Pre-Registration Report

Listener Perception of AI-Assisted Conversational Speech: Detection Accuracy, Acoustic Cues, and the Effect of Prior Knowledge

Project: Psyche Lab – Proactivity Tool Study
Date: March 2026
OSF #: *(to be completed upon submission)*
IRB #: *(to be completed)*

Prepared in conjunction with `generate_qualtrics.py`, which programmatically generates the Qualtrics `.qsf` survey file from `study_conditions.csv` and `optimized_valid_samples.csv`. All questionnaire items in this document are directly reflected in that script.

0. Contents

1. Background and Rationale

AI suggestion tools embedded in real-time communication platforms (e.g., Zoom-integrated sentence completion or idea prompting) are increasingly available to speakers during live conversations. Such tools may alter observable characteristics of speech—fluency, elaboration depth, response length, and naturalness—in ways listeners might perceive consciously or unconsciously.

This study examines whether naïve listeners can detect AI suggestion use from audio recordings of real conversational responses, what acoustic/paralinguistic cues drive those judgments, and whether explicit knowledge of AI tool access systematically shifts listener perceptions. All speakers are non-native English speakers (NNS) responding to food-culture questions in a structured meeting context, providing a naturalistic setting where AI suggestion use is plausible and variable.

The study contributes to the emerging literature on AI transparency, human-AI communication, and the social perception of AI-assisted speech, with implications for disclosure norms in AI-mediated communication.

2. Research Questions and Hypotheses

'1 **(Implicit detection):** Can naïve listeners, without knowledge of the AI context, differentiate clips by AI suggestion use level, as reflected in authenticity and external source use ratings?

'2 **(Acoustic cues):** Which speech features—disfluency level, elaboration, fluency—independently predict listener ratings of authenticity and perceived external source use in the blind phase?

'3 **(Knowledge effect):** Does explicit knowledge of AI tool access shift authenticity, speech quality, and AI reliance ratings, and does this shift vary by actual AI use level?

'4 **(Individual differences):** Do prior AI use experience in real-time conversation contexts and attitudes toward conversational AI moderate detection *sensitivity* (Group 2 vs. Group 0 contrast magnitude in Phase 1) and knowledge-induced rating shifts (Phase 1 vs. Phase 2 differences)?

2.1 Directional Hypotheses

H1a In Phase 1 (blind), Group 2 clips will receive lower authenticity ratings than Group 0 clips, controlling for disfluency level.

H1b In Phase 1 (blind), Group 2 clips will receive higher external source use ratings than

Group 0 clips, controlling for disfluency level.

H2a Higher disfluency level will be positively associated with authenticity ratings in both phases (disfluency as a naturalness cue).

H2b Higher disfluency level will be negatively associated with external source use / AI-reliance ratings in both phases. *(Phase 1 DV: Q4 "How likely using external sources; Phase 2 DV: Q5 /* `AI_detection_likelihood` *composite "How likely relying on AI suggestion tool. Prediction holds across phases: more disfluent speech is perceived as less likely to benefit from AI content assistance.)*

H3a Clips rated in the informed condition will receive lower authenticity ratings than equivalent clips rated in the blind condition. (Estimated between conditions at the stimulus level via counterbalancing; see Section 4.4.)

H3b Clips rated in the informed condition (Phase 2) will receive higher `AI_detection_likelihood` ratings than equivalent clips rated in the blind condition (Phase 1). Operationally: the `phase_condition` main effect on the `AI_detection_likelihood` composite (Q4 Phase 1 / Q5 Phase 2) is expected to be positive (Phase 2 > Phase 1).

H3c The knowledge-induced rating shift will be larger for Group 2 clips than for Group 0 clips (`phase_condition` × `suggestion_use_group` interaction). Directional specification by DV:

- *Authenticity (H3a component):* Phase 2 decrease is larger for Group 2 than Group 0 (credit discounting stronger for clips that actually used AI).
- `AI_detection_likelihood` *(H3b component):* Phase 2 increase is larger for Group 2 than Group 0 (informed listeners attribute AI use more accurately to high-use clips).

H4a Participants with more frequent prior AI use in real-time conversation contexts will show greater implicit detection *sensitivity* in Phase 1 — operationalized as a larger Group 2 vs. Group 0 contrast magnitude on Q3 (Authenticity) and Q4 (External Source Use). Tested as `Demo_AITaskFrequency` row 4 × `suggestion_use_group` interaction in Analysis 3.

H4b More negative attitudes toward real-time conversational AI (Q15g composite; reverse-coded items 2 and 4) will be associated with lower authenticity and higher AI detection ratings in Phase 2, over and above the Phase 1 blind-condition baseline. *Mechanism: negative real-time AI attitude primes motivated detection — participants uncomfortable with conversational AI scrutinize Phase 2 clips more actively for AI cues, amplify-*

ing knowledge utilization. Operationalized as `Demo_AIAttitude` × `phase_condition` *interaction in Analysis 3.*

H5a (Detection sensitivity) Native English speakers (NES) will show greater detection *sensitivity* than non-native English speakers (NNES) in Phase 1 — operationalized as a larger Group 2 vs. Group 0 contrast magnitude on Q3 and Q4. Tested as a `Demo_NativeSpeaker` × `suggestion_use_group` interaction on the Group 2 vs. Group 0 slope in Analysis 3.

H5b (Detection tendency — exploratory) Both NES and NNES are expected to show a significant detection tendency (Group 2 vs. Group 0 main effect), but NES sensitivity is predicted to be larger. A null H5a with significant Group 2 vs. Group 0 contrasts in both groups would indicate cross-linguistic generalizability of detection at equivalent sensitivity. *Note: "detection accuracy is not used; sensitivity refers to the magnitude of rating differentiation between AI-use groups.*

3. Study Design Overview

Design: Two-phase design combining within-person and between-subjects elements.

Each participant completes two phases using *different speaker sets*:

- **Phase 1 (Blind):** 9 audio clips rated without knowledge of AI tool availability. Captures implicit/incidental detection.
- **Phase 2 (Informed):** After an AI reveal, 9 clips from *different speakers* (same 3×3 stimulus space) are rated with an additional explicit AI reliance item.

Critical design point: Phase 1 and Phase 2 do *not* present the same stimuli twice. A direct within-person before/after comparison is therefore not possible. The knowledge effect (H3) is instead estimated *between participants* via counterbalancing: each sample appears in Phase 1 (blind) for one group of participants and in Phase 2 (informed) for a different group. The design is thus **between-subjects at the stimulus level** and within-person only in the sense that the same participant provides data in both phases.

What the within-person aspect contributes:

- Individual response tendencies (scale use, baseline rating level) are controlled via participant random intercepts in all LMMs.
- Individual difference moderators (AI experience, attitudes) are estimable because each participant contributes data to both phases.

Between-subjects factor: Condition assignment (24 conditions: 12 sample pairs × 2 counterbalance groups) determines which sample a participant receives in Phase 1 vs. Phase 2 and randomises clip order within each phase. Equal allocation across conditions is enforced by the Qualtrics Evenly Present Elements randomiser.

4. Stimuli

4.1 Source Corpus

Stimuli are audio recordings of NNS speakers responding to food-culture questions during a structured meeting in which an AI suggestion tool was embedded in Zoom. The corpus was coded for the following variables:

Suggestion use (`suggestion.used.group`): Three-level categorical variable:

- 0 = No use: suggestion ignored entirely
- 1 = Partial use: suggestion partially adopted or modified
- 2 = Full use: suggestion substantially adopted

Total disfluency length: Continuous measure (seconds) of total disfluency duration per clip (restarts, repetitions, filler words, pauses). Overall corpus: $M = 4.62\,\text{s}$, $SD = 4.47$, Range = 0–25 s.

Clip disfluency count: Discrete count of disfluency events per clip. Overall: $M = 3.84$, $SD = 4.45$, Range = 0–21.

Pitch obviousness (`pitch.obvious.level`): Observer-coded categorical variable (none/medium/high). Distribution: 75 none (83%), 12 medium (13%), 3 high (3%). Due to the highly skewed distribution, pitch is included as a control covariate in sensitivity analyses only.

A key corpus observation is that higher AI suggestion use is associated with *lower* disfluency, consistent with AI-generated sentence structures reducing production difficulty (Table 1). Total disfluency length and clip disfluency count are significantly correlated; pitch is independent of both disfluency measures.

Response–AI similarity (`ai_adoption_sim`): Per-clip cosine similarity (bag-of-words) between the speaker's transcribed response (Whisper `base.en` ASR) and the AI suggestion text shown to them for that topic. Computed via `clip_ai_similarity.py`; output saved to `clip_ai_similarity.csv`. Provides a continuous operationalization of AI adoption beyond the coarse three-level group coding. Used in sample pair quality control (see Section 4.4).

Table 1: Disfluency by suggestion use group.

Group	Suggestion Use	*M* Total Disfluency	*M* Disfluency Count
0	No use	5.17 s	3.88
1	Partial use	4.34 s	3.56
2	Full use	3.41 s	2.97

4.2 Stratification Scheme

Stimuli are selected via a 3×3 grid crossing two axes:

Axis 1 — Suggestion use group: 0, 1, 2 (as above).

Axis 2 — Total disfluency length group (within-corpus quantile boundaries):

Table 2: Disfluency length quantile groups used for stratification.

Group	Label	Range	*n*	%	*M*	Mdn
0	Low	0–2 s	99	41.1%	0.75 s	0
1	Medium	3–5 s	64	26.6%	3.92 s	4
2	High	6–25 s	78	32.4%	9.10 s	8

This yields **9 strata** (3×3). One clip is drawn from each stratum per sample, giving each participant 9 clips spanning all combinations of AI use and disfluency.

Rationale for crossing these axes: Suggestion use and disfluency are negatively correlated in the corpus. Without stratification, high-AI-use clips would be systematically more fluent, confounding the two variables. Crossing them in sample construction orthogonalises the two predictors at the sample level, enabling independent estimation of each variable's effect on listener ratings.

4.3 Sample Construction and Speaker Assignment

Within each sample:

- Exactly **3 speakers**, each contributing exactly **3 clips**.
- Each speaker's 3 clips span all 3 suggestion use levels (0, 1, 2).
- The 9 clips cover all 9 strata with no repetition.

This within-speaker structure allows control for individual speaker characteristics (accent, baseline fluency) when estimating suggestion use effects. 100 valid samples meeting these criteria were identified via an optimised sampling algorithm; 24 are used in the study.

4.4 Conditions and Counterbalancing

The study uses **24 conditions** = 12 sample pairs × 2 counterbalance groups:

- 12 pairs of samples are drawn from the 100 valid samples (no sample appears in more than one pair).
- Within each pair, two groups (X, Y) determine phase assignment:
  - Group X: Sample A → Phase 1, Sample B → Phase 2
  - Group Y: Sample B → Phase 1, Sample A → Phase 2

Every sample appears exactly once as Phase 1 and exactly once as Phase 2 across the 24 conditions. With equal condition assignment, each sample accumulates an equal number of blind-condition and informed-condition ratings across participants. This is the mechanism by which the knowledge manipulation is identified: *differences in ratings for the same sample between its Phase 1 and Phase 2 conditions reflect listener knowledge, not speaker differences.* Conditions are assigned with equal probability via the Qualtrics *Evenly Present Elements* randomiser. Within each phase, 9 clip blocks are presented in randomised order (Qualtrics BlockRandomizer).

Implication for analysis: Per-speaker global ratings Q6–Q9 (Phase 1) and Q10–Q13 (Phase 2) are responses to *different speakers* for every individual participant. Aggregate cross-phase differences (e.g. $\overline{Q12} - \overline{Q8}$ for authenticity, $\overline{Q10} - \overline{Q6}$ for quality, $\overline{Q13 \text{ row } 4} - \overline{Q9 \text{ row } 4}$ for AI-activity attribution) are interpretable as knowledge effects only because counterbalancing ensures the same samples contribute equally to both phases across conditions. Individual-level cross-phase difference scores confound the knowledge manipulation with between-sample speaker differences and are therefore not used as primary DVs.

Table 3: The 12 sample pairs used across the 24 conditions.

Pair	Samples	Pair	Samples
1	21 ↔ 50	7	62 ↔ 87
2	75 ↔ 78	8	2 ↔ 63
3	19 ↔ 72	9	55 ↔ 91
4	27 ↔ 43	10	13 ↔ 95
5	47 ↔ 71	11	15 ↔ 99
6	32 ↔ 51	12	70 ↔ 100

Similarity-based pair filtering: Because Phase 1 and Phase 2 samples are drawn from the same clip pool, some pairs inevitably share clip topics where one sample assigns the topic to Group 1 (partial use) and the other to Group 2 (full use). Although these clips

differ in usage group—constituting a valid cross-phase contrast—both responses may still closely mirror the AI suggestion wording if both speakers substantially adopted it. A listener hearing both clips across phases could recognise the content overlap, and this memory-based recognition would artificially deflate authenticity ratings for the later clip independent of actual AI detection. To mitigate this, the `ai_adoption_sim` scores for the two violating clips on each shared topic were averaged (`mean_viol_clip_ai_sim`) and combined with violation severity and speaker overlap into a composite `combined_risk` score. Pairs were ranked by `combined_risk` ascending; only the lowest-risk pairs were eligible for the final 12-pair selection (see `pairs_reranked.csv`).

4.5 Audio Format

Stimuli are MP3 audio files (72 KB–1 MB per clip). Audio-only presentation was chosen over video for the following reasons:

1. All primary coding variables (disfluency, pitch, fluency) are auditory; audio isolates the channel of theoretical interest and eliminates visual confounds (gaze, posture, facial expression).
2. File sizes allow direct upload to the Qualtrics media library, removing external hosting dependencies.
3. Many real-world AI-assisted communication contexts occur in audio or text-mediated settings, supporting ecological validity.

Participants are instructed to use headphones.

5. Participants

Recruitment: Undergraduate students recruited via SONA.

Target N: $N = 216$ (9 participants per condition $\times$ 24 conditions).

5.1 Power Analysis

Power analysis was conducted via simulation (2,000 iterations per cell; `power_analysis.py`) and analytical approximation, covering three primary analysis blocks. Key assumptions: 7-point scale $SD = 1.5$; $\alpha = .05$ two-tailed; within-person correlation across clips $\rho_w = 0.30$ (conservative).

Binding constraint and rationale: A2 (knowledge-effect interaction) requires the largest N at the assumed effect size. However, the interaction test in A2 is conservative: it asks whether knowledge *differentially* amplifies the Group 2 vs. Group 0 gap, which is a harder

Table 4: Required N for 80% power per primary analysis.

Analysis	Test	Assumed Effect	N for 80% Power
A1: Phase 1 detection (H1)	Paired t, G2 vs. G0	$d = 0.30$	**48**
A2: Knowledge interaction (H3)	Phase × Group	$d_{P1} = 0.30$, $d_{P2} = 0.45$	**288**
A3: Moderator correlation (H4a)	Pearson r	$r = 0.20$	**204**

Table 5: Power at recommended N values for primary contrasts.

N	A1 ($d = 0.30$)	A2 ($d_{P1} = 0.30$)	A3 ($r = 0.20$)	A3 ($r = 0.25$)
144 (6/cond)	>.999	.52	.67	.86
192 (8/cond)	>.999	.64	.80	.94
216 (9/cond)	**>.999**	**∼.68**	**∼.84**	**∼.96**
288 (12/cond)	>.999	.80	.94	>.99

test than the main knowledge effect or the main group effect. At $N = 216$, A2 is adequately powered for detection of a moderate interaction ($d_{P1} = 0.40 \rightarrow 87\%$ power at $N = 168$) and is designated **exploratory** if $d_{P1} < 0.35$. A1 and A3 are the pre-registered confirmatory analyses.

If recruiting to $N = 288$ is feasible, this brings A2 to 80% power and A3 ($r = 0.20$) to 94% power; $N = 288$ is the preferred target. $N = 216$ is the minimum pre-registered sample.

Effect size assumptions: $d = 0.30$ for A1 reflects moderate-small listener sensitivity to AI suggestion use from audio cues alone (no visual channel). $r = 0.20$ for A3 reflects a small-to-moderate partial correlation between domain-specific AI experience and detection accuracy, conservative given the specificity of the moderator (real-time conversation AI use only).

5.2 Inclusion and Exclusion Criteria

Inclusion:

- Age 18–25
- Self-reported normal hearing
- Completion of both phases without interruption

Exclusion (applied post-hoc):

All primary analyses are run twice: on the full analytic sample (E1–E4 applied) and with a stricter within-phase completion threshold (all 9 clips rated, no missing items).

Table 6: Post-hoc data exclusion criteria.

Code	Criterion	Scope
E1	`RevealComprehension` $\neq$ option 3 (correct answer)	Phase 2 excluded; Phase 1 retained
E2	Fewer than 9 per-clip rating sets completed in either phase	Full participant excluded
E3	Identical responses to all per-clip items within a phase (straight-lining)	Full participant excluded
E4	Per-clip page dwell time (`PageSubmit`) < 15 s on ≥ 3 clips in either phase	Full participant excluded (insufficient audio engagement)

6. Procedure

6.1 Consent

Participants read a consent form describing the study as examining "how people perceive conversational speech." The true AI-detection purpose is withheld from Phase 1 to preserve the blind condition. Full debriefing is provided after completion. Estimated total duration: 45–60 minutes.

6.2 Phase 1 — Blind Condition

Participants are told that each clip is an excerpt from a separate conversation between two college students brainstorming menu items, that they will hear **two speakers** per clip (one asking, one responding), and that all conversations share the same topic. Participants are explicitly instructed to **focus on and rate the responding speaker only**; all per-clip and per-speaker questions refer to the responding speaker. No information about speakers' language background is disclosed to participants, to prevent demand characteristics that could bias proficiency ratings. Each audio page also displays a reminder: *"Focus on the speaker who is responding to the question."*

The 9 clips are organised into **3 groups of 3** from the same responding speaker; a brief reminder screen appears before each new group. After each clip, questions Q1, Q2, Q3, Q4 are administered. After the third clip of each group, per-speaker questions Q6, Q7, Q8, Q9 are administered (Q9 = activity likelihood, repeated once per speaker group). No mention of AI tools is made at any point during Phase 1.

6.3 AI Reveal

A disclosure screen informs participants that all responding speakers had access to an **AI suggestion tool embedded in their Zoom meeting platform** that automatically displays two types of suggestions: *idea suggestions* (content prompts) and *sentence suggestions* (pre-formulated starters/phrases). Screenshots of each type are displayed vertically at full width. A prominently styled callout box explicitly states that **having access to the tool does not mean the speaker used it**, and that each speaker **may or may not have chosen to follow the suggestions — some may have ignored the tool entirely**.

6.4 Comprehension Check

A neutral recall question confirms the participant processed the reveal:

> *"Before we continue to the next set of clips, which of the following best describes what the speakers had available to them during their conversations?"*

Five options are presented; the correct answer is option 3: *"An AI suggestion tool built into the meeting platform."* The question is framed as a natural recall prompt rather than an explicit test to avoid reactance. Failure ($\neq$ option 3) flags the participant for E1 exclusion post-hoc; the survey does not branch on this response to preserve Phase 1 data.

6.5 Phase 2 — Informed Condition

Phase 2 instructions restate the two-speaker structure and the responding-speaker focus, introduce the AI suggestion tool, and repeat the speaker-grouping structure (3 groups of 3; reminder screen before each group). Participants listen to 9 new clips from different responding speakers, answer Q1–Q5 after each clip, and Q10, Q11, Q12, Q13 after the third clip of each group (Q13 = activity likelihood, per speaker). Q14 is administered once at end of phase. **No per-clip AI reminder banner is displayed**—the informed context is established at the instruction and reminder level; per-speaker questions Q10–Q12 are identically worded to Phase 1 Q6–Q8; Q13 (activity likelihood) parallels Phase 1 Q9, to avoid demand characteristics from repeated AI framing.

6.6 Demographics

Questionnaire Q15a–Q15h is administered at session end.

7. Measures

7.1 Per-Clip Measures — Both Phases

Q1. Speech Impressions Matrix `P{phase}_C{clip}_SpeechQuality`

"How would you describe the speech of this response?"

Scale: 1 (Not at all) – 7 (Very much) | Rows: **Clear**, **Elaborate**, **Fluent**, **Coherent**

The four adjectives each capture a distinct, directly audible dimension:

- **Clear** — acoustic intelligibility and articulateness
- **Elaborate** — depth and richness of content; a primary cue for AI suggestion use (AI-assisted responses tend to be more content-rich)
- **Fluent** — temporal smoothness of delivery; negatively correlated with disfluency (the primary covariate in this study)
- **Coherent** — logical flow and topical relevance; subsumes the dropped Q3 response-alignment item

Q2. Speaker Performance Impression `P{phase}_C{clip}_SpeakerImpression`

"To what extent do the following words describe this speaker's performance in this clip? (1 = not at all, 7 = very much)"

Format: 7-point matrix (consistent with Q1). 3 items (agency, affect, engagement); each analyzed as a standalone secondary DV. Flat is reverse-coded.

#	Item	Dimension	Valence
1	Confident	Agency	+
2	Flat	Affect	−
3	Engaged	Engagement	+

Engaged replaces the dropped Q3 engagement items, capturing behavioral engagement audible from speech prosody and conversational responsiveness.

Q3. Authenticity — Primary DV `P{phase}_C{clip}_Authenticity`

"How authentic and natural did this response sound?"

Scale: 1 (Not at all authentic) – 7 (Very authentic) — 7-point to match per-speaker Q8/Q12 for direct cross-level comparison.

Standalone item (not embedded in matrix) to maximise measurement sensitivity and analytical independence as the primary dependent variable.

Q4. External Source Use — Implicit AI Detection Proxy `P1_C{clip}_ExternalSource` **(Phase 1 only)**

"How likely do you think they were using information from external sources?"

Scale: 1 (Very unlikely) – 7 (Very likely)

Standalone; implicit Phase 1 probe for AI detection without naming AI. Removed from Phase 2 where "external sources" conflates with the now-known AI tool. Cross-phase comparison with Q5 captures the implicit→explicit detection shift (Analyses 2 and 5).

7.2 Per-Clip Measure — Phase 2 Only

Q5. AI Reliance Rating — Explicit Detection `P2_C{clip}_AIDetection` **(Phase 2 only)**

"How likely do you think this speaker relied on the AI suggestion tool when responding?"

Scale: 1 (Very unlikely) – 7 (Very likely)

Direct counterpart to Phase 1 Q4. Both use matched "How likely" wording and 7-point likelihood scale. The construct shift (implicit "external sources" → explicit "AI suggestion tool") is by design and represents the knowledge manipulation effect (Analyses 2 and 5).

7.3 Phase 1 Post-Study Questionnaire

Questions Q6–Q9 repeat three times per phase — once after each speaker's 3 clips. Speaker slot order is pre-randomised across the 24 conditions via cyclic permutation of all 6 orderings.

Q6. Global Speech Quality `P1_S{1-3}_GlobalQuality`

"How would you perceive the overall quality of this speaker's responses?" Scale: 1 (Very low) – 7 (Very high)

Repeated per speaker slot; yields 3 speaker-level quality ratings per participant per phase.

Q7. Speaker Assessment `P1_S{1-3}_Assessment`

Format: 7-point matrix (consistent with Q6/Q8). One item per dimension (alternating valence); each analyzed as standalone DV in Analysis 5.

"To what extent do the following words describe this speaker across the three clips you just heard? (1 = not at all, 7 = very much)"

Rows: Proficient (+) · Unclear (−) · Informative (+) — one item per dimension, alternating valence.

Repeated per speaker; parallel to Phase 2 Q11. Enables blind vs. informed comparison (Analysis 5).

Q8. Global Authenticity `P1_S{1-3}_GlobalAuth`

"How would you perceive the overall authenticity of this speaker's responses?" Scale: 1 (Not at all authentic) – 7 (Very authentic)

Mirrors per-clip Q3 at speaker level. Aggregate across slots = primary global DV for Analysis 4.

Q9. Activity Likelihood Matrix `P1_S{1-3}_ActivityLikelihood`

Repeated 3×, once after each speaker group (parallel to Q6–Q8).

"Thinking about the 3 clips from this speaker — how likely were they doing the following?" Scale: 1–7

Rows:

- Looking at on-screen information (e.g., notes, reference materials, web search)
- Looking elsewhere off screen (e.g., physical notebooks, wall)
- 3rd-party platform distraction (e.g., social media, messaging, email)
- **Using an online tool or digital assistance** *(generic; no AI mention)*

Row 4 uses generic wording to preserve the blind condition fully; no AI or suggestion language appears anywhere in Phase 1. Compared to Q13 row 4 ("Using an AI or digital assistance tool") in Analysis 5: the cross-phase row 4 shift isolates AI-specific attribution over generic digital tool use. Rows 1–3 are manipulation-check DVs. Export tags: `P1_S{1-3}_ActivityLikelihood`*.*

7.4 Phase 2 Post-Study Questionnaire

Questions Q10–Q13 repeat three times — once after each speaker's 3 Phase 2 clips (parallel to Phase 1 Q6–Q9). Q10–Q12 are worded identically to Q6–Q8; Q13 parallels Q9. Phase 2 context is established at the instruction level.

Q10. Global Speech Quality `P2_S{1-3}_GlobalQuality`

"How would you perceive the overall quality of this speaker's responses?" Scale: 1 (Very low) – 7 (Very high)

Parallel to Phase 1 Q6 per slot. Aggregate compared to Q6 aggregate = quality knowledge effect.

Q11. Speaker Assessment `P2_S{1-3}_Assessment`

Format: 7-point matrix — same rows as Q7.

"To what extent do the following words describe this speaker across the three clips you just heard? (1 = not at all, 7 = very much)"

Parallel to Phase 1 Q7; enables per-speaker blind vs. informed comparison on proficiency,

clarity, and content depth.

Credit-attribution note (Analysis 5 predictions): Proficient and Informative predicted to *decrease* (Q11 < Q7; credit-discounting: quality attributed to AI, not speaker); Unclear predicted to *increase* (Q11 > Q7; disfluency attributed to speaker, becoming more salient once AI assistance is known).

Q12. Global Authenticity `P2_S{1-3}_GlobalAuth`

"How would you perceive the overall authenticity of this speaker's responses?" Scale: 1 (Not at all authentic) – 7 (Very authentic)

Parallel to Phase 1 Q8 per slot. Aggregate $Q12 - Q8$ = primary authenticity knowledge effect (Analysis 4).

Q13. Activity Likelihood — Informed `P2_S{1-3}_ActivityLikelihood`

Repeated 3×, once after each speaker group (parallel to Q10–Q12).

"Thinking about the 3 clips from this speaker — how likely were they doing the following?" Scale: 1–7

Rows:

- Looking at on-screen information (e.g., notes, reference materials, web search)
- Looking elsewhere off screen (e.g., physical notebooks, wall)
- 3rd-party platform distraction (e.g., social media, messaging, email)
- **Using an AI or digital assistance tool**

Parallel to Phase 1 Q9 (per-speaker). Primary cross-phase comparison on row 4 at the speaker-slot level (Analysis 5). Rows 1–3 vs. Q9 rows 1–3 as manipulation-check DVs. Export tags: `P2_S{1-3}_ActivityLikelihood`.

Q14. AI Tool Utility `P2_AIUtility`

Once, end of Phase 2. Three-item matrix.

"To what extent do you agree with the following statements about the AI suggestion tool?" Scale: 1 (Strongly disagree) – 7 (Strongly agree)

Rows:

1. Using this tool helps the conversation flow better *(conversational quality)*
2. Using this tool leads to more informative and well-organized responses *(content competence)*
3. Using this tool helps the speaker sound more fluent and articulate *(linguistic*

competence)

(Q15h placed in demographics — see below.)

7.5 Demographics and Individual Difference Moderators

Q15a. Native speaker of English `Demo_NativeSpeaker` — Yes / No / Prefer not to disclose

If "No" is selected, the following four questions are shown via a conditional branch:

Q15a-i. First (primary) language `Demo_NNS_Language` — Open text (required)

Q15a-ii. Years lived in an English-speaking country `Demo_NNS_YearsInEngCountry`
Options: I do not currently live in one · Less than 1 year · 1–2 years · 3–5 years · More than 5 years

Q15a-iii. Daily English exposure hours `Demo_NNS_DailyEngHours`
"On a typical day, roughly how many hours are you exposed to English conversations? This includes both hearing and speaking English."
Options: Less than 30 minutes · 30 min–1 hour · 1–3 hours · 3–6 hours · More than 6 hours

Q15a-iv. Self-rated English proficiency `Demo_NNS_SelfRatedProf`
Options: Basic · Intermediate · Upper-intermediate · Advanced · Near-native

Q15b. Age `Demo_Age` — Open text (optional — leave blank if you prefer not to disclose)

Q15c. Gender `Demo_Gender` — Male / Female / Non-binary / Prefer to self-describe / Prefer not to disclose

Q15d. Ethnicity `Demo_Ethnicity` — Multi-select

Q15e. Field of study or professional expertise `Demo_Field` — Open text

Q15f. AI Task-Scenario Frequency `Demo_AITaskFrequency`
"How often do you use AI tools in each of the following scenarios?"
Scale: Never / Rarely / Sometimes / Often / Always | Rows:

1. Writing or drafting text (e.g., emails, messages, reports)
2. Looking up information or getting explanations
3. Rephrasing or improving the tone of your communication
4. Real-time conversation or meeting assistance *(primary moderator, H4a)*

Rationale: Task contexts are more stable and theoretically meaningful than tool names. Scoped to 4 communication-proximal scenarios. Rows 3–4 are most directly relevant to the study task; coding/creative/image-generation scenarios excluded as theoretically distal.

Q15g. AI Attitude Toward Real-Time Conversational AI `Demo_AIAttitude`

"To what extent do you agree with the following statements about AI tools embedded in real-time conversations?"

Scale: 1 (Strongly disagree) – 7 (Strongly agree) | Rows:

1. I trust AI tools embedded in conversations to provide helpful and accurate suggestions
2. I feel uncomfortable when people use AI suggestions during live conversations *(reverse-coded)*
3. It is appropriate to use real-time AI assistance while talking with others
4. I am concerned about people becoming overly reliant on AI during live social interactions *(reverse-coded)*

Scale balance: 2 positive items (1, 3) and 2 reverse-coded items (2, 4). Composite = items 1 + 3 + R(2) + R(4). Covers cognitive trust, normative appropriateness, affective discomfort (R), and over-reliance concern (R).

Q15h. Familiarity with Limited-Proficiency English Speakers `Demo_NNSFamiliarity`

"How familiar are you with conversing with people who are not fully fluent in English?"

Scale: 1 (Very unfamiliar) – 7 (Very familiar)

Moderator variable; placed at end of demographics, outside any phase context.

Note on scale consistency: All per-clip items (Q1–Q5) and all post-study/global items (Q6–Q14, Q15g) use **7-point scales** throughout. Uniform scaling eliminates between-question response-range confounds and aligns per-clip items with per-speaker items for direct cross-level comparison. Q4 and Q5 use matched "How likely" 7-point wording. Power analyses use Cohen's d (scale-invariant).

8. Key Design Decisions and Rationale

9. Planned Statistical Analyses

Framework: Linear mixed-effects models (LMM; R `lme4` or equivalent). Unless stated otherwise, all models include **participant** and **speaker** (`sona_id`) as crossed random intercepts, accounting for individual response tendencies and speaker baseline differences.

Each subsection is organised as: *Purpose · Model specification · DVs · Contrasts & predictions.*

9.1 Research Question – Hypothesis – Analysis Mapping

9.2 Analysis 1 — Phase 1 Implicit Detection (RQ1, H1, H2)

Do listeners implicitly detect AI use from blind Phase 1 ratings? Does disfluency moderate that detection?

Model `DV ~ suggestion_use_group × disfluency_group + (1 | participant) + (1 | sample)`

Both predictors 3-level; reference = Group 0. Phase 1 data only.

Primary DVs Q3 Authenticity; Q4 External Source Use (Phase 1 only)

Key secondary Q1_Elaborate — directional: Group 2 > Group 0 (AI-assisted responses tend to be more content-rich; cf. Q6 entering Analysis 4)

Secondary Q2_Confident, Q2_Flat, Q2_Engaged; Q1_Clear, Q1_Fluent, Q1_Coherent *(exploratory; no directional prediction)*

Contrasts Group 2 vs. Group 0 on primary DVs (H1a, H1b); disfluency Group 2 vs. Group 0 (H2a, H2b). Q1_Elaborate: one-tailed $\alpha = .05$. All other secondary contrasts exploratory.

Group 1 note Included in full model; reported descriptively. Dose-response pattern (Group 1 intermediate between Groups 0 and 2) noted if observed (exploratory).

9.3 Analysis 2 — Knowledge Effect (RQ3, H3)

Does revealing AI access shift authenticity and detection ratings (H3a, H3b)? Is the shift larger for high-AI-use clips than low-AI-use clips (H3c)?

Identification Phase 1 and Phase 2 use *different* speakers; the phase effect is identified **between conditions at the stimulus level** via counterbalancing. Each sample appears in Phase 1 (blind) for one counterbalance group and Phase 2 (informed) for the complementary group. Participant random intercepts absorb response tendencies; sample random intercepts absorb speaker baseline differences.

Model `DV ~ phase_condition × suggestion_use_group + disfluency_group + (1 | participant) + (1 | sample)`

`suggestion_use_group` effect-coded ($-1/0/+1$) so that the `phase_condition` coefficient = marginal phase shift averaged across SU groups, making H3b and H3c orthogonally estimable. `disfluency_group` is an additive covariate (no preregistered 3-way interaction).

Primary DVs Q1 dimensions, Q3, `AI_detection_likelihood` composite (Q4 Phase 1 /

Q5 Phase 2 pooled)
Robustness: separate models for Q4 (Phase 1 only) and Q5 (Phase 2 only) reported alongside the pooled model.

Speaker-level parallel Q6/Q10 global quality shift — analyzed at speaker-slot level in a parallel model (same structure as Analysis 4); reported as secondary replication.

Secondary DVs Q2_Confident, Q2_Flat, Q2_Engaged *(does AI knowledge shift perceived agency, affect, or engagement?)*

Contrasts `phase_condition` × `suggestion_use_group` interaction (H3c); informed > blind more strongly for Group 2 on Q3 (H3a); marginal `phase_condition` main effect on `AI_detection_likelihood` (H3b).

9.4 Analysis 3 — Moderator Analyses (RQ4, H4, H5)

Do individual differences in AI experience, attitudes, or native-speaker status amplify or attenuate detection sensitivity and knowledge effects?

Each moderator is entered separately (one model run each). Three-way interactions are exploratory only.

9.5 Analysis 4 — Global Perception Shift (H3a at Speaker Level)

Cross-level complement to Analysis 2: replicates the H3a authenticity shift using per-speaker ratings (Q8/Q12) rather than per-clip ratings (Q3).

Aggregation Q8 and Q12 are collected per speaker slot (3× per phase). Because Phase 1 and Phase 2 stimuli are *different* speakers, the analysis aggregates across all 3 slots per participant ($\overline{Q12}_{S1..S3}$ vs. $\overline{Q8}_{S1..S3}$). Counterbalancing ensures each stimulus contributes equally to both phases.

Model Linear regression with `(1 | pair_condition)` random intercept (24 levels; absorbs stimulus-level baseline variance). Participant random intercepts are not identifiable here (one aggregated observation per person).

Cross-level rule If Analysis 2 (Q3, per-clip) and Analysis 4 (Q8/Q12, per-speaker) diverge, the per-clip Q3 estimate from Analysis 2 is primary (larger N, greater sensitivity). Divergence is reported as a level-of-aggregation effect.

Primary DV $[\,\overline{Q12}_{S1..S3} - \overline{Q8}_{S1..S3}\,]$ authenticity shift

Secondary DV $[\,\overline{Q10}_{S1..S3} - \overline{Q6}_{S1..S3}\,]$ quality shift

Predictors `Demo_AIAttitude` composite; `Demo_AITaskFrequency` row 4; `Demo_NativeSpeaker`

9.6 Analysis 5 — Noticeability and Cue Attribution

Does AI attribution (row 4) increase once listeners are informed? Does it track actual AI-use level (`suggestion_use_group`)? Does knowledge shift speaker assessment dimensions (Proficient, Unclear, Informative)?

Q9 and Q13 (Activity Likelihood Matrix) are collected per speaker slot (3× per phase), enabling direct linkage to each speaker's known `suggestion_use_group`. Q9 row 4 uses generic wording ("online tool or digital assistance"); Q13 row 4 is explicit ("AI or digital assistance tool"). The cross-phase row 4 shift therefore isolates AI-specific attribution.

Model `DV ~ phase_condition × suggestion_use_group + (1 | participant) + (1 | speaker_sample_id)`

Analyzed at the **speaker-slot level** ($3 \times N$ observations per phase).

Primary DV Q9/Q13 row 4 (7-point generic-to-AI shift)

Manip. check Q9 rows 1–3 vs. Q13 rows 1–3 (identically worded): if unchanged, the row-4 shift is AI-label-specific, not a generic increase in digital-tool attribution

Secondary DVs Q7→Q11 shifts (Proficient, Unclear, Informative), each at speaker-slot level using the same model

Moderator Q14 AI Tool Utility composite: does believing AI improves communication amplify Q7→Q11 shifts? *(exploratory)*

Predictions Q13 row 4 > Q9 row 4; larger increase for Group 2 than Group 0 (credit-attribution theory).

Proficient and Informative: Q11 < Q7 (credit-discounting: quality attributed to AI, not speaker).

Unclear: Q11 > Q7 (disfluency attributed to speaker, more salient once AI assistance known).

Higher Q14 amplifies all shifts *(exploratory)*.

9.7 Multiple Comparisons Policy

Tests are organised in three tiers:

Bonferroni correction for primary confirmatory tests is applied within each analysis block (not across all analyses), because each block addresses a distinct research question.

9.8 Effect Size and Reporting

Cohen's d for pairwise contrasts; marginal and conditional R^2 for LMMs; 95% CIs reported throughout. Bonferroni-corrected thresholds reported as supplementary for primary tier.

9.9 Timing Data (Exploratory)

Each clip block includes a Qualtrics Page Timer (`PageSubmit`, in seconds; hidden from participants). Planned exploratory uses:

- **Empirical clip-length estimation:** median `PageSubmit` minus fixed rating time (≈85 s) estimates mean audio duration per clip for survey duration calibration.
- **Attention-check operationalization:** E4 threshold (<15 s) is derived from `PageSubmit` data.
- **Rating quality analysis:** correlation between `PageSubmit` and within-participant rating variance across clips (exploratory; longer dwell → more differentiated ratings?).
- **Replay detection:** `ClickCount` > expected baseline may indicate audio replay; reported descriptively.

9.10 Prespecified Sensitivity Analyses

S1 Replace `disfluency_group` (categorical) with continuous `total.disfluency.length` in all LMMs.

S2 Add `pitch.obvious.level` (0/1/2) as a covariate.

S3 Stricter completion threshold: all 9 clips rated per phase, no missing items within any block.

S4 Restrict to native English speakers only.

A. Stimulus Corpus Characteristics

Suggestion use distribution (full corpus):

- Group 0 (No use): 76 clips (31.5%)
- Group 1 (Partial use): 86 clips (35.7%)
- Group 2 (Full use): 79 clips (32.8%)

Intercorrelation note: Total disfluency length and clip disfluency count are significantly correlated. Pitch obviousness is independent of both disfluency measures and is therefore not included as a primary stratification variable. Disfluency type (filled vs. unfilled pause) is not included in primary stratification due to balancing feasibility concerns; it may be captured as a covariate in follow-up analyses.

B. Survey Flow Summary

```
Consent -> Study Overview -> Sound Check (demo audio)
-> Condition assignment (BlockRandomizer: 24 conditions)
```

```
-> Phase 1 Instructions ("Part 1 - Audio Clips")
-> Per speaker slot (3x):  [slots 1, 2, 3]
     Reminder: "Group X of 3"
     3 x Clip Block (randomised within slot):
          Audio player (MP3)
          Q1  Speech (Clear/Elaborate/Fluent/Coherent) -- 1-7 matrix
          Q2  Speaker impression (Confident, Flat, Engaged) -- 1-7
             matrix [3 items]
          Q3  Authenticity                                  -- 1-7
             standalone
          Q4  External Source Use ("How likely")            -- 1-7
             standalone [Phase 1 only]
     Per-speaker Qs: Q6 global quality | Q7 assessment | Q8 global
        auth
                   | Q9 activity likelihood (blind; row 4 = "online
                      tool or digital assistance")
-> AI Reveal screen [+ tool illustration]
-> Comprehension Check (neutral recall MC; exclude if != option 3)
-> Phase 2 Instructions ("Part 2 - Audio Clips")
-> Per speaker slot (3x):  [slots 1, 2, 3]
     Reminder: "Part 2 - Group X of 3"
     3 x Clip Block (randomised within slot):
          Q1, Q2, Q3 (same as Phase 1) + Q5 AI reliance ("How likely
             ", 1-7) [Phase 2 only]
     Per-speaker Qs: Q10 global quality | Q11 assessment | Q12
        global auth
                   | Q13 activity likelihood (informed; row 4 = "AI
                      or digital assistance tool")
-> End of Phase 2: Q14 AI utility (3-item matrix; once only)
-> Demographics (Q15a-Q15h; NNS branch -> Q15a-i to Q15a-iv; Q15h at
    end)
-> Closing: open-ended comments (optional) + thank-you screen
```

Listing 1: Full survey flow as implemented in Qualtrics.

Conditions: 24 (12 pairs × 2 counterbalance groups), evenly distributed via Qualtrics randomiser. **Total questions per participant:** ∼115 (native speakers) / ∼119 (non-native speakers; additional Q15a-i to Q15a-iv via branch). Includes closing comments screen (optional, unscored). **Estimated duration:** 45–60 minutes (conservative; `PageSubmit`

timing data will yield an empirical estimate post-pilot). **SONA compensation:** 2 credits (1 credit = 30 min; compensated at the 60-min rate).

Table 7: Summary of key design decisions and their rationale.

Decision	Rationale
Audio-only stimuli	All primary coding variables are auditory; eliminates visual confounds (gaze, posture); file sizes (72 KB–1 MB) allow Qualtrics direct upload
Authenticity as standalone item	Primary DV requires maximum sensitivity; matrix embedding creates halo effect and analytical multicollinearity
External source use as standalone	Functions as implicit AI detection proxy DV; embedding in engagement matrix would suppress sensitivity through response assimilation
No per-clip AI banner in Phase 2	Banner on every clip creates demand characteristic artificially inflating Q5 ratings across all clips, compressing variance in the key DV
Comprehension check as neutral recall	Avoids test-like reactance affecting subsequent ratings; achieves the same psychometric function without signalling evaluation
Different speakers across phases (counterbalanced)	Eliminates memory and repetition effects; counterbalancing ensures each sample is rated blind by half the participants and informed by the other half, making the knowledge effect identifiable between conditions rather than as a within-person difference on the same stimuli
Uniform 7-point scale for all per-clip items (Q1–Q5)	Eliminates response-range confounds between items; enables direct per-clip to per-speaker cross-level comparison; Q4/Q5 use matched "How likely" wording for implicit→explicit cross-phase comparison
AI demographics by task context	Task contexts more stable and theoretically meaningful than tool names; scoped to real-time conversation to match study context
Attitude scale scoped to real-time AI	Minimises construct-stimulus mismatch vs. general AI attitude scales; increases predictive specificity for Phase 2 ratings
Deliberate Q9/Q13 row 4 label shift ("online tool" → "AI tool")	Q9 uses generic wording to preserve the Phase 1 blind condition fully. Q13 uses the explicit AI label. The row 4 shift across phases therefore isolates AI-specific attribution; rows 1–3 (non-AI, identically worded) serve as within-comparison manipulation checks
Q15g reliability criterion ($\alpha \geq .70$)	Internal consistency computed post-collection; if $\alpha < .70$, items used individually rather than as composite; H4b composite test designated exploratory

Table 8: Cross-reference of RQs, hypotheses, analyses, and inference tiers.

RQ	Hypotheses	Analysis	Primary DV(s)	Tier
RQ1	H1a, H1b	Analysis 1	Q3 Authenticity; Q4 Ext. Source Use	Primary confirmatory
RQ1 ext.	H1 ext.	Analysis 1	Q1_Elaborate	Key secondary
RQ2	H2a, H2b	Analysis 1	Q3, Q4/`AI_det_lik.` w/ `disfl_group`	Primary confirmatory
RQ3	H3a, H3b	Analysis 2	Q3, `AI_detection_likelihood`	Primary confirmatory
RQ3	H3c	Analysis 2	Q3, `AI_detection_likelihood`	Primary confirmatory
RQ3 global	H3a (spkr)	Analysis 4	Q8/Q12 Auth. shift	Exploratory replication
RQ4	H4a	Analysis 3	`Demo_AITaskFreq.` × SU slope	Primary confirmatory
RQ4	H4b	Analysis 3	`Demo_AIAttitude` × phase	Exploratory
RQ4	H5a	Analysis 3	`Demo_NativeSpeaker` × SU slope	Key secondary
RQ4	H5b	Analysis 3	SU main effect per lang. group	Exploratory

Table 9: Moderator mapping: base model, interaction term, and hypothesis.

Moderator	Base model	Interaction term	Hypothesis
`Demo_AITask-Frequency` row 4	Analysis 1	× `suggestion_use_group`	H4a: higher AI-task frequency → larger Group 2 vs. 0 gap on Q3/Q4
`Demo_AIAttitude` composite	Analysis 2	× `phase_condition`	H4b: more negative attitude → larger Phase 1→2 shift on Q3 and `AI_detection_likelihood`
`Demo_NativeSpeaker`	Analysis 1	× `suggestion_use_group`	H5a: NES → larger Group 2 vs. 0 contrast on Q3/Q4
`Demo_NNS_YearsInEngCountry` / `Demo_NNS_DailyEngHours`	Analysis 1	× `suggestion_use_group`	Exploratory: English exposure (years in country; daily hours) moderates detection sensitivity for NNES participants
`Demo_NNSFamiliarity`	Analysis 1+2	main effect + × `phase_condition`	Exploratory: familiarity with NNS speech modulates baseline ratings

Table 10: Prespecified inference tiers for all tests.

Tier	Tests included	Standard
Primary confirmatory	Q3 (H1a, H3a), Q4/`AI_detection_likelihood` (H1b, H3b), `disfluency_group` on Q3/Q4 (H2a, H2b), `phase_condition` × `suggestion_use_group` on Q3 and `AI_detection_likelihood` (H3c), `Demo_AITaskFrequency` × `suggestion_use_group` (H4a)	$\alpha = .05$ two-tailed; Bonferroni within each analysis block
Key secondary (directional)	Q1_Elaborate (H1 extension, one-tailed); `Demo_NativeSpeaker` × `suggestion_use_group` on Group 2 vs. 0 slope (H5a, one-tailed); Q9/Q13 row 4 × `suggestion_use_group` (Analysis 5, speaker-slot level)	$\alpha = .05$ one-tailed as prespecified
Exploratory	All Q2 items, Q1_Clear/Fluent/Coherent, Q6/Q10 quality shift, Q7→Q11 shifts, H4b (`Demo_AIAttitude` × `phase_condition`), H5b (detection tendency null), Analysis 4 quality shift, `Demo_NNSExperience`/ `Demo_NNSFamiliarity` interactions, Group 1 dose-response	$\alpha = .05$ uncorrected; reported with effect sizes and 95% CIs

Table 11: Full disfluency statistics by suggestion use group.

Group	*M* Total Disfluency	*M* Count	*M* Sentence Disfluency
0 (No use)	5.17 s	3.88	1.16
1 (Partial use)	4.34 s	3.56	1.43
2 (Full use)	3.41 s	2.97	1.30