directional bias of the response distribution, derived from the eigenvalues ($\lambda_{1,2}$) of the 2×2 covariance matrix. $AI_j = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$ *Interpretation:* Response "shape." $0 \rightarrow$ isotropic (diffuse); $1 \rightarrow$ linear (structured coupling).
Polarization Index	PI_j	*Definitions:* The Euclidean distance between the mean vectors ($\boldsymbol{\mu}$) of a 2-component Gaussian Mixture Model (GMM). Note: $PI_j > 0$ only if $\Delta\text{BIC} \geq 6$ (conservative threshold) and cluster separation ≥ 0.5. $PI_j = \lVert \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rVert_2$ *Interpretation:* The structural splitting. $0 \rightarrow$ unimodal consensus; $0.8 - 1.0 \rightarrow$ moderate polarization (likely adjacent quadrants); $1 \rightarrow$ presence of distinct clusters (likely opposing quadrants).
Categorical Ambiguity		
Quadrant Entropy	H_j^Q	*Definitions:* The Shannon entropy of the response distribution across the four affective quadrants (e.g., Stress, Excitement), where p_k is the relative count in quadrant k. $H_j^Q = -\sum_{k=1}^{4} p_k \ln p_k$ *Interpretation:* Categorical conflict. High values indicate that responses span multiple affective states rather than a single quadrant.
Cue Entropy	H_j^{cue}	*Definitions:* The Shannon entropy of tagged environmental factors, where p_m is the proportion of factors falling into category m (positive or negative). $H_j^{\text{cue}} = -\sum p_m \ln p_m$ *Interpretation:* The diversity of environmental attribution, which measures the consistency of the environmental signal indicated by respondents.

3.3 Analysis Workflow

We operationalized affective response distributions as geometric objects to quantify: (i) the affective position of cycling scenarios; (ii) the magnitude and structural organization of response heterogeneity; and (iii) the stability of these patterns across observer covariates. Prior to analysis, responses were screened for incomplete submissions and completion times that were too short (< 15 minutes). Validated responses were projected into a continuous 2D coordinate system $[-1, 1]$ for analysis. Several demographic variables were aggregated into broader categories to ensure sufficient subgroup sizes for stable estimation (e.g., cycling frequency into Frequent vs. Infrequent). To establish a baseline of within-subgroup agreement in scenario-specific appraisals, we computed the intraclass correlation coefficient ICC(2,1), separately for valence and arousal across the 30 scenarios. This two-way random-effects coefficient estimates the reliability of single ratings, quantifying the expected agreement between two randomly selected participants from the same subgroup, and penalizes both random disagreement and systematic differences in scale use (rater bias). The workflow proceeded in three stages, following the order of the study's research questions.

Stage 1: Mapping the affective response landscape

We first established a global baseline by aggregating all responses to characterize the overall utilization of the valence–arousal space in the context of cycling scenarios. We then characterized each scenario by its affective centroid $\mathbf{C}_j = (\bar{V}_j, \bar{A}_j)$, providing a descriptive overview of how environments are distributed across the affect space. For each scenario, we applied Principal Component Analysis (PCA) to the scatter of individual ratings to identify the dominant axis along which individual responses for that scenario vary.

Stage 2: Diagnosing geometric appraisal signatures

We quantified appraisal heterogeneity at the scenario level using geometric metrics capturing Dispersion (D_j), Anisotropy (AI_j), and Polarization (PI_j). For each scenario, PI_j was computed by fitting one- and two-component Gaussian Mixture Models (GMMs) to the valence–arousal responses and comparing them via BIC. If a two-component model was preferred, PI_j was set to the Euclidean distance between the means of the two clusters; otherwise, it was set to zero. GMMs used full covariance matrices and 10 random initializations for stability. To distinguish between diffuse stochastic variance and structured signals, we examined the relationships among these metrics and performed split-sample reliability tests across random and subgroup partitions. To examine the environmental context of this heterogeneity, we analyzed the association between Quadrant Entropy (H_j^Q) and Cue Entropy (H_j^{cue}), testing whether conflicting affective states co-occur with the diverse perception of environmental cues. To further characterize the role of these cues, we calculated conditional displacement vectors, where each vector represents the affective shift associated with a specific tagged environmental factor. For each scenario, we computed the mean valence–arousal response of the subset of participants who tagged that factor (as positive or negative) and subtracted the scenario centroid, then aggregated these centered shifts across scenarios to yield one characteristic vector per factor.

Stage 3: Testing subgroup robustness of appraisal structure

We evaluated whether the observed appraisal structure reflected stable properties of the cycling scenarios or varied systematically across observer subgroups. To distinguish scenario-driven signals from observer-driven variance, we recomputed the scenario-level metrics

Table 2 Mean ratings for valence ($\bar{V}$) and arousal ($\bar{A}$), calculated from raw, unaveraged individual responses, alongside inter-rater agreement within demographic and cycling-related subgroups across the 30 scenarios. Higher values indicate stronger collective within-subgroup agreement on scenario-specific ratings.

Category	Count (n)	%	$\bar{V}$	$\bar{A}$	$\bar{V}$ ICC(2,1)	$\bar{A}$ ICC(2,1)
Gender						
Female	122	51.0	0.00	0.12	0.281	0.177
Male	117	49.0	0.04	0.14	0.321	0.203
Age						
18–25 years	92	38.5	0.07	0.05***	0.209	0.159
26–35 years	82	34.3	-0.01	0.17	0.390	0.262
36–45 years	39	16.3	0.01	0.16	0.340	0.201
46+ years	26	10.9	-0.03	0.25	0.342	0.114
Cycling Frequency						
Infrequent	126	52.7	0.03	0.11	0.231	0.148
Frequent	113	47.3	-0.03	0.16	0.398	0.246
Cycling Purpose						
Commuting	114	47.7	0.03	0.14	0.390	0.246
Recreational	98	41.0	0.02	0.12	0.234	0.135
I do not cycle	27	11.3	-0.05	0.14	0.228	0.186
Cycling Confidence						
Confident	196	82.0	0.03	0.13	0.326	0.209
Not confident	43	18.0	-0.02	0.14	0.200	0.120
Cycling Environment						
Urban area	162	67.8	0.06*	0.14	0.304	0.190
Rural area	49	20.5	-0.05	0.10	0.314	0.208
I do not cycle	25	10.5	-0.05	0.14	0.250	0.171
Both	3	1.3	-0.12	0.22	–	–
Familiarity						
Swiss	87	36.4	0.02	0.13	0.422	0.270
Non-Swiss	152	63.6	0.02	0.13	0.246	0.152

Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (post-hoc tests).

(D_j, AI_j, PI_j, H_j^Q, and H_j^{cue}) separately for each participant subgroup (cycling frequency, confidence, purpose, gender, age, and environmental familiarity). Uncertainty was estimated using $n = 1000$ bootstrap resamples over the scenarios to calculate 95% confidence intervals. If subgroup estimates tracked the overall (all-participant) values with substantial overlap in confidence intervals, it would indicate that the appraisal "shape" is an intrinsic, scenario-specific property. On the contrary, systematic shifts and little overlap in confidence intervals would suggest that appraisal geometry is primarily moderated by observer background.

4 Results

This section presents the findings from our online survey, structured around the three research questions (see Section 1). We first describe sample characteristics, then analyze scenario-specific appraisal signatures based on the defined set of metrics. Finally, we partition the observed variance to determine whether the geometric structure of these appraisals is driven by the cycling scenario properties or moderated by observer characteristics.

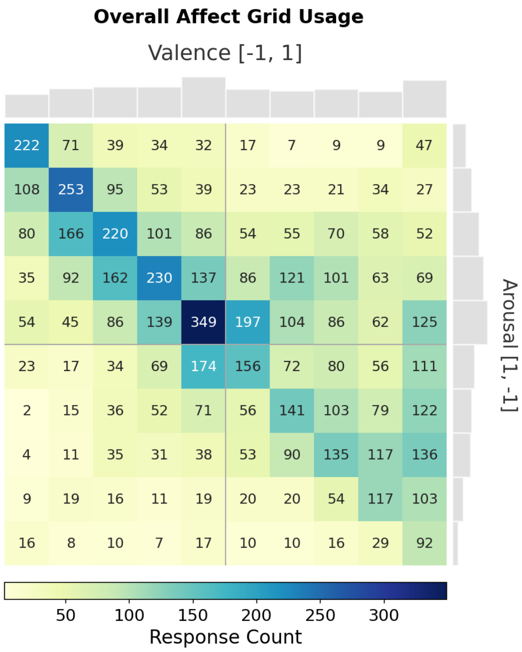

Figure 3 Overall usage of the Affect Grid across all cycling scenarios. Cell intensity indicates how frequently each valence–arousal cell was selected.

4.1 Sample, Measurement Scale Coverage, and Baseline Agreement

After quality checks, $N = 239$ valid responses were retained. Overall, mean valence was clustered around neutrality (≈ 0), while mean arousal was consistently positive, reflecting the vigilant nature of cycling (see Table 2). Arousal baselines vary by age, with the youngest cohort (18–25) reporting significantly lower arousal than older groups, which might reflect

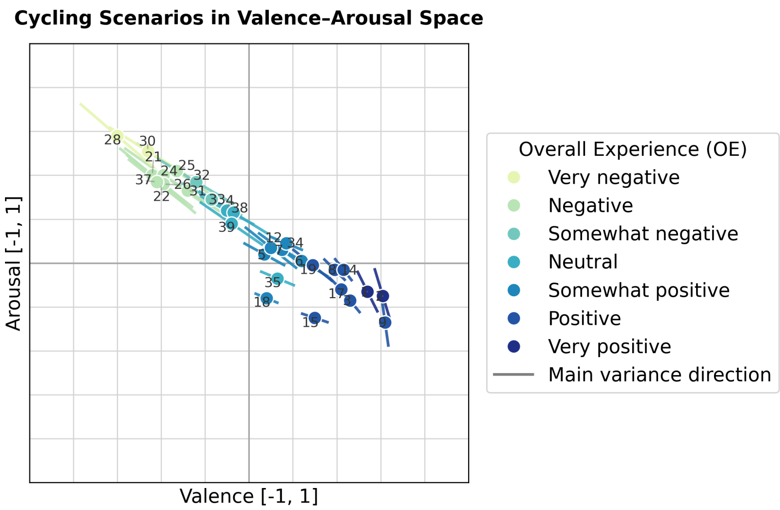

Figure 4 Cycling scenarios in valence–arousal space. Points indicate cycling scenario centroids (means), colored by modal Overall Experience (OE) rating. Line segments represent the main axes along which response varies (first principal component of the valence-arousal covariance matrix).

Table 3 Scenario-level geometric signatures. $(\bar{V}, \bar{A})$ and SD represent the affective centroid and 1-dimensional spread. ρ_{VA} denotes the valence–arousal correlation; D is the 2-dimensional mean dispersion; AI is anisotropy (directional structure); PI is polarization (clustering); and H^Q is quadrant entropy (categorical ambiguity). OE is the modal Overall Experience rating.

Scenario	$\bar{V} \pm SD$	$\bar{A} \pm SD$	OE (Mode)	ρ_{VA}	D	AI	PI	H^Q
1	0.54 ± 0.55	-0.13 ± 0.64	Very positive	-0.20	0.77	0.27	1.11	1.12
2	0.61 ± 0.46	-0.15 ± 0.63	Very positive	-0.27	0.71	0.37	1.17	1.05
3	0.46 ± 0.50	-0.17 ± 0.52	Positive	-0.29	0.66	0.21	0.94	1.13
4	-0.10 ± 0.55	0.24 ± 0.45	Neutral	-0.35	0.63	0.39	0.92	1.22
5	0.07 ± 0.57	0.04 ± 0.49	Somewhat positive	-0.27	0.69	0.30	1.03	1.35
6	0.24 ± 0.61	0.01 ± 0.51	Somewhat positive	-0.31	0.74	0.39	1.22	1.25
7	0.15 ± 0.54	0.07 ± 0.48	Somewhat positive	-0.35	0.67	0.37	0.92	1.33
8	0.39 ± 0.52	-0.03 ± 0.52	Positive	-0.18	0.68	0.18	0.95	1.24
9	0.62 ± 0.41	-0.27 ± 0.56	Positive	-0.15	0.62	0.32	0.85	0.95
12	0.10 ± 0.59	0.07 ± 0.54	Somewhat positive	-0.39	0.74	0.41	1.10	1.32
14	0.43 ± 0.48	-0.03 ± 0.50	Positive	-0.21	0.63	0.16	0.88	1.18
15	0.30 ± 0.57	-0.25 ± 0.50	Positive	-0.12	0.68	0.18	1.04	1.22
17	0.42 ± 0.54	-0.12 ± 0.53	Positive	-0.25	0.70	0.22	0.93	1.23
18	0.08 ± 0.56	-0.16 ± 0.50	Somewhat positive	-0.09	0.69	0.16	0.89	1.35
19	0.29 ± 0.55	-0.01 ± 0.52	Positive	-0.31	0.70	0.30	0.97	1.29
21	-0.44 ± 0.52	0.40 ± 0.47	Negative	-0.62	0.59	0.55	1.10	0.84
22	-0.40 ± 0.54	0.36 ± 0.49	Negative	-0.58	0.60	0.56	0.86	0.87
24	-0.39 ± 0.57	0.40 ± 0.46	Negative	-0.54	0.60	0.60	1.12	0.84
25	-0.33 ± 0.60	0.42 ± 0.48	Negative	-0.62	0.66	0.64	1.06	0.95
26	-0.39 ± 0.56	0.36 ± 0.50	Negative	-0.47	0.62	0.48	1.17	0.87
28	-0.60 ± 0.55	0.58 ± 0.50	Very negative	-0.63	0.61	0.60	0.94	0.73
30	-0.46 ± 0.58	0.51 ± 0.45	Very negative	-0.60	0.61	0.57	1.08	0.77
31	-0.28 ± 0.57	0.33 ± 0.45	Negative	-0.56	0.62	0.53	0.94	1.00
32	-0.24 ± 0.60	0.37 ± 0.46	Somewhat negative	-0.51	0.65	0.56	1.11	1.00
33	-0.17 ± 0.56	0.29 ± 0.43	Somewhat negative	-0.42	0.61	0.45	0.94	1.09
34	0.17 ± 0.55	0.09 ± 0.46	Somewhat positive	-0.18	0.64	0.23	0.00	1.32
35	0.13 ± 0.60	-0.07 ± 0.51	Neutral	-0.19	0.72	0.24	0.00	1.36
37	-0.42 ± 0.52	0.37 ± 0.47	Negative	-0.50	0.59	0.46	0.98	0.92
38	-0.07 ± 0.57	0.23 ± 0.45	Neutral	-0.48	0.65	0.51	0.95	1.21
39	-0.08 ± 0.58	0.18 ± 0.49	Neutral	-0.45	0.68	0.47	0.92	1.25

higher risk tolerance. Urban cyclists report significantly higher valence than rural cyclists. We see few differences in mean affect across gender, frequency, or confidence. However, utility-oriented traits (commuting as primary purpose, high confidence and frequency) and environmental familiarity (Swiss vs. non-Swiss) are associated with higher collective agreement ($ICC(2,1) \approx 0.33 - 0.42$), though differences are not significant. This suggests that while experience and familiarity do not alter the nature of the affective response (i.e., the means remain stable), they may drive more aligned expectations for cycling infrastructure. Figure 3 shows that overall grid usage is anisotropic, with responses clustering along a diagonal from calm–pleasant to tense–unpleasant. The two affective dimensions exhibit a moderate negative correlation ($r = -0.49$). The low-valence/low-arousal quadrant is rarely selected, likely due to the cycling task being inherently active and vigilant. An additional factor may be the stimulus design: the 30-second video clips did not include static congestion and extended delays, which would be more likely to elicit states of boredom or fatigue.

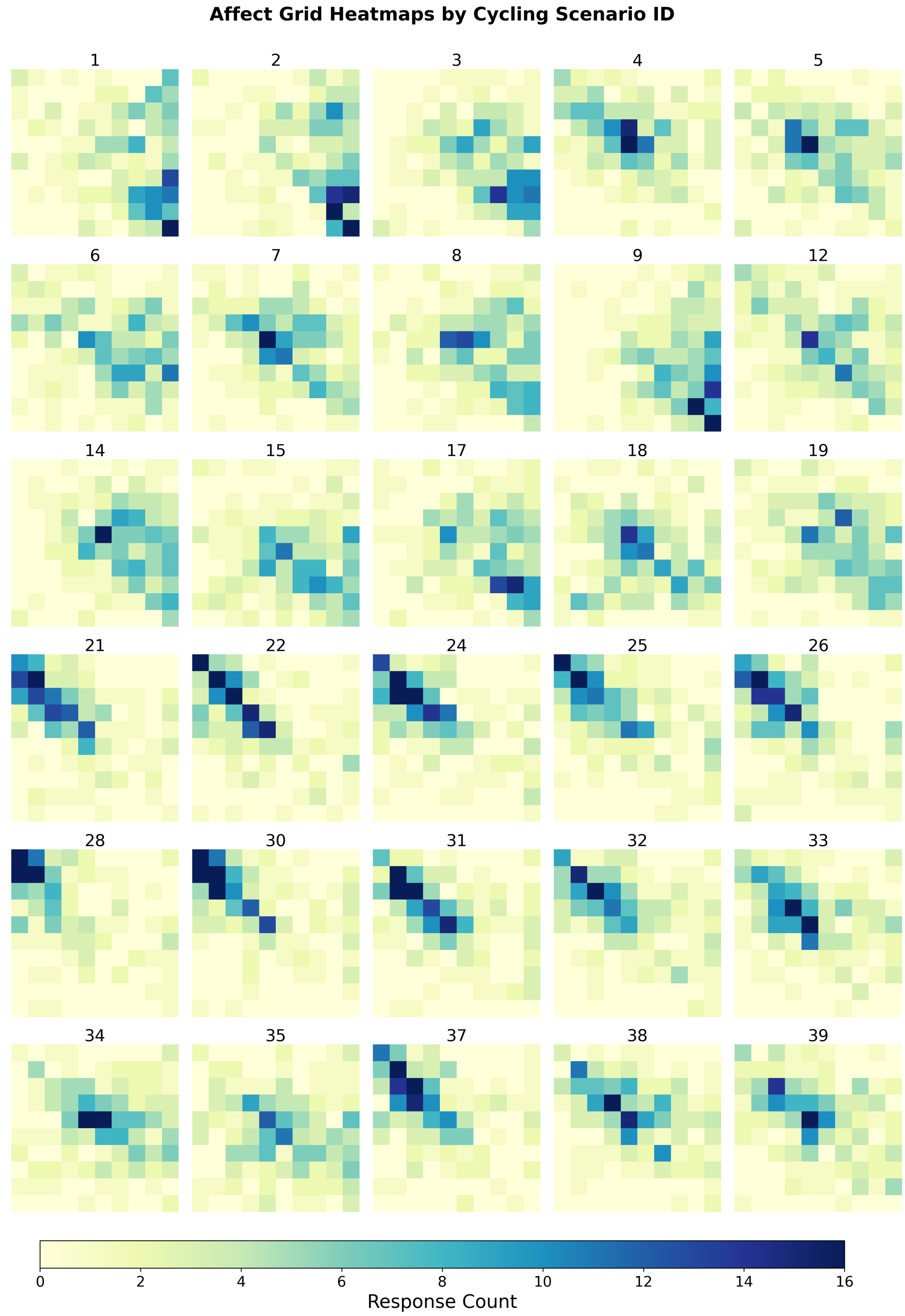

Figure 5 Response distribution across 30 cycling scenarios. Videos for all street scenarios are available online (see Supplementary Material). Darker cells indicate higher concentrations of responses for that specific scenario.

4.2 Stage 1: Mapping the Affective Cycling Experience (RQ1)

Scenario centroids and their dominant axes of variation reveal distinct structural response properties (see Figure 4). While responses generally align along the valence–arousal diagonal, the geometry of consensus differs across cycling scenarios. Negatively rated scenarios (e.g., 22, 28, 30) manifest as tight, linear streaks anchored in the high-arousal quadrant. In contrast, positively rated (e.g., 1, 2, 9) and neutrally rated scenarios (e.g., 12, 18, 35) appear as diffuse clouds, suggesting that while stress elicits uniform appraisal, positive experiences elicit more heterogeneous interpretations (see Figure 5). In stressful environments, a clear signal tightly couples valence and arousal along a shared axis, whereas in positive environments, these dimensions vary more independently. This complements the broader findings of Mattek et al. [20], who showed that the correlation between valence and arousal weakens as stimulus ambiguity increases.

4.3 Stage 2: Appraisal Signatures in Affect Space (RQ2)

Geometric Structure of Affective Responses. Table 3 quantifies the internal structure of appraisal across scenarios. While total dispersion (D) remains relatively stable (0.59–0.77), other metrics vary significantly. Negatively rated scenarios (e.g., 25, 28) show high anisotropy ($AI \geq 0.60$) and strong negative valence–arousal correlations ($\rho_{VA} \approx -0.60$), indicating that responses align tightly along a single affective axis. Positive and neutral scenarios show lower anisotropy, with responses spread more evenly across both affect axes. The polarization index (PI) further isolates specific cases, such as Scenario 6 ($PI = 1.22$), where moderate dispersion ($D = 0.74$) reflects not a shared moderate experience but a split into distinct clusters (e.g., stress vs. excitement) that the centroid alone would mask.

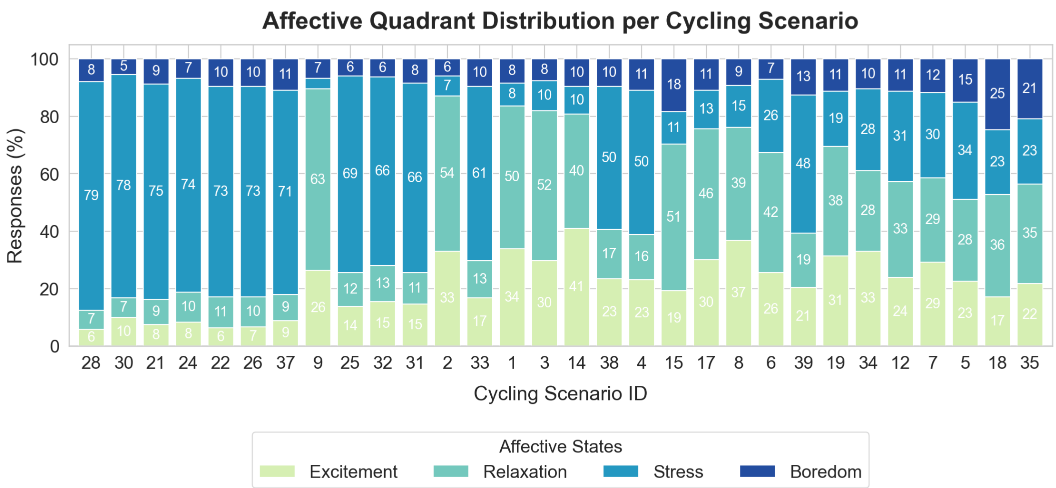

Figure 6 Response distributions across four affective quadrants for each cycling scenario. Scenarios are arranged from left to right in ascending order of Quadrant Entropy (H^Q), with the lowest-entropy (most consensus) scenarios on the left and the highest-entropy (most ambiguous) scenarios on the right.

Environmental Drivers and Directional Shifts. As shown in Figure 6, neutral scenarios exhibit high Quadrant Entropy (H^Q) and span multiple affective quadrants, whereas negative scenarios are dominated by the "Stress" quadrant. H^Q correlates moderately with H^{cue}

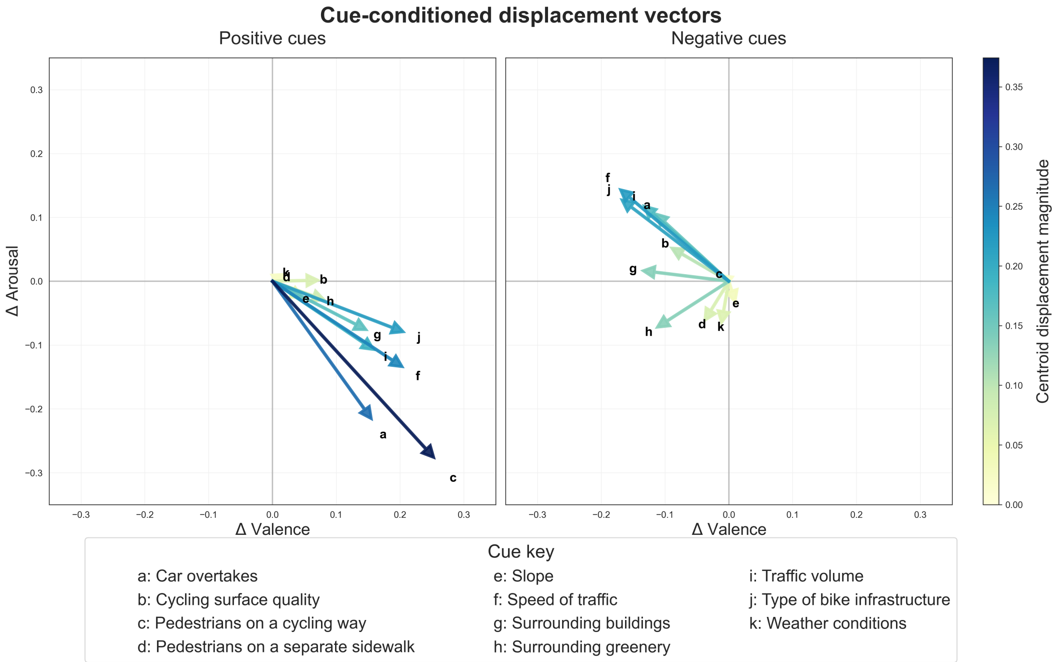

Figure 7 Conditional displacement vectors of tagged environmental factors. Vectors represent the directional shift in valence and arousal associated with specific positive (left) and negative (right) cues relative to the scenario-level mean.

($r = 0.61$), suggesting that scenarios with diverse environmental tags (e.g., conflicting perceptions of "traffic volume" versus "bike infrastructure") are associated with higher affective ambiguity. The conditional displacement vectors (see Figure 7) isolate the directional influence of these factors. Negative factors form a tight cluster pointing toward the high-arousal/negative-valence quadrant. Positive factors shift appraisals toward the relaxation quadrant. "Pedestrians on a cycling way" and "Car overtakes" point toward relaxation (when tagged as positive), suggesting an "absence-of" signal.

Overall, these patterns converge into three recurring structural signatures (see Figure 8). Negative scenarios (e.g., 28, 21) exhibit strong directional structure and low affective ambiguity ($H^Q \approx 0.73 - 0.84$), where high traffic volumes (up to 15,301 AADT) and lack of protected infrastructure align with broad agreement on negative appraisal. Mixed-signal environments (e.g., 6, 12) instead show higher dispersion ($D = 0.74$) and high quadrant entropy ($H^Q \geq 1.32$): stress-mitigating factors (lower speeds, greenery) co-occur with stressors (car overtakes), preventing a single affective consensus. Finally, consensus-positive environments (e.g., 9, 1) shift toward positive valence and lower arousal driven by separated infrastructure and high greenery share ($> 50\%$), but differ in PI profiles (0.85 vs. 1.11).

4.4 Partitioning Sources of Variation (RQ3)

To distinguish scenario-driven signals from observer-driven variability, we stratified the selected metrics by participant subgroups (see Figure 9). The analysis reveals a difference: D, PI, and H^{cue} are systematically sensitive to observer characteristics, whereas AI and H^Q remain largely stable across subgroups. Non-cyclists and infrequent riders show higher values in D, PI, and H^{cue} compared to commuters and frequent riders, which suggests

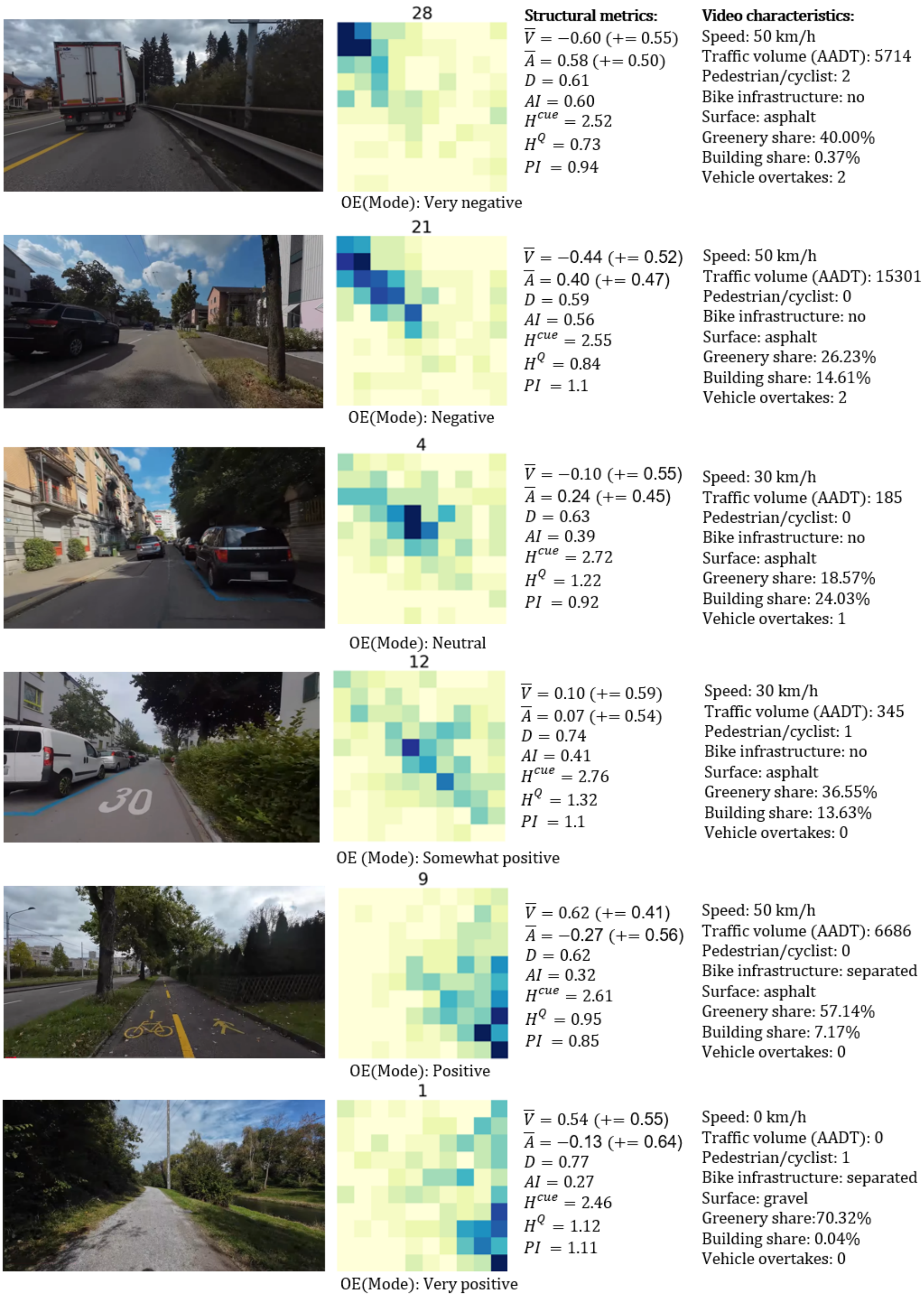

Figure 8 Representative cycling scenarios with varying combinations of physical infrastructure, traffic cues, and geometric properties ($D, AI, PI, H^{cue}, H^{Q}$) that characterize three distinct experience signatures: consensus-positive, structured-stress, and ambiguous-mixed signals.

that experience acts as a tuning mechanism that fosters more shared interpretations. The youngest cohort (18–25) displays significantly higher values across these metrics, consistent with less structured appraisal. Polarization is also lower among rural, recreational, less confident, and female participants. In contrast, the stability of AI and H^Q suggests that directional orientation and affective categorical ambiguity are intrinsic properties of cycling scenarios. In turn, while observer traits modulate the magnitude of variation, the shape of that variation serves as a robust, observer-independent diagnostic of the environmental signal.

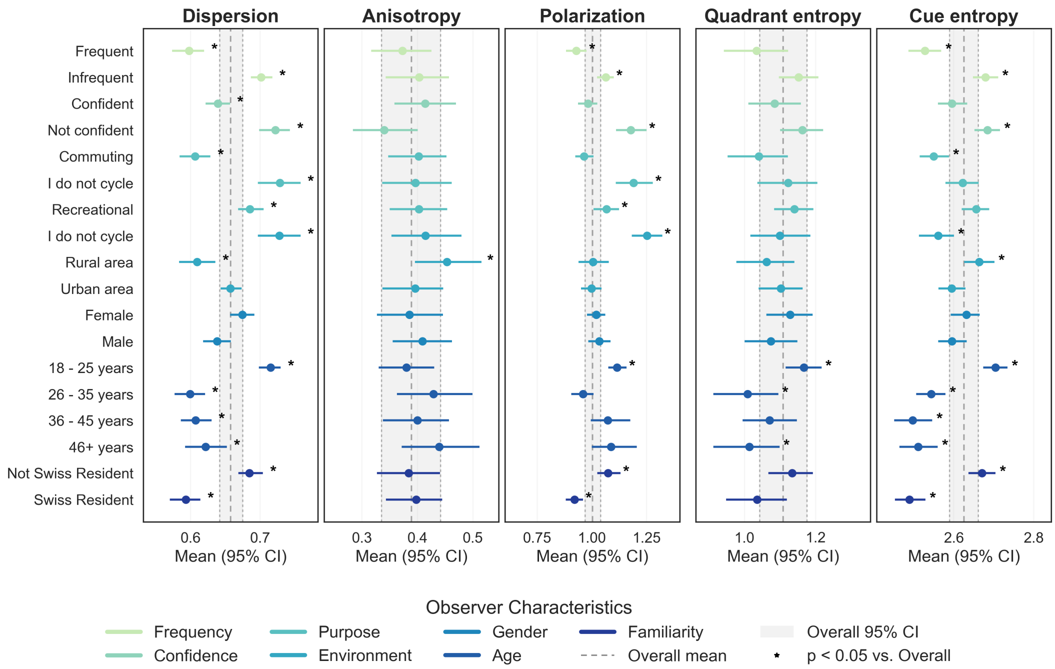

Figure 9 Stability of geometric appraisal metrics across participant subgroups. Individual dots represent subgroup means, while the vertical dashed lines indicate the aggregate sample baseline (mean and 95% CI). Asterisks (*) denote significant deviations from the overall sample mean ($p < 0.05$) based on post-hoc testing.

5 Discussion

This study treats appraisal heterogeneity in cycling experience as a diagnostic signal. We demonstrate that cycling scenarios can be characterized by distinct geometric signatures when projected onto affect space. These signatures provide a novel audit layer for distinguishing between environments that evoke clear stress and those that elicit contested interpretations.

Key Findings. First, we demonstrate that cycling scenarios differ not only in mean affect but in the underlying geometry of appraisal. Our metric set shows that appraisal heterogeneity within the subjective cycling experience is multidimensional with largely independent properties (see Table 3). Second, we identify a structural difference in response heterogeneity. High-stress scenarios typically manifest as clear, directional appraisals (high AI, low D), while positive scenarios result in diffuse appraisals (low AI, high D), appearing as scattered clouds (see Figure 8). Third, we find that affective state entropy correlates with tagged cue entropy, suggesting an association between signal heterogeneity and competing visual cue interpretations. Fourth, we show a "magnitude vs. shape" dissociation regarding sources of variation. Specifically, while observer experience acts as a tuning mechanism that changes the magnitude of appraisal heterogeneity (often consisting of polarized clusters), the response shape (AI) remains largely robust across participant subgroups (see Figure 9).

Possible Theoretical and Practical Implications. The presented structural response patterns provide complementary evidence for the valence ambiguity model by Mattek et al. [20], who showed across seven datasets of brief affective stimuli (faces, images, words, sentences, sounds; total $N = 195$) that the valence–arousal correlation depends on valence ambiguity, and we replicate this pattern in the cycling domain. More specifically, we found that unambiguous negative stimuli result in valence–arousal coupling (tight linear structure along the stress diagonal). More ambiguous responses to positive stimuli result in decoupled affective dimensions (diffuse, multi-directional variance), with a more heterogeneous response palette (e.g., excitement, relaxation, contentment). For practical intervention design, this difference could mean non-linear improvement dynamics. For example, adding amenities (e.g., greenery) to already "decoupled" (i.e., positive environments) may produce smaller, more scattered shifts as responses diffuse across multiple positive quadrants. On the other hand, removing stressors (e.g., through traffic volume reduction or physical separation) may target the "stress streaks," likely yielding a shift to a more dispersed consensus. This suggests that intervention impact is state-dependent: improving "somewhat positive" infrastructure may not yield linear affective gains, as the underlying geometry shifts from structured to diffuse. This geometric diagnostic enables planners to transition from modeling average satisfaction to auditing environmental legibility, better matching intervention strategies to the underlying structural diversity of cycling experiences. Importantly, such dispersion should not be a reason for concern to planners, as we show it can be attributed to individual differences rather than environmental characteristics.

Contributions. First, we reconceptualized appraisal heterogeneity as an informative signal. We introduced a distributional perspective on cycling experience that preserves the geometric information typically lost in measures of central tendency by separating affective response centroids from their structural geometry (D, AI, PI, H^Q). Second, we demonstrated a dissociation where observer characteristics (e.g., experience) moderate the *magnitude* of variance, yet the *geometric structure* remains largely scenario-driven and robust. Third, we provided a reusable metric set (see Table 1) supported by open-source code that demonstrates how to calculate these structural properties from raw survey data using standard statistical techniques. Finally, our approach complements existing facility typologies and cycling archetypes by enabling a more nuanced evaluation of urban interventions, where a successful street redesign may leave average valence unchanged while significantly reducing affective quadrant or environmental cue entropy, thereby improving the legibility and shared interpretability of the cycling environment.

Limitations. First, the generalizability of our findings is constrained by the sampling strategy. As all clips were recorded in Zurich, the observed signatures reflect the specific infrastructure typology and variability of the local urban form. Broader replication is needed to determine whether these geometric patterns hold in other traffic cultures. Similarly, our results are skewed toward a younger ($72.8\% \leq 35$ y.o.) and more confident (in terms of cycling) population. Given this snowball-recruited sample, we treat the present study as exploratory and foundational rather than confirmatory. Second, appraisal heterogeneity may be influenced by unmeasured observer traits. Although we accounted for cycling experience and confidence, latent variables such as individual attitudes, risk tolerance, sensation seeking, or prior crash history could further influence appraisal heterogeneity structure. Third, the Affect Grid is bounded by design, which constrains variance at the affective extremes. Its standard quadrant labels (e.g., “stress,” “boredom”) were also developed for general affect research and may not perfectly capture the specific sensations of cycling, though using validated terminology keeps our metrics comparable across studies. We also treat the absence of a dominance dimension (sampled by SAM but not by the Affect Grid) as a deliberate scope choice, traded against respondent fatigue across our 30 scenarios. Fourth, while first-person video captures dynamic visual cues, it lacks the vestibular feedback, physical effort, and personal control and risk that shape in-situ arousal, which was a deliberate trade-off to favor a broader outreach over full physical immersion. This, combined with the exclusion of prolonged static traffic congestion from our stimuli, likely explains the absence of the “boredom” quadrant (low valence, low arousal) in our dataset. Translating the framework to in-situ or immersive VR settings is therefore a necessary next step. The physical effort of real or active-VR cycling could elevate baseline arousal and shift or compress the geometric signatures we observe, leaving it unclear whether our structural patterns persist under fully embodied conditions.

6 Conclusions and Outlook

This paper introduces a geometric framework for analyzing subjective cycling experiences, moving beyond central tendency to quantify the structure of appraisal heterogeneity. The subset of metrics utilized proved largely robust across observer subgroups, suggesting they reflect intrinsic scenario properties rather than subgroup preference. We identified three distinct geometric signatures: diffuse consensus (positive environments), linear structured stress (negative environments), and polarized or ambiguous appraisals (mixed-signal environments). This lets practitioners distinguish between environments that are consistently stressful and those that potentially elicit multiple interpretations.

Future research should extend this geometric framework to in-situ experiments and diverse traffic cultures to test the universality of these structural signatures. Confirmatory follow-up studies should also employ a priori power and sample-size calculations and apply confirmatory statistical models (e.g., mixed-effects regressions) to formally test which environmental and observer variables predict the geometric signatures identified here. While our video-based methodology effectively captures dynamic visual-cognitive cycling experience appraisals, incorporating the physiological effort and vestibular feedback of real-world cycling or developing a cycling-relevant affect grid terminology remain the next research steps. Another direction is adding a dominance dimension to the appraisal. It could help differentiate our “structured-stress” signature into active stress (anger, frustration; high dominance) and passive stress (fear, helplessness; low dominance), though whether the resulting gain in resolution outweighs the added respondent burden in repeated-measures designs remains an open question. Ultimately, moving from a model of average satisfaction to a model of environmental legibility will support the design of cycling environments that are both inclusive and affectively coherent.

References

1 Frauke Luise Berghoefer and Mark Vollrath. Prefer what you like? Evaluation and preference of cycling infrastructures in a bicycle simulator. *Journal of Safety Research*, 87:157–167, 2023. doi:10.1016/j.jsr.2023.09.013.

2 Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. doi:10.1016/0005-7916(94)90063-9.

3 Laura Cabral and Amy M. Kim. An empirical reappraisal of the four types of cyclists. *Transportation Research Part A: Policy and Practice*, 137:206–221, 2020. doi:10.1016/j.tra.2020.05.006.

4 Ana Karina Christ, Miguel Costa, Manuel Marques, Carlos Roque, and Filipe Moura. Perceiving objective cycling safety: a systematic literature review. *Transportation Research Procedia*, 72:1380–1387, 2023. doi:10.1016/j.trpro.2023.11.601.

5 Miguel Costa, Manuel Marques, Felix Wilhelm Siebert, Carlos Lima Azevedo, and Filipe Moura. Bridging the gap: a scoping review exploring the relationship between objective and subjective cycling safety. *Transport Reviews*, 45(6):993–1015, 2025. doi:10.1080/01441647.2025.2537200.

6 Tineke De Jong and Aslak Fyhri. Spatial characteristics of unpleasant cycling experiences. *Journal of Transport Geography*, 112:103646, 2023. doi:10.1016/j.jtrangeo.2023.103646.

7 Jennifer Dill and Nathan McNeil. Revisiting the Four Types of Cyclists: Findings from a National Survey. *Transportation Research Record: Journal of the Transportation Research Board*, 2587(1):90–99, 2016. doi:10.3141/2587-11.

8 Haining Ding, Chenxi Wang, and Michal Gath-Morad. How built environment shapes cycling experience: A multi-scale review in historical urban contexts, 2025. arXiv:2509.15657. doi:10.48550/arXiv.2509.15657.

9 J. J. Gibson. The theory of Affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

10 Ayda Grisiute, Nina Wiedemann, Pieter Herthogs, and Martin Raubal. An ontology-based approach for harmonizing metrics in bike network evaluations. *Computers, Environment and Urban Systems*, 113:102178, 2024. doi:10.1016/j.compenvurbsys.2024.102178.

11 Eva Heinen, Bert Van Wee, and Kees Maat. Commuting by Bicycle: An Overview of the Literature. *Transport Reviews*, 30(1):59–96, 2010. doi:10.1080/01441640903187001.

12 Anja Katharina Huemer, Luzie Marianne Rosenboom, Melina Naujoks, and Elise Banach. Testing cycling infrastructure layout in virtual environments: An examination from a bicycle rider's perspective in simulation and online. *Transportation Research Interdisciplinary Perspectives*, 14:100586, 2022. doi:10.1016/j.trip.2022.100586.

13 Anjani Kalra, Tommy Lim, Lauren Pearson, and Ben Beck. Methods used to capture subjective user experiences in adults while riding bicycles: a scoping review. *Transport Reviews*, 43(3):453–477, 2023. doi:10.1080/01441647.2022.2123064.

14 Fatima Kchour, Salvatore Cafiso, and Giuseppina Pappalardo. Understanding Cyclists' Visual Behavior Using Eye-Tracking Technology: A Systematic Review. *Sensors*, 25(1):22, 2024. doi:10.3390/s25010022.

15 Peter Kuppens, Francis Tuerlinckx, Michelle Yik, Peter Koval, Joachim Coosemans, Kevin J. Zeng, and James A. Russell. The Relation Between Valence and Arousal in Subjective Experience Varies With Personality and Culture. *Journal of Personality*, 85(4):530–542, 2017. doi:10.1111/jopy.12258.

16 Tommy Lim, Anjani Kalra, Jason Thompson, Joanne Caldwell Odgers, and Ben Beck. Physiological measures of bicyclists' subjective experiences: A scoping review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 90:365–381, 2022. doi:10.1016/j.trf.2022.09.007.

17 Tommy Lim, Lauren Pearson, Jason Thompson, Joanne Caldwell Odgers, and Ben Beck. Establishing normative scores of affect, safety, and comfort for bicycling infrastructure and traffic conditions. *Transportation Research Part F: Traffic Psychology and Behaviour*, 114:337–353, 2025. doi:10.1016/j.trf.2025.06.006.

18 Tommy Lim, Jason Thompson, Lauren Pearson, Joanne Caldwell Odgers, and Ben Beck. Effects of within-trip subjective experiences on travel satisfaction and travel mode choice: A conceptual framework. *Transportation Research Part F: Traffic Psychology and Behaviour*, 104:201–216, 2024. doi:10.1016/j.trf.2024.05.024.

19 Kevin Lynch. *The Image of the City.* Publications of the Joint Center for Urban Studies. MIT Press, Cambridge, MA, 1960.

20 Alison M. Mattek, George L. Wolford, and Paul J. Whalen. A Mathematical Model Captures the Structure of Subjective Affect. *Perspectives on Psychological Science*, 12(3):508–526, 2017. doi:10.1177/1745691616685863.

21 Andrii Matviienko, Hajris Hoxha, and Max Mühlhäuser. What does it mean to cycle in Virtual Reality? Exploring cycling fidelity and control of VR bicycle simulators. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3544548.3581050.

22 Jun Qi, Xueqiong Tang, and Yongju Luo. Affordances of scenic cycleways: How recreational cyclists interact with different environments. *Urban Forestry & Urban Greening*, 64:127270, 2021. doi:10.1016/j.ufug.2021.127270.

23 Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M. Parkhi. EgoBlur: Responsible innovation in Aria, 2023. arXiv:2308.13093. doi:10.48550/arXiv.2308.13093.

24 Martin Raubal. Wayfinding: Affordances and Agent Simulation. In Shashi Shekhar and Hui Xiong, editors, *Encyclopedia of GIS*, pages 1243–1246. Springer US, Boston, MA, 2008. doi:10.1007/978-0-387-35973-1_1469.

25 James Russell, Anna Weiss, and G. Mendelsohn. Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, 57:493–502, 1989. doi:10.1037/0022-3514.57.3.493.

26 Sonja Schneider and Klaus Bengler. Evaluating behavioral validity in traffic simulators. In *Proceedings of TRA2020, the 8th Transport Research Arena*, Helsinki, Finland, 2020.

27 Justin Spinney. Cycling the City: Movement, Meaning and Method. *Geography Compass*, 3(2):817–835, 2009. doi:10.1111/j.1749-8198.2008.00211.x.

28 Markus Tomzig, Johanna Wörle, Alexandra Neukum, and Martin Baumann. The arousal paradox in critical task performance in automated driving during sleep inertia using a quasi experimental approach. *Scientific Reports*, 15(1):20338, 2025. doi:10.1038/s41598-025-08726-4.

29 Ran Zhang, Marco Te Brömmelstroet, Anna Nikolaeva, and George Liu. Cycling subjective experience: A conceptual framework and methods review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 101:142–159, 2024. doi:10.1016/j.trf.2023.12.021.