

When the Interviewer Supplies the Construct: Quantifying construct priming in AI-moderated qualitative interviewing

[Author One],¹ [Author Two],¹ and [Author Three]²

¹*Dialogue Research*

²*[Second affiliation]*

Corresponding author: [email]

Abstract

AI-moderated conversational interviewing is scaling rapidly, but the field lacks empirical accounts of how such instruments fail. We report a case in which an AI moderator supplied a study's central construct to participants before they produced it themselves, in 46 of 56 interviews (82.1%), through a single scripted probe delivered at a median of 2 minutes 30 seconds. The study's own design documents prohibited this behaviour explicitly and twice. We show that the failure is measurable in a way its human-moderated equivalent is not: because the moderator is a single invariant agent across all sessions and every turn is timestamped, the provenance of a construct can be established per interview, prompted agreement can be separated from unprompted production, and the resulting bias can be bounded rather than merely acknowledged. Four participants (7.1%) produced the construct's vocabulary spontaneously; two objected to the moderator's framing inside the interview. We propose a provenance-based analytic response in which affirmative responses to a primed frame are treated as low-confidence and refusals of that frame as high-confidence, and we demonstrate that the study's substantive conclusions survive under this restriction. The demonstration case concerns reciprocity: contrary to the study's stated hypothesis, participants did not describe an internalised ledger of obligation, and of 51 participants asked directly how they decide whether to request help, 41 stated a capability rule and one named debt avoidance. We argue that construct-provenance auditing should be a standard reporting requirement for AI-moderated qualitative work, and that the auditability of these instruments is a methodological advantage rather than only a liability.

Keywords: AI-moderated interviewing, conversational agents, demand characteristics, construct priming, qualitative methods, interviewer effects, research instrument validity, reciprocity

1. Introduction

Conversational agents are now used to conduct qualitative interviews at a scale that human moderation cannot reach. Recent work has demonstrated that AI-conducted interviews can elicit reasoning and mental models that closed-ended surveys miss, and that participants rate the experience

comparably to conventional survey modes (Wuttke et al., 2026). The methodological literature on what these instruments do *wrong*, however, is thin relative to the speed of adoption.

This paper reports one failure mode in detail. During a study of reciprocity and help-seeking, an AI moderator introduced the study's core theoretical construct to participants through a scripted probe, before those participants had produced the construct themselves. This occurred in 46 of 56 interviews. Both of the study's design documents prohibited exactly this behaviour, in explicit terms, and the prohibition was reiterated in a pre-fielding revision.

The interest of the case is not that priming occurred. Priming through question wording is among the oldest documented threats in social measurement (Loftus & Palmer, 1974; Schuman & Presser, 1981), and interviewer-attributable variance is a well-mapped component of total survey error (West & Blom, 2017). Orne's account of demand characteristics, as the totality of cues that convey a hypothesis to a participant, describes the mechanism precisely (Orne, 1962), and its contemporary reformulation continues to treat the participant as an active interpreter of what the researcher wants (Corneille & Lush, 2023).

The interest is that in an AI-moderated setting the priming became **measurable at the level of the individual interview**. Three properties of the instrument make this possible, and none of them holds for conventional human-moderated qualitative work:

1. **Moderator invariance.** A single agent conducted all 56 interviews with an identical script, so the priming stimulus is constant and identifiable rather than distributed across a team of human interviewers with idiosyncratic habits.
2. **Complete turn-level records with timestamps.** Every utterance by both parties is recorded and time-indexed, so first use of any term can be attributed to a speaker and located in the interview.
3. **A stated, pre-registered prohibition.** The design documents specified in advance that the moderator must not introduce this vocabulary, which converts an ordinary methodological worry into a testable compliance claim.

Together these permit a procedure we call **construct-provenance auditing**: for a given theoretical construct, determine per interview which party introduced its vocabulary first, and at what point, and partition the evidence accordingly.

1.1 Contribution statement

This paper makes a **methodological contribution of the descriptive and procedural type**. It contributes: (a) an empirical, quantified case of construct priming by an AI moderator, with a violation rate (46/56, 82.1%), an onset distribution (median 2:30), and participant-side detection (2 of 56 objected in session); (b) a construct-provenance auditing procedure that establishes which party introduced a construct, per interview; and (c) an asymmetric-confidence analytic response, in which agreement with a primed frame is treated as low-confidence evidence and refusal of it as high-confidence, with a worked demonstration that a study's substantive conclusions can be rebuilt under that restriction.

We claim no causal estimate of the priming effect. The study contains no unprimed control arm, and the seven interviews in which the probe did not fire are far too few to serve as one. We claim no

generalisation beyond this instrument, this script, and this sample. What we claim is that the failure was detectable, quantifiable, and bounded, and that the same is not true of its human-moderated equivalent.

1.2 Roadmap

Section 2 positions the case against prior work on demand characteristics and interviewer effects. Section 3 gives the design, sample, moderation protocol, and analytic procedure, including the interview quality instrument. Section 4 reports the priming result and then the substantive result that survives it. Section 5 discusses implications for practice and reporting. Section 6 states limitations, which in a paper of this kind are load-bearing.

2. Related work and background

Demand characteristics. Orne (1962) argued that experimental participants are active interpreters of a study's purpose, and that the cues most powerful in shaping behaviour are those which convey the hypothesis effectively without being obvious. The contemporary reformulation by Corneille and Lush (2023) develops a conceptual framework for the subjective experiences that such cues elicit. An interview probe that names the study's central construct is a demand characteristic of an unusually direct kind: it does not merely imply the hypothesis, it supplies its vocabulary.

Question wording and leading questions. Loftus and Palmer (1974) demonstrated that a single verb substitution in a post-event question altered both speed estimates and, a week later, the reported presence of an object that was never shown. Schuman and Presser (1981) established through systematic experiment that form, wording, and context effects are pervasive rather than exceptional in attitude measurement. Both literatures concern closed-ended instruments; the mechanism, that supplied language reorganises subsequent report, transfers directly to open-ended conversational interviewing but has been much less systematically studied there.

Interviewer effects. West and Blom (2017) synthesise the interviewer-effects literature within a total survey error framework, covering variance introduced at recruitment, measurement, and processing. A recurring difficulty in that literature is attribution: separating interviewer-induced variance from respondent variance generally requires interpenetrated sample assignment, which most field studies cannot afford. An AI moderator inverts this problem. Because the interviewer is invariant, between-interviewer variance is eliminated by construction, and any systematic instrument effect is a property of the script rather than of a person, which makes it locatable in text.

AI-moderated interviewing. Wuttke et al. (2026) report AI conversational interviewing at scale (571 participants) and find that transcripts surface reasoning that standard surveys miss, with participant experience ratings at or above conventional modes. Broader reviews of LLM use in qualitative research note both promise and persistent concerns about depth and reliability, and emphasise continued human oversight [CITATION NEEDED: systematic review of LLM validity/reliability in qualitative data analysis; and CHI 2025 study of LLM uses, tensions and intentions in qualitative research]. To our knowledge, the specific failure mode reported here, an AI moderator

supplying a study's central construct against an explicit design prohibition, with a measured violation rate, has not been documented. [CITATION NEEDED: literature check for prior reports of construct priming by conversational agents in qualitative interviewing.]

The substantive domain. The demonstration case concerns reciprocity and the experience of receiving unpaid help, a domain framed classically by Mauss (1925/1990), whose account treats gift exchange as generating obligation rather than discharging it. The study under examination was designed to test whether market logic has displaced this capacity in contemporary US adults.

The gap. The methodological literature establishes that supplied language biases response, and that interviewers introduce error. It does not yet describe how these threats manifest in AI-moderated qualitative interviewing, how large they can be, or what analytic response is available when they occur mid-study and are discovered only at analysis. This paper addresses that gap with a case.

3. Methods

3.1 Design

Single-wave, AI-moderated voice interviews conducted on a commercial research platform (Dialogue) in August 2026. The instrument comprised a seven-question screener followed by a six-section semi-structured interview with a fixed, non-randomised section order. Section order was deliberately fixed because the design required a stated value (Section 1) to be captured before the behavioural episode (Section 2) against which it would be compared.

The moderation protocol specified a per-section turn ceiling (18 moderator turns across the interview), a target runtime of 11 to 13 minutes, a prohibition on restatement, four banned acknowledgment constructions, and a single rationed verbatim mirror per interview. Most relevant here, the protocol stated: *"The moderator never introduces the tally language first."* A pre-fielding revision restated this twice, in the form *"Never introduce the words owe, repay, or debt first."*

3.2 Sample and recruitment

N = 56 US adults aged 25 to 64, panel-recruited. The screener comprised an instruction-check item, a work-context quota item, a multi-select on help received in the past year, an over-claiming item containing a fictitious behaviour ("Hosted a stranger through the 'Kindred Roads' hospitality network", terminating if selected), an income-consistency item shown conditionally, a comfort-speaking gate, and a candour pledge.

Realised composition: 28 men, 27 women, 1 other. Age bands 25-34 (n=18), 35-44 (n=17), 45-54 (n=15), 55-64 (n=6). Fifty-six distinct US cities across 33 states, no metropolitan area appearing more than twice. Median interview duration 11 minutes 57 seconds (range 8:02 to 14:53).

The 55-64 band is materially under-filled relative to the intended quota design, which constrains age comparison (see Section 6).

3.3 Analysis

Transcripts were analysed using a three-phase emergent coding procedure (open, axial, synthesis) conducted on a **participant-only view** of each transcript. Moderator turns were excluded from the evidence base and used only to establish what was asked and to score moderation quality. All 56 interviews were read in full; no sampling was applied at any stage.

A **priors ledger** discipline was applied. Every outcome predicted by the study's design documents was extracted verbatim before coding began, together with a statement of what evidence would disconfirm it, and then closed. The ledger was reopened only after the codebook was complete, at which point each prediction was scored confirmed, disconfirmed, irrelevant, or reframed. This procedure is the study's structural defence against findings that merely restate the plan, and it is what made the priming failure visible: the ledger recorded the "never introduce the tally language" rule as a design commitment before coding, which prompted the provenance check.

3.4 Construct-provenance audit

For the construct of interest (repayment or score-settling obligation), we defined a lexical set: *owe, owed, owing, repay, repayment, pay (it/them/him/her) back, debt, indebted, ledger, tally, keeping score, even the score, even things out, settle up, IOU*. For each interview we then determined, by turn-level scan of the full dialogue, which speaker first produced any member of the set, and at what timestamp.

Interviews were partitioned into: **moderator-first**, **participant-first (spontaneous)**, and **neither**. For moderator-first interviews we additionally extracted the participant's speech *prior to* the first moderator use, as an uncontaminated window, and the participant's immediate response to the priming probe.

The lexical set was deliberately restricted to unambiguous obligation vocabulary. An initial broader set including *even* and *tab* produced false positives ("even though", "a tab open") and was discarded before analysis.

3.5 Interview quality

Every interview was scored on three components (0 to 5): **Participant Signal Depth**, **Moderator Effectiveness**, and **Data Usability**, combined at weights of 50/30/20 into a composite Interview Quality Score (IQS). Weights were agreed before scoring and are reported with the results rather than held as a hidden parameter. Moderator Effectiveness was scored from the full dialogue; the other two components from the participant-only view and the record metadata.

Study-level IQS was 4.10 (range 1.5 to 5.0), with component means of 4.18 (signal), 3.77 (moderator), and 4.41 (usability). No session was dead or void. Nine interviews scored below a composite of 3.0; none anchors a reported finding. Per-interview scores are given in Appendix C.

The platform's own automatically generated quality score was retained for comparison rather than used as an input. It correlates with the computed IQS at $r = 0.66$ ($n = 56$ pairs) and never fell below 3 of 5, including for a transcript containing 208 participant words.

3.6 Verbatim verification

Every quotation reported in this paper was audited against the source transcript on five criteria: existence at the cited timestamp; attribution to the correct participant; production by the participant rather than the moderator; single continuous utterance with no cross-turn compositing; and removal-only trimming with word order preserved. The audit was implemented as a token-subsequence check rather than by inspection. Of 122 quotations in the study's verbatim bank, all 122 passed after ten defects were identified and corrected (four cross-turn composites, two clause reorderings, three word alterations, one incorrect timestamp).

4. Results

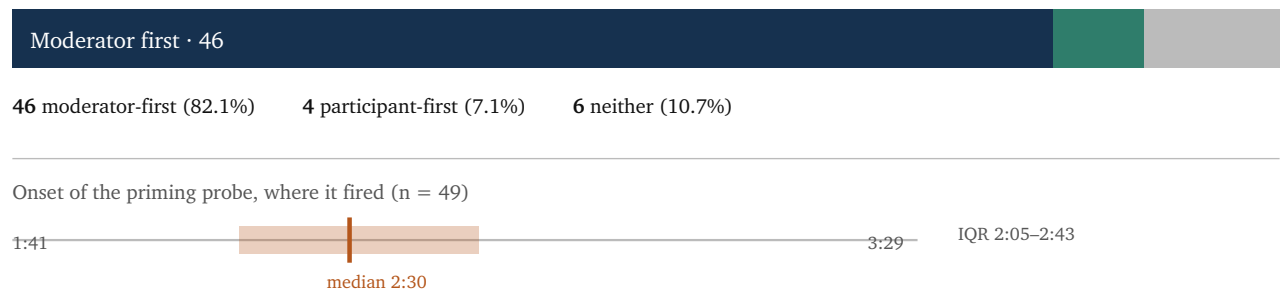
4.1 The priming result

The moderator introduced the study's core construct first in 46 of 56 interviews (82.1%).

Provenance of obligation vocabulary	n	% of 56
Moderator produced it first	46	82.1
Participant produced it first (spontaneous)	4	7.1
Never produced by either party	6	10.7

Figure 1. Provenance of the obligation construct, by interview (N = 56).

Which party produced the construct's vocabulary first



Source: turn-level provenance scan of all 56 transcripts (study number ledger). The probe did not fire in 7 interviews. Onset measured from interview start.

The mechanism was a single scripted probe, delivered in near-identical wording across interviews immediately after the participant completed their account of receiving help:

"Did any part of you want to pay it back or even the score?"

The probe fired in 49 of 56 interviews. Onset was tightly clustered: median 2 minutes 30 seconds, interquartile range 2:05 to 2:43, full range 1:41 to 3:29. At the point of firing, participants had produced a median of 156 words (range 57 to 371).

Two consequences follow. First, the uncontaminated window is short, which bounds what can be inferred from the *absence* of the construct in participant speech before the probe. Second, the priming is early and uniform, arriving before the majority of each interview and at a structurally identical moment, which makes it a near-constant treatment rather than a scattered irregularity.

Participant-side detection. Two participants identified the framing as the moderator's and said so during the interview.

"It's not so much even the score. You used the words even the score, so I used it back."
(P33, woman, 55-64, @ 03:23)

"not to even the score, I don't like that expression necessarily, but to just help pay back and show that you're really there for your friends." (P51, woman, 55-64, @ 02:49)

The first is, to our knowledge, an unusually explicit instance of a participant narrating the priming mechanism while inside it. It is also evidence that the cue was salient rather than subliminal, which bears on the Orne (1962) framing: the hypothesis was conveyed effectively and, for at least two participants, obviously.

Uncontaminated pre-probe language. In the window before the probe, the dominant register was gratitude, present in 35 of 56 interviews. Obligation-family concepts were rare: discomfort (3), returning a favour (3), burden (2), guilt (2), debt lexicon (1). Given the short window, these counts establish what came to mind *first*, not what participants were incapable of expressing, and we do not treat them as prevalence estimates.

4.2 The asymmetric-confidence response

We propose the following analytic rule for a primed corpus:

- **Affirmative responses to a primed frame are low-confidence.** A participant agreeing with a construct they were just handed is minimally informative, since agreement is the path of least resistance and is exactly what a demand characteristic produces.
- **Refusals of a primed frame are high-confidence.** Rejecting a frame supplied by the interviewer is effortful and mildly face-threatening. A participant who does it is providing evidence *against* the direction of the demand characteristic, which makes their refusal more trustworthy than an equivalent unprimed statement would be.

This is an asymmetry in the same direction as the priming, and it is what permits a corpus to be salvaged rather than discarded. It restricts what can be claimed: any conclusion requiring participants to *affirm* the primed construct becomes unavailable. Conclusions resting on refusals, on unprimed production, and on questions asked before the probe remain available.

We applied this rule to the study's substantive analysis. The results below are stated under it.

4.3 The substantive result, under restriction

The study's stated hypothesis had two branches: either people knowingly operate distinct gift and market economies and can receive freely in the right context, or market logic is so internalised that unearned generosity produces debt-anxiety and a compulsion to settle. A secondary proposition held that this same obligation accounting is why people have stopped asking for help.

Stated value (asked before any priming could occur). The opening question, delivered in the first minute, asked whether participants think of their own time as something to give or something to protect.

Response	n	% of 56
Protect	27	48.2
Both or mixed	17	30.4
Give	8	14.3
Rejected the binary	2	3.6
Deflected or unclear	2	3.6

The design anticipated a generous public claim ("on a form everyone says 'I give freely'") that later behaviour would contradict. The claim was already defensive at the point of asking. This result is unaffected by the priming, since it precedes the probe by roughly 90 seconds.

Help-seeking (asked after priming, but the response is not an affirmation of the primed frame). Fifty-one participants were asked how they decide whether to request help or handle a task themselves.

Stated decision rule	n	% of 51 asked
Capability ("only when I cannot do it myself")	41	80.4
Cost to the other person	4	7.8
Relationship history	4	7.8
Anticipated obligation	1	2.0
Efficiency	1	2.0

Two cautions are required and we state them prominently because the headline figure invites over-reading. First, **this is the rule participants state, not the rule their behaviour reveals**. Self-sufficiency is close to the only socially available answer to a direct question of this form, so the base rate for such an answer in any population is presumably high, and the 80.4% figure alone supports little. Second, the informative quantity is the asymmetry: in a study designed around obligation, that had supplied obligation vocabulary to 82.1% of the sample a few minutes earlier, **one participant of 51 named obligation as the operative rule**.

Behavioural corroboration is available but not sized. Participants described consequential non-asking: managing postpartum depression without requesting a parent's help, spending six months unemployed without activating a professional network, absorbing a financial fraud in silence, and going without vehicle repair. These are case reports, not a rate.

Where the construct did appear. Obligation is not absent from the corpus. It appears as an outward-directed forecast concerning specific individuals with a history of presenting favours back as claims, described in 11 interviews, and the language recurs closely across unrelated relationships ("hold it over my head"). One participant articulated it as an operating rule for whom to ask:

"if I feel like that person's personality is not the type of person that a pay me back person, then I'll ask them. If I feel like they are a pay me back person, I probably won't ask them anything." (P56, woman, 45-54, @ 06:21)

Another named the forecast and simultaneously disavowed its accuracy:

"I always feel like someone might hold it over my head even though people technically or typically don't do so." (P40, woman, 25-34, @ 04:54)

Genuine internalised ledgers occur and are a minority (4 interviews). A counter-directional pattern also occurs: nine participants described summoning market instruments (a contract, a payment, a rental) to discharge a gift experienced as too heavy, which runs opposite to the colonisation hypothesis. At $n = 9$ we report this as a raw count and a pattern, not as a structural claim, and we note that four participants described the inverse, resenting a market gesture applied to a functioning gift.

Scoring against the priors ledger. Of nine extracted predictions, five were disconfirmed, two confirmed with the mechanism relocated, one confirmed, and one untestable in this data.

4.4 What the priming cost, and what it did not

The priming rendered unavailable any claim of the form "participants report feeling obligated," which was the study's intended primary finding. It did not compromise: the pre-probe stated-value distribution; the help-seeking decision rule (a different question, answered in participants' own terms); the spontaneous-production count; the refusal evidence; or the qualitative characterisation of where obligation does appear.

The study's designed headline was therefore lost. A different and better-supported result survived. We regard this as the ordinary outcome of a well-instrumented failure: the damage is bounded and locatable, so the salvage is principled rather than opportunistic.

5. Discussion

5.1 The auditability argument

The central methodological claim of this paper is not that AI moderators prime participants. It is that **when they do, it can be counted, and when human moderators do the same thing, it generally cannot.**

Human moderators have supplied vocabulary to interview participants for as long as qualitative interviewing has existed. The literature treats this as a known hazard (Orne, 1962; Schuman & Presser, 1981; West & Blom, 2017), and standard mitigations are procedural: training, guides, reflexivity. What

has been largely unavailable is measurement at the level of a specific construct in a specific corpus, because doing so requires complete turn-level records across every session, an invariant interviewer, and a stated prior commitment against which compliance can be assessed. Field studies rarely have any of the three.

AI-moderated instruments supply all three by construction. The failure documented here was therefore not merely detectable but quantifiable to a rate, an onset distribution, and a per-interview provenance assignment. We suggest this reframes the methodological position of these instruments. They introduce a systematic bias risk that scales (one script error affects every interview identically, where a human team's errors are heterogeneous and partially self-cancelling), and they simultaneously provide the instrumentation to detect it.

5.2 Implications for practice

Construct-provenance auditing should be routine. For any study with a central theoretical construct, the audit is inexpensive: define the lexical set, scan turn-level records, report the provenance table. We suggest it belongs in the methods section of AI-moderated qualitative work as a matter of course, in the way that inter-rater reliability is reported for human coding.

Design prohibitions require enforcement paths. The prohibition in this case was stated clearly, in two documents, and reiterated during revision. It failed anyway, because it existed as prose in a planning document with no mechanism connecting it to the runtime script. A term blocklist enforced at build time would have prevented it. The general principle: a methodological constraint that cannot be checked mechanically will be violated at a rate unrelated to the care of the people involved.

Monitor turn count, not only duration. In this study, runtime landed on target (median 11:57 against a 12-minute goal) while the moderator turn ceiling was exceeded in 54 of 56 interviews (median 28 turns against a cap of 18). The instrument met the duration target by shortening turns rather than reducing them. Any dashboard monitoring duration alone would have reported the instrument as healthy.

Automated quality scores may be measuring completion. The platform's own quality score correlated with independently assessed IQS at $r = 0.66$ and over-read every one of the study's weakest interviews. Its failure direction is the dangerous one: it reports adequacy where adequacy is absent.

5.3 Implications for the substantive question

We advance the substantive finding tentatively, as a hypothesis better supported than the one the study set out to test rather than as an established result. In this sample, the accounting participants describe is directed outward, at specific individuals with a demonstrated history of converting favours into claims, rather than inward at a generalised internal ledger. Help-seeking is gated primarily on self-assessed capability and on the social cost of disclosing incapacity, not on anticipated obligation. Both propositions require testing on an uncontaminated instrument before they should be relied on.

The Maussian framing (Mauss, 1925/1990), in which the gift generates obligation, is not contradicted by these data so much as relocated: obligation attaches to a subset of relationships identified by a learned forecast, rather than being a general property of receiving.

6. Limitations

No control condition. The study contains no unprimed arm. The seven interviews in which the probe did not fire differ on other dimensions and are far too few to serve as a comparison group. We therefore report a violation rate and its analytic consequences, and offer **no estimate of the priming effect size**. A properly powered A/B design (primed probe versus open probe) is the obvious next study and we have not run it.

The uncontaminated window is short. Participants produced a median of 156 words before the probe. Low counts of obligation language in that window are weak evidence of its absence from participants' thinking.

The headline decision rule is self-report. The 41-of-51 capability finding is a stated rule with a high presumptive base rate, as we state in Section 4.3. Its corroboration is case-based and unsized.

Qualitative prevalence is a floor. Counts throughout describe the number of interviews in which a pattern appeared in speech. They do not estimate incidence in any population. Absence of a code is weak evidence of absence of the underlying disposition.

Cross-sectional and directional. No causal inference is available. The study was designed as generative, to prospect hypotheses and language, and its own design documents state this.

Sample. Single market (US), single language, panel-recruited, $n = 56$. Panel recruitment is specifically relevant to one reported observation (five participants described paid survey participation as their alternative to requesting help), which is structurally biased by the sampling frame. The 55-64 band ($n = 6$) is under-filled and no age comparison should be drawn.

Verification is against transcripts, not audio. Quotation auditing confirmed speaker, wording, and order against the platform transcript. Prosody, overlap, and interruption were not assessable, and the transcript is itself a machine artifact. One participant's closing section was unintelligible in transcription and was excluded from citation.

Moderator scoring is investigator judgment. Moderator Effectiveness scores were assigned by the analysis team and are reproducible from the transcripts, but a second scorer would differ at the margins. The bimodality of the distribution is robust; individual adjacent scores are not.

Novelty is asserted, not established. We have not conducted a systematic literature search sufficient to establish that construct priming by a conversational agent is undocumented. Section 2 carries an explicit placeholder for that check, which should be completed before submission.

7. Conclusion

An AI moderator supplied a study's central construct to 82.1% of its sample, in violation of an explicitly stated design prohibition, at a median of two and a half minutes into each interview. Two participants noticed and objected on the record.

The contribution is not the failure. It is that the failure was measurable: a rate, an onset distribution, a per-interview provenance assignment, and a principled partition of the evidence into what the priming compromised and what it did not. The study's designed conclusion was lost. A different conclusion, better supported and contrary to the original hypothesis, survived under an

asymmetric-confidence restriction that treats agreement with a primed frame as weak and refusal of it as strong.

We suggest that construct-provenance auditing be adopted as a standard reporting element for AI-moderated qualitative research, and that the field treat the auditability of these instruments as a methodological asset. Human interviewers have been introducing constructs to participants for a century without anyone being able to count it. Here it was counted.

References

Corneille, O., & Lush, P. (2023). Sixty years after Orne's *American Psychologist* article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, 27(1), 83–101. <https://doi.org/10.1177/10888683221104368>

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5), 585–589.

Mauss, M. (1990). *The gift: The form and reason for exchange in archaic societies* (W. D. Halls, Trans.). Routledge. (Original work published 1925)

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Academic Press.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175–211.

Wuttke, A., Lang, M. M., Klamm, C., Würschinger, Q., & Kreuter, F. (2026). *AI conversational interviewing: Scaling up semi-structured and in-depth interviews*. arXiv:2606.20064.

[CITATION NEEDED] Systematic review of LLM validity and reliability in qualitative data analysis (referenced in Section 2). Candidate identified during drafting but not verified to publication record; the author should confirm venue, year, and authorship before submission.

[CITATION NEEDED] CHI 2025 study of LLM uses, tensions, and intentions in qualitative research (referenced in Section 2). A paper of this description appears in the CHI 2025 proceedings; full author list and page range require verification.

[CITATION NEEDED] Prior reports, if any, of construct priming by conversational agents in qualitative interviewing (Section 2 novelty claim). A systematic search is required to support or retract the novelty assertion in Section 1.1.

Appendix A. Interview protocol (summary)

Section	Job	Turn budget	Seed question as fielded
Opening	Set tone	Statement only	Framing statement, no question
1. Value prime	Stated value	Seed + 1 probe	"Do you think of your own time more as something to give or something to protect?"
2. Receiving episode	Narrative and the construct	Seed + 4 probes	"Tell me about the last time someone did something generous for you that you didn't pay for."
3. Conditions	What would ease receiving	Seed + 2 probes	"What would have made it easier to receive?"
4. Asking for help	Whether obligation suppresses asking	Seed + 2 probes	"How do you decide whether to ask someone for help, or just handle it yourself?"
5. Giving contrast	Say-do comparison	Seed + 2 probes	"Do those two line up, or is there a gap?"
6. Frame question	Polarity read	Seed only	"Some people say we've turned time into money so completely that we can't accept it as a gift anymore. Does that ring true for you or not at all?"
Closing	Open floor	Seed only	"Do you have any additional thoughts you'd like to share?"

Note on Section 2. The priming probe reported in Section 4.1 ("Did any part of you want to pay it back or even the score?") was delivered within this section. It does not appear in the design documents in this form, and the documents prohibit its content.

Note on Section 6. The seed question states the study's hypothesis as a proposition and requests assent. It is a leading question as fielded, and responses to it were excluded from the evidence base.

Question delivery rates. Section 2 seed 55/56; Section 3 (any of three) 53/56; Section 4 decision question 51/56; Section 5 mandatory closing question 14/56; Section 6 frame question 55/56.

Appendix B. Codebook summary

Code	Type	Interviews
Invoicing forecast (specific person expected to present a claim)	Decision logic	11
Self-reliance gate on asking	Rule cluster	41 of 51 asked
Standing threat (receiving as status exposure)	Tension pair	Not sized
Refusal ritual (decline, then giver overrides)	Rule cluster	7
Contract as discharge (market instrument summoned)	Decision logic	9
Want versus must (desire to reciprocate distinguished from obligation)	Meaning split	6 explicit
Script availability (equality or hierarchy both license receiving)	Meaning split	Contested, not sized
Routinisation (repetition dissolves both obligation and gratitude)	Mechanism	4
Means-testing the giver	Rule cluster	8
Ask budget (rationing one's own requests)	Decision logic	4
Receiving as a gift to the giver	Reframe	4
Two economies, explicitly zoned	Decision logic	5 explicit
Internalised ledger (the hypothesised pattern)	Decision logic	4
Moderator-supplied construct vocabulary	Instrument artifact	46

Appendix C. Interview quality summary

Weights: Signal 0.50, Moderator 0.30, Usability 0.20. N = 56, no void sessions.

Component	Mean	Range	Distribution (5/4/3/2/1)
Participant Signal Depth	4.18	1–5	26 / 18 / 9 / 2 / 1
Moderator Effectiveness	3.77	1–5	11 / 32 / 3 / 9 / 1
Data Usability	4.41	2–5	30 / 21 / 3 / 2 / 0
Composite IQS	4.10	1.5–5.0	27 high / 16 strong / 4 adequate / 9 thin

Nine interviews below composite 3.0: P02, P06, P12, P20, P26, P27, P35, P36, P42. None anchors a reported finding. Platform-generated quality score: mean 4.43, range 3–5, correlation with computed IQS $r = 0.66$.

Appendix D. Curated verbatim selection

All quotations verified against source per Section 3.6. Participant identifiers are study-internal; no names were collected.

ID	Quotation (trimmed, removal only)	Locator
P33	"It's not so much even the score. You used the words even the score, so I used it back."	03:23
P51	"not to even the score, I don't like that expression necessarily"	02:49
P56	"if I feel like that person's personality is not the type of person that a pay me back person, then I'll ask them."	06:21
P40	"I always feel like someone might hold it over my head even though people technically or typically don't do so."	04:54
P15	"when it comes to receiving help, I do have difficulty... But when I'm giving help, I don't think of any of those things."	11:33
P29	"it's like you're saying that your efforts didn't really meet your expectations"	07:42
P34	"I would have felt better about it if he had made up a contract or some sort of like trade"	04:57
P52	"it would be taking away from them their opportunity to be generous if I withheld the willingness to receive"	04:30
P45	"on some level, everyone keeps an accounting of the transactional nature of assistance to ensure that they're not being taken advantage of"	04:58
P41	"Yes, I definitely mentally keep a note of it"	07:59

Appendix E. Demographics

Dimension	Composition
N	56
Gender	Man 28, Woman 27, Other 1
Age	25-34: 18 · 35-44: 17 · 45-54: 15 · 55-64: 6
Work context	Professional/office 26 · Self-employed 12 · Not working for pay 6 · Service/hourly 3 · Student 3 · Skilled trade 3 · Caregiving 2 · Healthcare hourly 1
Geography	56 distinct US cities, 33 states, no metro more than twice
Duration	Median 11:57, range 8:02–14:53