PREPARED FOR SUBMISSION TO THE *JOURNAL OF MARKETING RESEARCH* · DIALOGUE AI × THE OBSERVATORY WORKING PAPER · JUNE 2026

# When the Instrument Builds Types and the People Refuse Them: Segmentation Falsification in a Large AI-Moderated Interview Study

Nicole Zeng,[1] Benjamin Lo,[2] Justin Hoang,[2] and Christian Lansang[2]

[1] *The Observatory* · [2] *Dialogue AI*

Corresponding author: Nicole Zeng (gnezelocin@gmail.com)

**Abstract.** Psychographic segmentation assumes a population divides into stable types. We test that assumption on a dataset purpose-built to satisfy it: 495 AI-moderated interviews with U.S. air travelers, instrumented with a five-item calibration battery and forced-choice items explicitly designed to "create separation between people." Running the study's own prescribed pipeline (clustering the calibration responses toward five to seven segments), we find no natural segmentation: k-means silhouette never exceeds 0.21 at any k from two to eight, Ward linkage and the Bayesian information criterion concur, and adding the instrument's further variables lowers separation rather than raising it. Principal components reveal a continuous two-dimensional structure (a stakes-intensity dimension and a delegate-trust-versus-guard-grieve dimension) explaining 56% of variance, with one soft density pole and no discrete boundaries. Qualitative coding (plan-blind, three-phase, to saturation) explains the null: the dominant patterns are cross-cutting rules shared at 57 to 70% prevalence, while the differentiating variation is a within-person "movable dial" that shifts by trip context, so the unit that varies is the trip, not the person. We frame the methodological contribution (segmentation falsification under a segmentation-optimized instrument), correct the study's a-priori hypothesis (the felt injury during disruption is uncertainty, "being left in the dark," more than powerlessness), and discuss implications for survey-based and AI-moderated segmentation practice. The contribution is descriptive, not causal.

**Keywords:** market segmentation, psychographics, cluster validity, attitude-behavior gap, AI-moderated interviews, mixed methods, interview quality

# 1. Introduction

Segmentation is the default deliverable of consumer research: a population is partitioned into a small number of types, each with a profile and a name, and decisions are made per type. The approach presumes the partition exists in the data. We report a case in which it does not, and in which the absence is the result.

**Contribution.** This paper contributes a methodological finding: a study explicitly designed and instrumented to produce a psychographic segmentation can falsify its own premise, and we show how to detect that honestly rather than imposing a typology the data will not carry. We further contribute a within-person account of *why* such instruments fail (contextual dial-shifting that places within-person variance on par with between-person variance) and a corrected substantive hypothesis for the studied domain. The contribution is descriptive and bounded to the sample; we make no causal or population-level claim.

# 2. Background

Cluster-based market segmentation has a long methodological literature on whether recovered clusters are natural or imposed (Wedel and Kamakura, 2000). Internal-validity indices such as the silhouette coefficient (Rousseeuw, 1987) and model-selection criteria such as the Bayesian information criterion (Schwarz, 1978) provide standard tests for whether a partition reflects structure or merely tessellates a continuum. Separately, the gap between stated attitudes and enacted behavior is among the oldest findings in social science (LaPiere, 1934), and the dramaturgical account of self-presentation (Goffman, 1959) predicts that respondents perform stances under questioning that they do not hold across contexts. Our results sit at the intersection: a calibration instrument elicits stated stances that neither cluster nor predict the behavior recounted in the same interview. Prior work on AI-moderated interviewing is now developed enough to anchor our measurement contribution. Geiecke and Jaravel (2024) build an open-source platform that runs single-agent LLM interviews at scale and validate it with respondent-based quality metrics, reporting performance comparable to human experts. Chopra and Haaland (2023) field 381 AI-led interviews and show the elicited responses predict behavior months later, while noting that the substantive mental models surface in later probes rather than in top-of-mind answers, a pattern that mirrors the say-do gap we observe. Both establish that AI-led interviews can be high quality and behaviorally valid at scale. Neither asks whether a calibration battery embedded in such interviews recovers a latent segmentation, which is our question, and our Interview Quality Score extends their respondent- and expert-rated checks into a per-interview, three-component instrument (signal, moderator, usability).

# 3. Methods

Data are 495 AI-moderated interviews (three fielding rounds of 205, 96, and 194; round retained as metadata) with U.S. adults who had flown in the prior 24 months. Each interview ran a fixed 10 to 12 minute script (self-image, choice process, a cheap-versus-predictable tradeoff, a disruption narrative, a control-versus-delegation probe, a fairness probe) and a calibration tail of two forced-choice items and five 1 to 7 agreement ratings. Participants were anonymized P001 to P495; the recurring AI moderator was separated from participant turns, and only participant turns served as evidence.

**Quantitative.** The five calibration ratings (complete cases n = 365) were standardized and submitted to k-means (k = 2 to 8, 20 restarts), Ward agglomerative clustering, and Gaussian-mixture/BIC, with the silhouette coefficient as the cluster-validity index, then to principal-components analysis. A richer feature set adding the two forced-choice items (18 features) was tested identically.

**Qualitative.** Coding was three-phase (open, axial, synthesis) and run blind to the study's predictions, which were quarantined in a priors ledger and scored only after the codebook stabilized. A stratified deep-read sample (n = 18, spanning the latent dimensions, frustration choice, round, and signal quality) reached code saturation (under 5% new codes across the final reads); the stabilized codebook was applied across all 486 usable transcripts for prevalence, reported as presence floors.

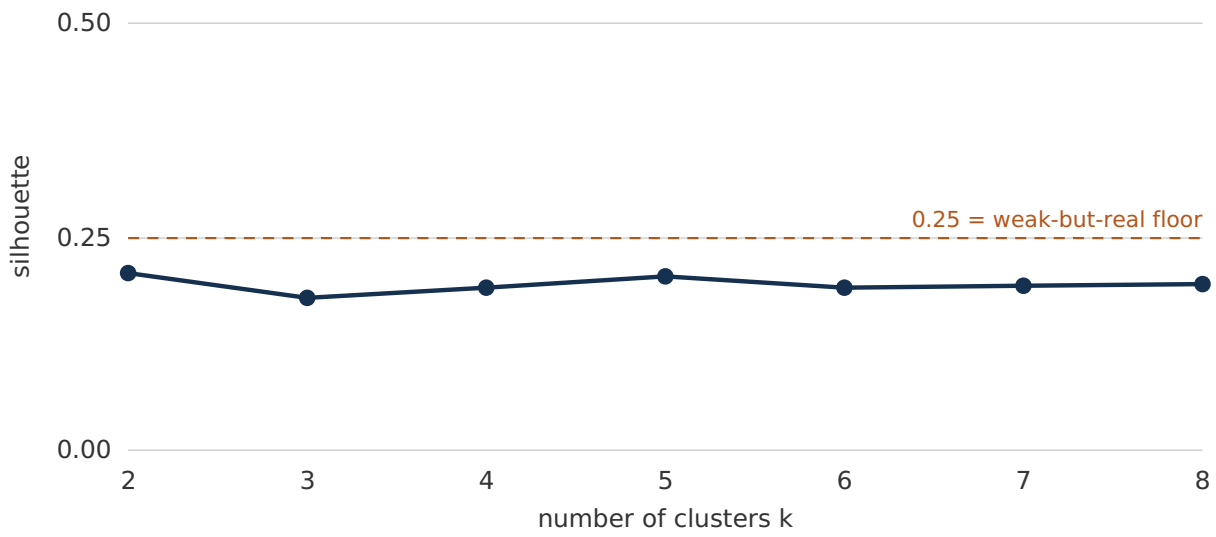
**Credibility.** Each interview was scored on an Interview Quality Score with components Signal Depth, Moderator Effectiveness, and Data Usability (weights 0.60 / 0.15 / 0.25; moderator down-weighted because the fixed AI script is uniform, SD 0.55). Study-level composite mean was 3.76 (n = 486 scored, 9 dead/void excluded). Authenticity used behavioral signals (response-latency rhythm from transcript timestamps, self-correction, idiosyncrasy) alongside textual signals, with a language-register caveat applied to non-native-English speakers.

# 4. Results

**4.1 No natural segmentation.** Across k = 2 to 8, k-means silhouette peaked at 0.208 (k = 2) and sat at 0.204 at the study's target of k = 5, never approaching the 0.25 weak-but-real floor (Exhibit 1, Table 1). Ward linkage agreed (0.13 to 0.17); BIC declined without a clean elbow. Adding the forced-choice items lowered silhouette to 0.12 to 0.17. PCA returned a continuous structure: PC1 (34% variance) loaded positively on all five items (a stakes-intensity dimension); PC2 (22%) contrasted options-stress and trust-airline against fees-anger and trust-recovery (a delegate-trust-versus-guard-grieve dimension: high scorers guard against the airline and keep grievance score); cumulative variance 56%, with one soft density pole and no gaps. An imposed median-split overlay

yields four near-equal quadrants (27/26/24/23%), a spacing consistent with a gradient rather than clusters.

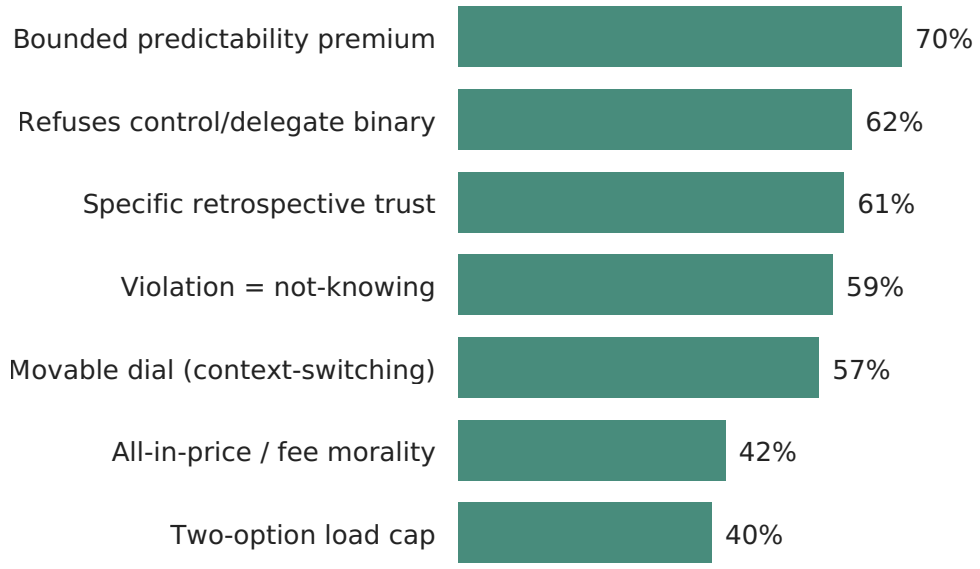
**Exhibit 1. Cluster validity by k (k-means on 5 calibration items, n = 365)**




Silhouette never reaches the 0.25 weak-but-real floor at any k. The study targeted 5 to 7 segments.

**4.2 A shared rulebook explains the null.** The most prevalent qualitative codes are cross-cutting rules, not segment markers (Exhibit 2): a bounded predictability premium (70%), refusal of the control-versus-delegate binary (62%), specific retrospective brand trust (61%), the disruption violation located in not-knowing (59%), and a within-person movable dial that shifts stance by trip context (57%). Two calibration items ceiling out (hidden-fee anger mean 6.09, trust-recovery 5.65), carrying little separating power; only options-stress and trust-airline show wide spread (Table 2).

**Exhibit 2. Prevalence of the shared rulebook (presence floors, n = 486)**




The dominant patterns are shared rules, not segment-defining traits. Floors undercount diversely-phrased codes.

**4.3 The differentiating variation is within-person.** Coded narratives show the same participant adopting different stances by companions (solo versus children), route (domestic versus long-haul), and budget state. Demographic cuts do not rescue a partition: the refused-binary rate is identical (61%) among frequent and infrequent flyers, and options-stress is flat across age bands. A stated-versus-revealed gap compounds this: 232 participants endorsed "the cheapest flight that works" and 263 "pay more to avoid hassle" (all 495 answered this forced-choice item, so the two sum to the full sample), yet narrated behavior favored paying up even among self-described price-firsts.

**Table 2. Calibration distributions (1 to 7)**

| Item | n | mean |
|---|---|---|
| Pay more to avoid uncertainty | 425 | 5.47 |
| Too many options = stressful | 410 | 3.62 |
| Trust airline to decide for me | 405 | 3.69 |
| Hidden fees bother me > upfront price | 392 | 6.09 |
| Once trust lost, hard to win back | 384 | 5.65 |

Source: package.json/distributions. Two items (fees, trust-recovery) are ceiling-pinned.

# 5. Discussion

The instrument did what segmentation instruments are built to do (force stances) and the population declined. The substantive lesson for the domain: the study's a-priori hypothesis ("travelers hate being made powerless") is better stated as a plural injury in which uncertainty, being left in the dark, leads a field that also includes disrespect and control-loss; "control" itself often functions as a dignity signal rather than an agency demand. The methodological lesson is more general: when a calibration battery converges and a clustering pass returns sub-0.21 silhouettes, the honest deliverable is the gradient and its rules, not an imposed typology. Where a navigational typology is required for action, it should be walled behind its validity caveat (here, roughly 57% individual-assignment reliability) and never presented as natural structure.

# 6. Limitations

The study is cross-sectional and qualitative-led; we claim association and description, not causation or population generalization beyond the sample. Calibration ratings are self-report. Two of the plan's named axes (control, price/certainty) were not asked as forced-choice in the delivered script and exist only as coded signal. Qualitative prevalence figures are presence floors, not survey incidence, and diversely-phrased codes are undercounted. The AI moderator, though competent, skipped the mandated consistency probe in 58% of interviews and was verbose enough to out-talk participants in roughly a third, introducing a mild leading risk tracked as low-confidence evidence. Cluster-validity conclusions are sample-specific.

# 7. Conclusion

A segmentation-optimized study can, and here did, falsify its own segmentation premise. The defensible contribution is to detect that honestly and to report the structure that is actually present: a shared rulebook on two continuous dials whose settings move within a person by context. We offer this as a descriptive and methodological result, bounded by its sample and its design.

## References

Chopra, F., & Haaland, I. (2023). Conducting Qualitative Interviews with AI. *CESifo Working Paper No. 10666.*

Geiecke, F., & Jaravel, X. (2024). Conversations at Scale: Robust AI-led Interviews. *SSRN Working Paper 4974382* (forthcoming, *Review of Economic Studies*).

Goffman, E. (1959). *The Presentation of Self in Everyday Life.* Doubleday.

LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces, 13*(2), 230–237.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.). Kluwer.

# Appendix A. Interview protocol (summary)

Probe areas and intent (not the verbatim script): travel self-image; flight-choice process and avoidance tradeoffs; a forced cheap-versus-predictable choice; a disruption narrative with an "annoying to not-okay" turning-point probe; control-versus-airline-handles-it; fairness and switching threshold. Calibration tail: two forced-choice items (most-frustrating problem; loyalty reason) and five 1 to 7 agreement ratings. Note: control and price/certainty were not asked as forced-choice in the delivered script.

# Appendix B. Codebook (summary)

| Code | Prevalence floor |
|---|---|
| G. Bounded predictability premium | 70% |
| A. Refuses control/delegate binary | 62% |
| H. Specific retrospective trust | 61% |
| C. Violation = not-knowing | 59% |
| B. Movable dial (context-switching) | 57% |
| E. All-in-price / fee morality | 42% |
| J. Two-option load cap | 40% |
| F. Options-as-care | 24% (undercounted) |
| D. Fault ledger | 22% (undercounted) |
| I. Role-contingent identity | 10% (undercounted) |

Presence floors, n = 486; not survey incidence.

# Appendix C. Interview Quality Score (summary)

| Component | Weight | Mean |
|---|---|---|
| Signal Depth | 0.60 | 3.71 |
| Moderator Effectiveness | 0.15 | 3.28 |
| Data Usability | 0.25 | 4.18 |
| Composite | — | 3.76 |

n = 486 scored; tiers High 230 / Good 138 / Thin 84 / Low 34; Dead 9 excluded. Flags: uniform-latency 6, straightline 2.

# Appendix D. Representative verbatims (curated selection, exact and located)

A confidentiality-safe selection; full transcripts are withheld to protect participant privacy. Quotes are exact and located (P## + timestamp).

P480 [12:42]: "I wanna say both. I want like the airline to take care of it for me, but I also want to be presented with options."

P490 [09:47]: "I just would like a frank discussion and just tell us the truth, instead of just keeping everyone in the dark."

P490 [11:02]: "It was an act of God. It was bad weather, so it had nothing to do with my choice of flights. So it was pretty irrelevant." (rates trust 1/7 over the communication failure)

P480 [13:41]: "Anything more than three options, that becomes overwhelming. Just give us two options."

P445 [15:48]: "I know what I'm getting with Delta."

# Appendix E. Demographics

| Dimension | Distribution (n = 495) |
|---|---|
| Gender | Female 251 / Male 242 / Other 2 |
| Age | 18-24: 52 · 25-34: 163 · 35-44: 145 · 45-54: 84 · 55-64: 42 · 65+: 9 |
| Flights / 2 yr | 2-4x: 202 · 5-9x: 176 · 10+x: 85 · 1x: 32 |
| Airline type | Major 342 · Mix 124 · Low-cost 29 |
| Stated booking priority | Timing 181 · Lowest price 168 · Comfort 76 · Points 63 · Don't choose 7 |