

The Moderator You Don't See: How AI-Moderator Execution Quality Changes the Texture, but Not the Conclusions, of Interview Research

Nicole Zeng, Benjamin Lo, and Justin Hoang

The Observatory × Dialogue AI

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

Abstract

AI-moderated interviewing lets a single interviewer conduct hundreds of conversations under one fixed script, which turns a variable that human research could never isolate, the moderator, into a controllable one. We exploit this to ask a question qualitative methodology has not been able to answer at scale: when the moderator itself varies in execution, does what we learn change? We analyze 495 airline-traveler interviews (486 with usable participant speech) run by one AI moderator. We extract objective moderator behaviors from transcripts, stratify interviews by execution quality, and re-derive the parent study's substantive outputs within each stratum, pre-registering what would count as a changed conclusion. We find a sharp asymmetry. Calibration means and the study's two-dimensional attitudinal structure reproduce across strata (largest standardized difference $d=0.26$, below the detectable threshold), and the leading themes are stable, while the texture of responses (length, concrete referents, code breadth) varies substantially with execution quality. We further report a negative methodological result: no simple behavioral count (moderator word share, turn count, restatement rate, or a holistic rating) validly grades an AI moderator, because each is either mechanically or judgmentally confounded. We argue that AI-moderator execution is a volume knob on richness, not a steering wheel on findings, and draw the practical implication for research teams: trust headline numbers across execution quality, but read transcripts to protect the load-bearing verbatims, the layer that actually varies.

Keywords: AI-moderated interviews, large language models, qualitative methods, research quality, interviewer effects, methodological rigor, measurement

1. Introduction

Human interview research has always carried an uncontrolled variable at its center. The moderator shapes what a participant says, and no two moderators are the same, so interviewer effects are estimated and worried over but never cleanly isolated: you cannot run the same interviewer, on the same script, across hundreds of participants and vary only how well they executed. AI-moderated

interviewing changes this. A large language model running a fixed script is, for the first time, a moderator you can hold approximately constant and then inspect for the ways it still varies.

Recent work has established that AI can conduct interviews at scale and has compared AI to human moderators, generally finding AI usable but lower in richness and thematic diversity. That literature asks whether to use an AI moderator. It does not ask the operational question a research team faces once it has decided to: given that the AI's execution varies from interview to interview, does that variation change the answer, or only its texture?

Contribution. This paper contributes, first, a descriptive result: across a fleet of 486 interviews run by one AI moderator, execution quality moved the texture of responses (length, specificity, thematic breadth) but not the study's substantive conclusions (calibration levels, attitudinal structure, leading themes), an asymmetry we state at the descriptive level the design supports and not as a causal claim. Second, we report a negative methodological finding with a direct practical consequence: simple behavioral counts do not validly grade an AI moderator (each is either mechanically or judgmentally confounded), so a team cannot cheaply rank its interviews by moderator quality, and the obvious counts would mis-rank them.

Section 2 positions the work; Section 3 gives the design and the pre-registered decision rules; Section 4 reports the asymmetry and the measurement result; Sections 5 to 7 discuss implications and the load-bearing limitations.

2. Related work

A fast-growing body of work treats the LLM as an interviewer. Jack, Cooper and Flower [1] examine generative-AI chatbots automating the qualitative interview, and platform and protocol papers such as AIInterviewer [4] and AI Conversational Interviewing [2] show that adaptive LLM interviewers can be fielded to many respondents and can, in places, match trained human interviewers on response quality while showing systematic limits elsewhere. A parallel caution comes from work examining LLMs as qualitative research participants [3], which warns against mistaking fluent, generic output for lived detail. Commentary has celebrated the scale this unlocks [5]. Across this literature the comparison is AI-versus-human, and the recurring empirical finding is that AI interviewers tend to yield lower richness and narrower thematic diversity than skilled humans.

What is missing is the within-AI question. Once a team commits to AI moderation, the relevant variance is no longer between an AI and a human but between one AI interview and the next. Because the moderator is now a controllable artifact, that variance is measurable in a way human interviewer effects never were. We take up that gap, and we borrow three tools from measurement and statistics to keep the analysis honest: the halo effect [7] and common-method bias [8] to reason about why a holistic quality rating cannot be trusted against a co-rated outcome, false-discovery-rate control [6] for the many-themes comparison, and Tucker's congruence coefficient [9] with a random-split null to test whether a latent structure truly reproduces. The interviewing craft the AI is being graded against is Spradley's [10].

3. Methods

3.1 Design and corpus

We conducted a secondary, within-moderator analysis of the "Traveler Under Pressure" corpus: 495 interviews with United States airline travelers, all moderated by a single fixed AI script ("Dialogue," self-introducing as "George"), fielded in three rounds (R1=205, R2=96, R3=194). The unit of analysis is the interview; nine dead or void transcripts (no usable participant speech) were excluded, leaving 486. Approximately 493 participants are unique; two appear in both R1 and R2 and are retained as separate interviews. The script covered flyer self-concept, a booking walk-through, a two-option tradeoff scenario, a disruption narrative, a control-versus-delegation module, five embedded 1 to 7 calibration items, and a script-mandated consistency probe that re-tested an early-stated preference against a later scenario.

3.2 Moderator-behavior extraction

We parsed each transcript into moderator and participant turns and computed 63 per-interview features. Candidate moderator-quality signals included moderator word share, turn count, verbosity per turn, reflective-restatement rate, and whether the consistency probe fired (a strict regex requiring a callback to a stated preference plus a re-test). Participant-side outcomes were computed length-normalized where possible (self-corrections per 100 words, lexical diversity as Guiraud's R, concrete referents per 100 words) alongside code breadth (number of the parent study's ten emergent codes an interview triggered).

3.3 Pre-registered decision rules

Before inspecting any moderator-by-outcome relationship, we registered what would count as a changed conclusion. Calibration means were deemed robust if no construct's low-versus-high stratum means differed by at least 0.5 scale points or a standardized effect of 0.30; theme prevalence robust if the top-three themes matched and no code moved by at least 10 points after Benjamini-Hochberg correction [6]; the attitudinal structure robust if a within-stratum principal-components solution reproduced the pooled one with Tucker congruence [9] at or above 0.90. Every robustness verdict is reported with its minimum detectable effect (MDE), so a null is never confused with an underpowered test.

3.4 Confound handling

Three confounds are handled explicitly rather than assumed away. Moderator word share is mechanically the inverse of participant volume, so it is never paired with a volume outcome. The holistic quality rating (the moderator component of the study's Interview Quality Score) was assigned by the same analyst who rated response quality, so by the halo and common-method logic [7,8] it is treated as an upper bound on apparent influence, not a clean measure. Because moderator behavior is reactive to the participant, all claims are associational and causal language is barred. Finally, the study was fielded in three rounds under a moderator prompt that was revised between rounds, so round is confounded with prompt version; we therefore make only within-round and by-moderator-quality comparisons and treat any cross-wave contrast (including changes in scripted-probe frequency) as out of scope.

4. Results

4.1 The asymmetry: conclusions hold, texture moves

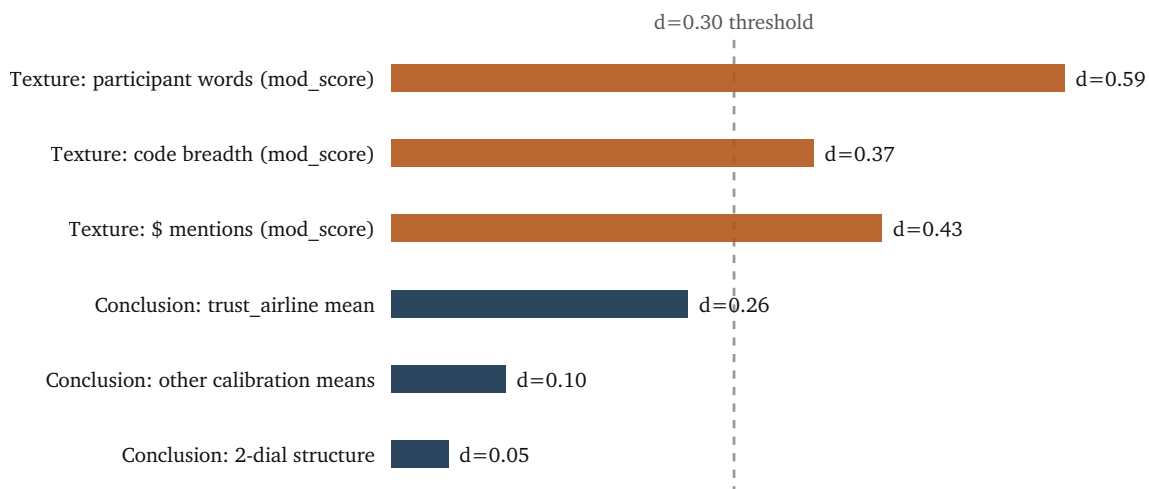
Calibration means reproduced across the moderator-quality range. On the holistic-quality split, no construct crossed the pre-registered threshold; the largest gap was trust in the airline ($d = +0.26$), still below its detectable effect (Table 1). The two-dimensional attitudinal structure also reproduced: the low-versus-high congruence was 0.896 on the first component and 0.761 on the second, and although the second value looks unstable, a 200-replication random-split null of the same cell sizes placed its fifth percentile at 0.485, so the wobble is sampling noise of a weak component, not a moderation effect. By contrast, texture varied: higher-rated interviews were longer, named more dollar figures, and triggered more codes (Figure 1). The texture effects are real but confounded (richer transcripts read as better moderated), whereas the conclusion stability holds on the clean, within-round axis as well.

Table 1. Calibration means by moderator-quality stratum (holistic split; low $n = 164$, high $n = 199$).

Construct (1 to 7)	Low	High	Diff	d	MDE (d)	Verdict
Options stress	3.58	3.56	-0.02	-0.01	0.319	robust
Trust in airline	3.44	3.85	+0.41	+0.26	0.321	robust
Pay uncertainty	5.58	5.45	-0.13	-0.09	0.315	robust
Fees anger	6.23	6.01	-0.22	-0.15	0.326	robust
Trust in recovery	5.67	5.68	+0.01	+0.01	0.328	robust

No construct crosses 0.5 points or $d = 0.30$. Source: analysis/results.json (conclusions_calibration).

Figure 1. The asymmetry: standardized effect sizes, texture versus conclusions (holistic split).



Texture outcomes (sienna) move; conclusion outcomes (navy) sit at or below the $d = 0.30$ detectability line. Source: analysis/results.json (texture, conclusions_calibration, conclusions_pca).

Theme prevalence was the one place the two axes diverged. On the clean axis (the scripted consistency probe present versus absent, within Round 1 only, where the prompt was constant), the

leading themes were stable and no code survived correction; honestly stated, this is robust to prevalence shifts larger than about 20 points, with smaller shifts unresolved at this subsample size (MDE 0.196). On the halo-confounded axis, two mid-tier codes shifted and survived correction, but higher-rated interviews triggered more codes overall (breadth $d = +0.37$), so the shift tracks interview richness rather than the moderator.

4.2 *The moderator does not admit a simple scorecard*

Every low-cost behavioral proxy failed a validity check. Moderator word share correlates with participant word count at $r = -0.88$ (it is largely the inverse of participant volume). Turn count is inversely endogenous: interviews with more moderator turns had fewer participant words ($d = -0.86$), so the count marks a moderator working harder on a thinner participant, not a better interview. Restatement rate proved ambiguous in valence: a hand-read of high- and low-restatement transcripts showed high restatement to be engaged mirroring, not leading. And the holistic quality rating correlates with the co-rated response-quality score at $r = 0.48$, the halo the design anticipated [7,8]. No single clean quality dial exists; the two least-compromised axes each retain one confound (the scripted probe is clean only within a fixed prompt, so it is narrow; the holistic score carries halo). The practical consequence is direct: a team that ranked its interviews by the obvious counts (words, length, turns) would mis-rank them, because those counts track the participant more than the moderator.

5. Discussion

The results resolve a practical question into an asymmetry. For decisions that rest on prevalence, calibration, and structure, AI-moderator execution behaved as fungible: the sloppy interviews and the pristine ones pointed the same way. For the parts of research that depend on the individual interview, the vivid, bounded, self-correcting account that a stakeholder actually remembers, execution mattered, because that is the layer that thins when the moderator does less. A single frame captures it: execution is a volume knob on richness, not a steering wheel on findings. The value is visible at the exact script moment, where good moderation converts a flat preference into a bounded one ("it kind of did, because they moved us to Chicago, which was actually going to get us in sooner"), the kind of qualification a perfunctory interview never surfaces.

The measurement result sharpens the practical stakes. Because no cheap count validly grades an AI moderator, a team cannot simply dashboard its way to interview quality; the volume and length statistics that are easy to monitor track the participant, not the moderator. The implication is not more counting but a different allocation of human attention: read transcripts to find and protect the load-bearing verbatims, since that is the layer execution actually moves, and do not spend that attention re-checking headline numbers that are robust to it.

6. Limitations

The findings are bounded in four ways. First, all 486 interviews were run by one competent AI script; there is no genuinely poor moderator in the data, so the results hold within the competent band and do not license "AI moderation is fine" in general. Second, the design is observational and the moderator is reactive to the participant, so all claims are associational. Third, the holistic-axis effects are halo-inflated upper bounds, and the study's quality instrument carries the platform vendor's name, so the

reassuring stability result is held to the higher bar of the within-round and null-baseline checks rather than the holistic score. Fourth, the study was fielded in three rounds under a prompt that was revised between them, so round is confounded with prompt version; we make only within-round and by-moderator-quality comparisons, and any cross-wave contrast (including changes in scripted-probe frequency across rounds) is out of scope and not interpreted as moderator drift.

7. Conclusion

Because an AI moderator is a controllable artifact, we could ask whether its execution changes what a study concludes, and answer at a scale human research cannot reach. Within one competent script, execution moved the texture of interviews but not their conclusions, and no simple count reliably tells the well-run interviews from the rest. The operational implication is modest and concrete: trust the headline numbers across execution quality, and spend transcript-reading attention where it pays off, on the load-bearing quotes that thin when the moderator does less.

References

- [1] Jack, T., Cooper, A., & Flower, L. (2026). Automating the qualitative interview? Using generative-AI chatbots in social science research. *Methodological Innovations*. <https://doi.org/10.1177/20597991261448157>
- [2] Wuttke, A., et al. (2025). AI conversational interviewing: transforming surveys with LLMs as adaptive interviewers. *Proceedings of LaTeCH-CLfL 2025* (also arXiv:2410.01824). [verify arXiv ID and author list before submission]
- [3] Examining large language models as qualitative research participants ("Simulacrum of Stories"). (2024). arXiv:2409.19430. [verify author list and title before submission]
- [4] AInterviewer: a platform for designing and conducting AI-led qualitative interviews. (2025). arXiv:2606.20588. [verify arXiv ID and authors before submission]
- [5] AI can carry out qualitative research at unprecedented scale. (2024). *LSE Impact of Social Sciences Blog*.
- [6] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289-300.
- [7] Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4), 250-256.
- [8] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research. *Journal of Applied Psychology*, 88(5), 879-903.
- [9] Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57-64.
- [10] Spradley, J. P. (1979). *The Ethnographic Interview*. Holt, Rinehart and Winston.
- [11] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.

Appendix A. Interview protocol (probe areas)

Fixed AI script, probe areas and intent (not the verbatim wording): (1) flyer self-concept; (2) recent-booking walk-through with decision-criterion probes; (3) two-option tradeoff scenario (cheaper-but-riskier vs. costlier-but-predictable) with a switching-threshold probe; (4) disruption narrative with a "what did you want from them" probe; (5) options-vs-delegation module; (6) the consistency probe, a callback to an early-stated preference re-tested against a later scenario; (7) five embedded 1 to 7 calibration items and two forced-choice items.

Appendix B. IQS summary (credibility spine)

Table B1. Interview Quality Score, component weights and distribution (n=486 scored; 9 void excluded).

Component	Weight	Mean	SD	Range
Composite IQS	--	3.76	--	1.6-4.7
Signal depth	0.60	3.71	1.06	0.6-5.0
Moderator effectiveness	0.15	3.28	0.55	2.2-4.2
Data usability	0.25	4.18	0.31	1.7-4.3

Moderator weight is down-weighted from the 0.30 default because the single AI script is near-uniform (SD 0.55), the premise of the measurement result in Section 4.2. Signal tiers: High 230, Good 138, Thin 84, Low 34, Dead 9. Source: package.json (iqs).

Appendix C. Curated verbatims (a selection; exact and located)

Responses from well-moderated Round 1 interviews, illustrating the bounded, specific reconsideration that constitutes the texture varying with moderator quality. Full transcripts withheld to protect participant confidentiality; these quotes are exact and located.

P047 (35-44 F, Sioux Falls SD), probe response: "It kind of did, yeah, because they moved us to Chicago for our connecting flight, which was actually going to get us in sooner than the Denver flight."

P136 (55-64 M, Chicago IL), probe response: "Maybe it's an OCD thing with me, but it didn't really change what I needed out of any transportation companies."

P091 (35-44 F, Arlington TX), probe response: "Next time I'm really gonna consider not dealing with a 30 to 40 minute flight within the same state. If I can afford to do the direct."

Appendix D. Sample

Table D1. Fielding rounds.

Round	Interviews	Alive
R1	205	201
R2	96	92
R3	194	193
Total	495	486

US airline travelers, mixed age/gender/geography. Source: package.json (demographics, interviews).