

Guide Design, Not Moderator Modality, Accounts for Register Variation in Sensitive-Topic Interviews, and Participants Cannot Tell

[AUTHOR NAME],¹ [AUTHOR NAME],² and [AUTHOR NAME]¹

¹*Dialogue AI*

²*[INSTITUTIONAL AFFILIATION]*

Corresponding author: [EMAIL]

Abstract

AI-moderated interviewing is widely expected to elicit franker disclosure on stigmatised topics than human moderation, an expectation grounded in decades of work on computer-administered self-report. We test the assumption directly on a maximally sensitive topic (adult content consumption) and find that it does not survive measurement. Across three corpora, we compare a human-moderated arm and an AI-moderated arm running an identical guide (Study 01, $n=21$ and $n=57$) against an AI arm running a guide rewritten for vocabulary elicitation (Study 02, $n=89$). Using an external, published lexicon applied wholesale and rarefied to equal corpus size, explicit-vocabulary breadth is statistically indistinguishable between the human and AI arms of the identical guide (34.0 versus 32.8 types), while the redesigned guide more than doubles it (72.5 types, 95% interval 63 to 82). The share of sexual-domain vocabulary naming concrete acts rather than taxonomic categories shows the same pattern (3.3%, 3.8%, 17.9%). Within interviews, register tracks which question is asked and not elapsed time or accumulated rapport (first-half versus second-half density 8.60 versus 8.59 per 1,000 words, paired $p=0.996$). Our principal contribution is a dissociation: 64.4% of participants report that the AI moderator changed what they were willing to say, yet those participants use no more explicit vocabulary than participants reporting no difference (5.50 versus 6.61 per 1,000 words, $p=0.489$). The dissociation is strongly gendered in report (women 83.3% versus men 52.0%, Holm-adjusted $p=0.006$) and absent in every measured linguistic feature. We state one boundary condition up front: no human-moderated arm was run on the Study 02 guide, so guide and modality are not fully separable for that study, and closing that cell is the defining design of the follow-up we specify in Section 7.

Keywords: AI-moderated interviewing, register, self-report validity, sensitive topics, qualitative methods, sociolinguistics, social desirability

1. Introduction

A durable finding in survey methodology is that removing a human interlocutor increases disclosure of sensitive information. Weisband and Kiesler's meta-analysis of 39 studies found computer administration raised self-disclosure relative to face-to-face interviewing, with the largest effects on

sensitive items (Weisband & Kiesler, 1996). Tourangeau and Yan's synthesis reaches a compatible conclusion: self-administration reduces misreporting on sensitive questions, and the mechanism is the removal of an audience to whom an impression is owed (Tourangeau & Yan, 2007). More recently, Lucas and colleagues showed that framing a virtual interviewer as automated rather than human-operated lowered evaluation apprehension and raised rated disclosure (Lucas, Gratch, King, & Morency, 2014).

Conversational AI moderation inherits this expectation. If a computer form elicits more than a clipboard, a conversational agent that never raises an eyebrow should elicit more than a person in the room. The applied research industry has largely adopted this reasoning, and it is the premise on which the study reported here was commissioned.

The premise is testable and, as stated, it is wrong. What follows is not a refutation of the disclosure literature, which concerns *whether people report a behaviour at all*. It is a demonstration that a distinct construct, the *register* in which people describe a behaviour they have already agreed to discuss, does not respond to the moderator at all. It responds to the instrument.

We then report a second and, we argue, more consequential result. Participants' own accounts of whether the modality changed their speech bear no statistical relationship to whether their speech actually differed. This places a large body of applied practice on unstable ground, because self-reported comfort is routinely used as evidence that a method elicited franker data.

2. Contribution

We make three claims, ordered by how strongly the evidence supports them.

2.1. Register in sensitive-topic interviews is a property of instrument design, not of moderator modality. Holding the guide constant and swapping a human moderator for an AI produces no measurable change in explicit-vocabulary breadth or kind. Holding the moderator constant and redesigning the guide produces a large one.

2.2. (Principal contribution.) Self-reported register shift is dissociated from measured register shift. Participants who report that the AI changed what they said do not speak measurably differently from those who report it did not. To our knowledge this is the first study to measure both constructs in the same participants on the same topic and report their independence. [CITATION NEEDED: verify no prior study measures self-reported disclosure change and lexical measures of register in the same sample; a literature check in the CHI/CSCW and survey-methods corpora is required before this novelty claim is asserted in submission.]

2.3. The dissociation is gendered. Women in this sample report the modality effect at substantially higher rates than men while producing statistically identical language. We report this as a descriptive result about experience, not as a claim about disclosure behaviour.

3. Related work

Disclosure and mode. The disclosure literature is consistent that self-administration and computer mediation reduce socially desirable responding on sensitive items (Weisband & Kiesler, 1996;

Tourangeau & Yan, 2007), and that anthropomorphic agents framed as automated can extend the effect to conversational settings (Lucas et al., 2014). Our results do not contradict these findings. We measure a different dependent variable: not whether a behaviour is reported, but the lexical register used to describe it once reported.

Register and audience. The sociolinguistic account holds that speakers shift style in response to their audience, and that the linguistic variable, together with its envelope of variation, is the proper unit of analysis rather than raw counts. [CITATION NEEDED: Labov's formulation of the linguistic variable and the envelope of variation; Bell's audience design account. Verify editions and page ranges before submission.] Our design exploits this: the AI moderator is held constant across all 89 interviews, so between-participant differences in interactional conduct are not confounded by interlocutor.

The limits of verbal report. Nisbett and Wilson's demonstration that people confabulate accounts of their own cognitive processes is the closest precedent for our principal finding (Nisbett & Wilson, 1977). Our result extends the concern to a specific applied case: participants' introspective reports about how an interview modality changed their own speech.

4. Methods

4.1 Corpora

Three corpora, all US, English, adults 18 and over recruited by online panel and screened for adult-content use in the prior three months.

Corpus	Moderator	Guide	Interviews	Participant words
Study 01, human arm	Human researcher	Guide A	21	68,031
Study 01, AI arm	AI (single persona)	Guide A	57	74,574
Study 02	AI (same persona)	Guide B	89	202,942

Guide A was a conventional product-experience guide. Guide B was rewritten to create room for participants' own vocabulary, with probes that decline to supply terms ("What would you actually call that?"). Study 02 comprised 89 interviews from 87 unique participants; two participants were interviewed twice across fielding waves, and their second sitting is excluded from all per-participant rates.

4.2 The lexical instrument, and a bias we corrected

Explicit vocabulary is measured against an external, published word list of 403 terms compiled independently of this research and applied *wholesale with no curation*. [CITATION NEEDED: the LDNOOBW word list, canonical repository URL and access date.]

We report a methodological failure of our own because it materially changed the effect size. Our first lexicon was hand-built by the analyst *after* reading Study 02 transcripts. Applied to the three corpora it produced a 2.9-fold guide effect. Re-run on a lexicon derived exclusively from the Study 01 human corpus, the same effect fell to 1.59-fold; on the external list it is 1.89-fold. Roughly half of the

original effect was analyst-construction bias. We report the external-list figures throughout and recommend that any lexical study of register pre-register its instrument or adopt a published one.

4.3 Rarefaction

Vocabulary type counts scale with corpus size, so cross-corpus comparison of breadth requires normalisation. We rarefy by document subsampling: for each corpus we draw documents in random order until the participant-word total reaches 68,031 (the size of the smallest corpus), count distinct types, and repeat 1,000 times. We report the mean and the 5th to 95th percentile interval.

4.4 Envelope-based variationist measures

Two linguistic variables were coded against explicit envelopes rather than as per-word rates, because a per-word rate confounds a speaker's use of a variant with how much they talked.

(*hedge*): envelope is every first-person preference or evaluation predicate (908 tokens, 87 speakers). Variants: hedged (a mitigator within the clause window) versus bald.

(*specificity*): envelope is every participant turn of at least 8 words answering a what-do-you-want question (853 codeable tokens, 683 excluded as neither variant, 89 speakers). Variants: names a concrete act, anatomy or genre, versus uses only a generic placeholder.

This distinction is load-bearing. Measured per 1,000 words, men in our sample hedge more than women (29.7 versus 23.5, $p=0.034$). Measured inside the envelope where hedging is an actual choice, the difference vanishes (50.2% versus 48.8%, $p=0.798$). We report only the envelope figures.

4.5 The register self-report measure

Every interview closed with: *"Did you find yourself saying things you might not have said if a human researcher were sitting across from you?"* All 88 codeable responses were hand-coded into five non-exclusive categories. Hand-coding was necessary: a keyword classifier recovered the effect in 12.4% of interviews where close reading recovered it in 64.4%, an error large enough to have inverted the study's conclusion had it gone unchecked.

4.6 Inference

Wilson score intervals on all proportions. Permutation tests for linguistic rates are computed at the *speaker* level (shuffling speaker labels, not tokens) to avoid pseudoreplication from unequal talk time. Fisher exact tests for categorical contrasts, with Holm-Bonferroni correction across the three gender comparisons. Small-cell rules: no percentage reported below $n=30$ without a directional label, raw counts only below $n=10$, and variationist cells suppressed below 5 speakers. All data are cross-sectional; every association reported is associational.

4.7 Data quality

Each interview was scored on a three-component index (participant signal 50%, moderator effectiveness 30%, data usability 20%), computed independently from the transcript rather than inherited from the platform's own score. Study mean 3.92 of 5 (range 2.79 to 4.81; signal 3.39, moderator 4.33, usability 4.62). No interview was void, out-of-market, or authenticity-capped. Eighteen carry a data-quality flag. Headline findings were re-run excluding flagged interviews and hold (Section 5.4).

5. Results

5.1 Modality does not move vocabulary; the guide does

Corpus	Explicit terms / 1,000 words	Distinct types at 68,031 words	Act-share of sexual vocabulary
Study 01, human	4.41	34.0	3.3% [2.3, 4.7]
Study 01, AI, <i>same guide</i>	4.26	32.8 [31, 34]	3.8% [2.9, 4.9]
Study 02, AI, <i>new guide</i>	8.34	72.5 [63, 82]	17.9% [16.7, 19.1]

Table 1. Lexical measures across the three corpora. External 403-term lexicon applied wholesale. Type counts rarefied by 1,000-draw document subsampling to equal participant-word totals; brackets give the 5th to 95th percentile interval. Act-share brackets are Wilson 95% intervals.

Swapping the moderator while holding the guide constant moves rarefied breadth from 34.0 to 32.8, a difference in the direction opposite to the hypothesis and well within sampling variation. Redesigning the guide while holding the moderator constant moves it to 72.5, with an interval that does not approach the others.

The act-share measure shows that the vocabulary changed in kind, not merely in quantity. Study 01 participants described the domain taxonomically under *both* moderators (most frequent terms: *content*, *videos*, *category*, *amateur*, *professional*). Study 02 participants named acts and anatomy. Terms were classified as taxonomic or act/anatomy by linguistic type before the Study 02 counts were inspected.

5.2 Register tracks the question, not the clock

If register were a product of accumulated rapport, explicit density should rise across an interview. It does not. Splitting each participant's turns at their own midpoint gives 8.60 per 1,000 words in the first half and 8.59 in the second (paired within-speaker permutation, $n=89$, $p=0.996$).

What moves density is the guide block. Against each participant's own opening-block baseline, the vocabulary-elicitation block sits +3.90 per 1,000 words ($p<0.001$) and the closing reflection block sits −2.29 below it ($p=0.024$); the remaining blocks do not differ. This replicates the guide effect *within* interviews, independent of the Study 01 comparison.

5.3 The dissociation

Self-reported response to the closing question	n (of 87)	%	95% CI	Measured explicit density / 1,000 words
The AI let me say things I would have withheld	56	64.4	[53.9, 73.6]	5.50
<i>of which: would not have participated at all</i>	9	10.3	[5.5, 18.5]	—
No difference	20	23.0	[15.4, 32.9]	6.61
Held back even with the AI	5	5.7	[2.5, 12.8]	—
A human would have been easier	6	6.9	[3.2, 14.2]	—

Table 2. Hand-coded register self-report against measured lexical behaviour. Categories are non-exclusive. Difference between the top and third rows: -1.10 per 1,000 words, permutation $p=0.489$.

Nearly two thirds of participants report that the modality changed what they were willing to say. Their measured vocabulary does not differ from that of participants reporting no difference, and the point estimate runs in the opposite direction ($p=0.489$).

The two extremes of the sample illustrate the independence. The participant with the highest explicit density in the corpus (44.53 per 1,000 words) reported no difference: "*Uh, not at all. I'm pretty open and straightforward. Uh, could be, uh, six females in front of me, I'm still blatantly honest.*" A participant who used no explicit terms at all (0.00 per 1,000 words) reported the strongest form of the effect: "*I mean, yeah, of course. I wouldn't talk about any of this with a human researcher.*"

We read the two constructs as follows. The self-report measures *felt permission*, an experience of reduced evaluative threat, consistent with the mechanism identified by Lucas et al. (2014). The lexical measures capture *register*, which our data indicate is elicited by the instrument. They are separable, and applied practice conflates them.

5.4 Gendered in report, null in language

Measure	Women	Men	Test
Reports the AI changed what they said	83.3% (30/36)	52.0% (26/50)	Fisher $p=0.003$, Holm-adj. 0.006
Reports no difference	5.6% (2/36)	36.0% (18/50)	Fisher $p=0.0008$, Holm-adj. 0.002
Would not have participated	16.7% (6/36)	6.0% (3/50)	Fisher $p=0.156$, <i>not significant</i>
Explicit terms / 1,000 words	8.01	8.14	null
(hedge), hedged share of envelope	48.8%	50.2%	$p=0.798$, null
(specificity), named share of envelope	37.6%	38.7%	$p=0.846$, null

Table 3. Gender contrasts. The two surviving effects are both self-report. Every measured linguistic feature is null. Excluding flagged interviews, the first row is 82.4% versus 52.2% ($p=0.009$).

The study was designed around the hypothesis that women's unedited vocabulary would differ from men's in kind. It does not differ on any measure we took. Women were marginally *more* likely to use at least one explicit term (94.6% versus 89.8%).

We note the sampling asymmetry explicitly. The realised sample is 51 men and 37 women against a 50/50 target. Undersampling reduces power to *detect* a difference; it does not manufacture a null. Since the gender results are nulls on language and a large surviving effect on report, the shortfall does not threaten either conclusion. It would matter had we claimed a small positive difference.

6. Discussion

For practice. Teams adopting AI moderation to obtain franker data on stigmatised topics should expect the gain to arrive through *participation*, not through vocabulary. Ten per cent of our participants said they would not have taken part with a human present, and that group skews female. If the effect is on who consents to be interviewed, then the methodological benefit is a coverage benefit, and it should be evaluated as one: against non-response bias, not against depth of disclosure.

For measurement. Our principal result is a caution. Self-reported comfort is a poor proxy for changed speech, and the two can be measured in the same participants and shown to diverge. Any study inferring elicitation quality from participants' accounts of their own candour should measure the behaviour as well.

For instrument design. The guide effect replicates within interviews, which locates the mechanism at the level of the individual probe rather than the session. A probe that declines to supply vocabulary and asks the participant to supply their own appears to be the operative move.

7. Limitations, and the study they define

The unmeasured cell. We ran no human-moderated arm on Guide B. We can therefore show that modality alone moved nothing in Study 01, and that Guide B moved a great deal with modality held constant, but we cannot say what Guide B would produce in a human's hands. A full 2×2 crossing moderator modality with guide design, on one topic and one recruit, is the study this paper defines. We state this in the abstract rather than here because it bounds the central claim rather than qualifying it.

Self-report about a counterfactual. Sections 5.3 and 5.4 rest on participants predicting their own behaviour in a situation they did not experience. Section 5.3 is precisely a demonstration that such predictions do not track behaviour, and that caution applies to the self-report measures themselves.

Presence-based coding. Qualitative prevalence in this study is a floor. A participant who did not raise a topic may not have been probed on it. These are not incidence estimates.

Corpus asymmetry. Study 01's human arm is $n=21$ against Study 02's $n=89$, and the Study 01 transcripts carry noisier automatic transcription, which suppresses rare-term recall. That bias runs against the human corpus, and the human corpus nonetheless matches the Study 01 AI arm, so the direction of our finding is not threatened by it.

Sample and generalisation. One market, one language, one AI persona, one topic domain, panel-recruited. Nothing here generalises to other personas, other topics, or other populations without replication.

Thin identity cells. Identity-specific questions were reached by 23 interviews (queer participants) and 3 (trans and non-binary participants). Material from these blocks is reported qualitatively and is

never quantified.

References

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883.

Weisband, S., & Kiesler, S. (1996). Self-disclosure on computer forms: Meta-analysis and implications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)* (pp. 3–10). ACM.

[CITATION NEEDED] Labov, W. The linguistic variable and the envelope of variation. Verify edition and pages before submission.

[CITATION NEEDED] Bell, A. Audience design. Verify journal, volume and pages before submission.

[CITATION NEEDED] LDNOOBW English word list. Verify canonical repository, maintainer and access date before submission.

[CITATION NEEDED] Literature check supporting the novelty claim in Section 2.2, that no prior study measures self-reported disclosure change alongside lexical register measures in the same sample.

Appendix A. Interview protocol (summary)

Block	Probe area and intent	Interviews reaching it
1. Arrival	Vocabulary of habit; unreflective routine narration	89 / 89
2. Finding	Core elicitation; probes decline to supply terms	85 / 89
3. Selection and rejection	Evaluative vocabulary; instant accept and reject	87 / 89
4. Gaps	Unmet need; distance between sought and found	89 / 89
5. Industry	Ecosystem, trust, reputation; lower sensitivity	89 / 89
Identity-specific	Representation; asked only on screener self-identification	23 / 89 and 3 / 89
Closing reflection	Participant awareness of the modality's effect	89 / 89

Twenty-three questions. Median 15 reached per interview (range 10 to 19). Verbatim script withheld; probe intent shown.

Appendix B. Coding scheme (summary)

Measure	Envelope or denominator	Variants / rule
Explicit density	All participant words	External 403-term list, wholesale
Vocabulary breadth	Rarefied to 68,031 words	Distinct types, 1,000 draws
Act-share	All sexual-domain tokens	Taxonomic vs act/anatomy, classified a priori
(hedge)	908 preference predicates	Hedged vs bald
(specificity)	853 want/seek turns	Named vs generic placeholder
Register self-report	87 codeable responses	Hand-coded, five non-exclusive categories

Appendix C. Data quality summary

Component	Weight	Mean	Range
Participant signal	0.50	3.39	2.1 – 4.8
Moderator effectiveness	0.30	4.33	3.0 – 5.0
Data usability	0.20	4.62	3.0 – 5.0
Composite	—	3.92	2.79 – 4.81

89 interviews: 28 high, 49 good, 12 adequate, 0 thin, 0 void. 18 carry a data-quality flag. 0 authenticity-capped.

Appendix D. Sample

Dimension	Composition
Gender	51 men, 37 women, 1 other (target was even; shortfall discussed in 5.4)
Age	34 aged 35–44; 29 aged 25–34; 16 aged 45–54; 6 aged 18–24; 4 aged 55–64
Orientation	62 straight; 13 bisexual; 14 across gay, lesbian, queer, pansexual, asexual and questioning
Use frequency	27 daily; 27 two to four times weekly; 9 five to six times weekly; 26 weekly or less
Participants	87 unique across 89 interviews; 2 interviewed twice, second sitting excluded from rates

Appendix E. Selected verbatims

"I feel like porn sites are very broad, and it's not necessarily about them understanding what I'm looking for, but just me actually knowing what I'm looking for and being able to search that. You know, it's like a skill thing." P02, 35–44, woman

"Uh, not at all. I'm pretty open and straightforward. Uh, could be, uh, six females in front of me, I'm still blatantly honest. I don't care. People jerk off. It's natural, so..." P25, 25–34, man. Highest measured explicit density in the corpus.

"I mean, yeah, of course. I wouldn't talk about any of this with a human researcher." P29, 18–24, man. Zero explicit terms used.

"I probably censored myself a little bit and probably avoided using, uh, more explicit terms." P52, 45–54, woman. Top band for measured explicit vocabulary.

"I know that I would've returned this study if there had been a real human sitting across from me. I would not have been able to get through this study if there had been a human sitting across from me." P87, 45–54, woman

Verbatims are exact and located in the source corpus. Participants are identified by number and broad demographics only.