Useful, Not Human: A Feasibility Study of User Trust and Credibility in AI Interview Training

Dr Daniel Johnston
Cassette Group Ltd
dan.johnston@cassettegroup.com

Sophia Hudelist
Department of Languages and Linguistic Science
University of York

Alex Hannay
Cassette Group Ltd

AI-powered interview simulators are increasingly deployed as scalable training tools, yet little is known about whether users perceive them as credible sources of developmental feedback. This paper presents a mixed-methods feasibility study ($n = 14$) evaluating user trust, credibility, and feedback acceptance in a voice-based AI interview simulator, combining validated psychometric instruments with qualitative open-text responses in a pre–post design. Feedback delivery showed preliminary signs of being a failure-sensitive component, with participants who experienced feedback failures tending to rate helpfulness lower even when the interaction itself was perceived positively, though this rests on a small, non-randomised subgroup. Distrust decreased following a single session ($r = .68$) whilst active trust did not rise to the same degree, supporting a two-factor treatment of trust. Users also adopted functional mental models (Examiner, Tool, Coach) rather than relational ones. These preliminary findings suggest the success of AI interview training may depend less on human-like realism and more on reliable, contextually grounded feedback.

conversational AI, human-AI trust, credibility, AI feedback, interview training, mixed-methods

1. INTRODUCTION

Conversational AI systems are moving beyond task automation into roles that require interpersonal credibility, such as coaching, mentoring, and providing evaluative feedback. One area where this shift is most visible is interview preparation. AI mock interview tools are among the fastest-growing applications of conversational AI in professional development, with the global market projected to exceed USD 6.3 billion by 2031 (Verified Market Research 2025) and 70% of job seekers reporting use of generative AI in their job search (Indeed 2025). This scale of adoption implies that some trust- and credibility-relevant mechanisms already transfer; this paper examines which ones hold in a training setting.

Two underexplored questions, distinct from the RQs below, motivate this study. Firstly, trust and distrust may operate as separate constructs (Scharowski et al. 2025), such that a system reduces suspicion without building active confidence, yet this has not been tested empirically in AI training interactions. Secondly, a growing body of work has examined how candidates respond to AI interviewers in recruitment contexts, where they face judgement with real consequences (Clavel et al. 2024; Lavanchy et al. 2023; Wesche and Sonderegger 2021), but AI interview training is a fundamentally different context. The AI is a coach rather than a gatekeeper, and the user's goal is to learn rather than to perform. Whether these recruitment findings transfer to training remains an open empirical question.

Shi and Wang (2025) demonstrated that AI mock interviews can improve graduates' perceived employability, but did not examine how users perceived the AI interviewer itself, including whether they found it credible or accepted its feedback. Evaluation frameworks for conversational AI have been developed primarily for health and therapy chatbot contexts (Ding et al. 2024), where what matters most to users differs substantially from professional training tools such as interview training. What remains missing is evidence addressing a more fundamental question: *under what conditions do users accept feedback from a conversational AI interviewer, and what determines whether they perceive it as a credible training partner?* If users do not perceive the AI as credible or trust the system,

© Johnston et al. Published by BCS Learning and Development Ltd.

they will not engage meaningfully with its feedback, regardless of how natural the conversation feels.

This paper reports a mixed-methods feasibility study addressing three research questions: (RQ1) how users mentally model the AI interviewer and how this affects engagement; (RQ2) what conversational behaviours influence perceived credibility; and (RQ3) how feedback quality and delivery affect user trust. Given the feasibility scope ($n = 14$), all statistical findings are exploratory and effect sizes carry the main interpretive weight.

2. BACKGROUND

Trust and credibility. Trust in automated systems has been extensively studied since Jian et al.'s (2000) foundational scale development. Recent research recommends treating trust and distrust as separate constructs (Scharowski et al. 2025; McGrath et al. 2025), such that a system may reduce distrust without building active confidence, or vice versa. McCroskey and Teven's (1999) Source Credibility Scale, originally developed for human communicators, measures perceived competence, caring, and trustworthiness, dimensions directly relevant to AI systems providing evaluative feedback, though the scale predates the trust–distrust distinction and was not designed with it in mind. Trust in AI is also conditional and sensitive to performance stability (Lee and See 2004).

Conversational agents. The Computers Are Social Actors (CASA) paradigm proposes that users apply social heuristics to technologies presenting human-like attributes (Nass and Moon 2000). Gambino et al. (2020) extended this further to account for users developing distinct interaction scripts for media agents, a prediction explored in this paper.

Feedback. Effective feedback must address three questions: Where am I going? How am I going? Where to next? (Hattie and Timperley 2007). Kluger and DeNisi's (1996) meta-analysis found feedback interventions improve performance on average but worsen it in a minority of cases, pointing to a view in which specific, process-focused feedback tends to outperform generic praise. The Feedback Orientation Scale (Linderbaum and Levy 2010) captures dispositional feedback receptivity, including perceived utility, accountability, and self-efficacy. To our knowledge, no published work has yet examined how users receive and respond to AI-generated feedback in interview training contexts specifically.

3. METHOD

3.1. Participants and System

Participants were recruited through an application form distributed to university students. A total of 17 were recruited, with 14 completing both pre- and post-interview measures, yielding the matched pairs reported here. Each participant selected one of six graduate-level roles (Software Engineer, HR/People Operations, Research Assistant, Teaching Assistant, Project Coordinator, or Marketing) and submitted their CV in PDF format, enabling the system to tailor interview scenarios to their intended career pathway. Demographic data beyond student status were not collected (see Limitations).

The AI interview simulator was built using the ElevenLabs Conversational AI platform (accessed February–March 2026) with a large-language-model backend. Submitted CVs were processed into a vector store and made available to a retrieval-augmented generation (RAG) agent, orchestrated through n8n, allowing the system to retrieve and reference specific CV content during the interview. The selected job description was injected into a dynamic system prompt, enabling the system to tailor its questioning to both role requirements and individual experience. The system opened with standard introductory questions, branched adaptively into CV-grounded follow-ups, and generated evaluative feedback at the close of each session, typically addressing communication clarity, response structure, and content specificity. Sessions ran approximately 10–15 minutes. The full system prompt template is available as supplementary material on request from the corresponding author (dan.johnston@cassettegroup.com).

3.2. Measures

The study employed a pre–post design combining validated psychometric scales with targeted custom items. The pre-interview battery included the Trust in Automation Scale (Jian et al. 2000), scored as separate trust (7 items) and distrust (5 items) factors following Scharowski et al. (2025), alongside the Feedback Orientation Scale (Linderbaum and Levy 2010). The post-interview battery repeated the Trust scale and added the Source Credibility Scale (McCroskey and Teven 1999) (18 items, 7-point semantic differentials, reverse-coded per the original protocol) and the User Engagement Scale – Short Form (O'Brien et al. 2018) (items reworded to reference the AI interviewer explicitly; Aesthetic Appeal subscale excluded as not relevant to a voice-based agent). It also included a forced-choice mental model item (Coach, Examiner, Peer, Tool/Utility, Conversation Partner, Quizzer,

Interviewer, Other). Finally, interaction-quality items (responsiveness, naturalness) and feedback items (helpfulness, defensiveness, explanation quality) captured session-specific perceptions, with each Likert item paired with an open-text follow-up.

3.3. Procedure and Technical Variation

Participants completed the pre-interview questionnaire online ($\sim$7 minutes), engaged in a 10–15 minute voice-based interview, then completed the post-interview questionnaire ($\sim$12 minutes). During testing, two technical issues affected consistency: an intermittent feedback-generation timeout, and response degradation in later sessions (reduced coherence, weaker follow-ups). Four participants experienced one or both (the *degraded* group); the remaining ten experienced the system as intended (the *clean* group). This split was not a planned manipulation but the only naturally occurring performance variation observed during testing; we report the comparison because it offers a preliminary signal on how performance failures relate to user perceptions, not because groups were randomly assigned. Assignment was not random and group membership is confounded with session timing (later sessions were more likely to degrade), so this comparison is exploratory and hypothesis-generating rather than confirmatory.

3.4. Analysis

This study follows the feasibility framework outlined by Orsmond and Cohn (2015); 14 matched pairs is consistent with the widely cited recommendation of 12 per group for feasibility studies (Julious 2005). Quantitative analysis employed non-parametric tests (Wilcoxon signed-rank, Mann–Whitney U, Spearman ρ) with effect sizes $r = Z/\sqrt{N}$ interpreted per Cohen (1988) ($r = .10$ small, $.30$ medium, $.50$ large). Given the sample size, p-values should be read as descriptive, with effect sizes carrying the main interpretive weight; key analyses were repeated with the clean group only ($n = 10$) as a robustness check. Open-text responses were analysed via inductive thematic analysis (Braun and Clarke 2006), with initial coding by the first author refined through iterative comparison across participants and patterns. Given the brevity of responses and single-coder analysis, themes should be treated as preliminary; formal inter-rater reliability was not computed. The same author who developed the AI system conducted the initial qualitative coding (see Limitations).

3.5. Ethics and Use of AI

This study received ethical approval from the University of York (reference Reid_180825_Staff, approved 28 August 2025; amendment Gibb-Reid_090925_Amendment_Staff, approved 23 September 2025), and adhered to the British Psychological Society Code of Human Research Ethics (British Psychological Society 2021). All participants gave informed consent and retained the right to withdraw without consequence. No audio recordings were retained; transcripts were stripped of identifying information; CVs were deleted following data collection; and all responses were anonymised and stored in accordance with GDPR. Per the conference's AI-use declaration: idea generation, study design, and analytic approach were conceived by the authors without AI assistance; literature review used AI tools (Claude, Anthropic) for search, with all sources verified against original publications; AI tools (Claude, Anthropic) were also used for proofreading and language editing of author-written text, with all edits reviewed and approved by the authors; data collection involved no AI tools beyond the study artefact itself (the ElevenLabs platform and its n8n-orchestrated retrieval pipeline, evaluated rather than used as a research-conduct tool). All research design decisions, analytical interpretations, and conclusions are the authors' own.

4. RESULTS

Quantitative and qualitative findings are presented together, organised by research question. Table 1 summarises key scale scores and reliability estimates. Four themes emerged from thematic analysis (T1–T4, referenced below).

Table 1: *Key scale scores and reliability ($n = 14$). α reported where applicable; composites and single items marked '—'.*

Measure	M	SD	α
Pre-interview (1–7)			
Jian Trust (trust factor, 7 items)	3.14	0.94	.82
Jian Distrust (distrust factor, 5 items)	4.16	1.10	.84
Post-interview (1–7)			
SC Competence (6 items)	4.39	1.44	.91
SC Caring/Goodwill (6 items)	4.54	1.53	.94
SC Trustworthiness (6 items)	4.33	1.17	.85
SC Overall (18-item composite)	4.42	1.20	.94
Jian Trust (trust factor)	3.69	1.48	.95
Jian Distrust (distrust factor)	3.21	1.62	.93
Post-interview (1–5)			
UES Overall (9 items)	3.40	1.00	.91
Responsiveness (single item)	3.43	1.16	—
Naturalness (single item)	2.71	1.33	—
Feedback Helpfulness (single item)	3.21	1.25	—
Feedback Defensiveness (single item)	2.21	1.05	—

4.1. Feedback Is the Most Failure-Sensitive Component

The qualitative evidence for this finding is the most compelling, and so is presented first.

T1: Feedback failure as a trust-breaking event. One participant in the degraded group rated responsiveness 5/5 and naturalness 4/5, yet rated feedback helpfulness 1/5 because no feedback was generated; another wrote: 'I was a little apprehensive and then rather disappointed when it provided no room for me to expand.' In both cases, a positive interaction did not compensate for the absence of feedback.

To triangulate the single-item helpfulness measure, all open-text feedback responses ($n = 6$) were coded into three categories: *absent* ('feedback not generated'; helpfulness = 1/5), *present but generic* ('used the same line 2–3 times so it felt a little bit fake'; helpfulness = 3/5), and *present and valued* ('disconcertingly realistic, supportive, professional'; helpfulness = 4/5), with two participants in each category. This gradient aligns with Hattie and Timperley's (2007) distinction between praise, process, and self-regulation feedback, suggesting the single item captures a meaningful signal rather than noise.

An exploratory Mann–Whitney comparison between the clean ($n = 10$) and degraded ($n = 4$) groups found feedback helpfulness was substantially lower in the degraded group ($M = 2.00$, $SD = 1.41$) compared to the clean group ($M = 3.70$, $SD = 0.82$; $U_1 = 33.5$, $n_1 = 10$, $n_2 = 4$, $r = .53$, medium–large effect; $p = .049$, exploratory; see Limitations for confidence interval caveats). Feedback explanation quality trended in the same direction ($M = 2.25$ vs. 3.70). No equivalent group differences were observed for overall trust, credibility, or engagement, suggesting that the feedback pipeline is the component whose failure is most directly felt by users (Figure 1).

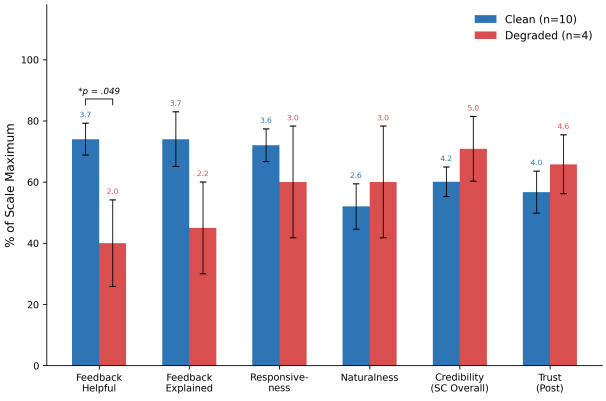

Figure 1: *Clean ($n = 10$) vs. degraded ($n = 4$) group comparison across key measures. Feedback helpfulness shows the largest group difference. Group sizes and non-random assignment limit inference; this comparison is hypothesis-generating.*

Notably, one participant who experienced peripheral technical issues still produced the most positive evaluation in the entire dataset. This points to a rough hierarchy of how much failures actually hurt: missing feedback seems to wreck the experience, cutting the interaction short does real damage, and peripheral glitches that don't touch core functions are mostly fine.

4.2. Trust: Reduced Distrust, Not Increased Confidence (RQ3)

Distrust decreased from pre-interview ($M = 4.16$, $SD = 1.10$) to post-interview ($M = 3.21$, $SD = 1.62$; Wilcoxon $W = 12.0$, $n_{nonzero} = 14$, $r = .68$, large effect; $p = .011$). The trust sub-factor trended upward ($M = 3.14$ to 3.69) but did not reach conventional significance ($p = .12$); see Figure 2 (an overall composite is shown alongside the two factors only as a bridge to prior single-factor literature; the separated factors remain primary). Restricting the analysis to the clean group ($n = 10$) as a robustness check preserved the direction and magnitude of all trust effects ($r = .52$–.59), consistent with reduced power rather than an absence of effect ($p = .065$–.099). Pre-interview trust predicted post-interview trust ($\rho = .58$), and post-trust correlated strongly with Source Credibility ($\rho = .77$). Together, these results suggest the interaction primarily reduced participants' suspicion and wariness rather than building active confidence, echoing the two-factor model advocated by Scharowski et al. (2025).

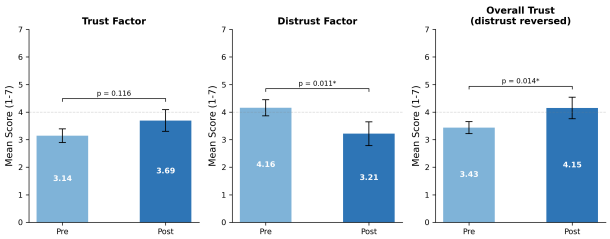

Figure 2: *Pre–post trust changes on the Jian Trust in Automation Scale ($n = 14$). Distrust decreased with a large effect ($r = .68$); the trust factor trended upward but did not reach significance. Error bars represent standard error.*

Feedback defensiveness was low ($M = 2.21/5$) and uncorrelated with dispositional feedback orientation ($\rho = -.06$), suggesting that the low defensiveness observed was a response to the AI interaction itself rather than a reflection of participants' dispositional receptivity to feedback.

4.3. Credibility: Built by Personalisation, Broken by Conversational Failures (RQ2)

Source Credibility scores were moderate across all three factors: Competence ($M = 4.39$), Caring ($M = 4.54$), and Trustworthiness ($M = 4.33$),

with substantial individual variation ($SD = 1.20$ overall). Naturalness, the lowest-scoring single item ($M = 2.71$), correlated strongly with overall credibility ($\rho = .74$), whereas responsiveness ($M = 3.43$) did not ($\rho = .37$), suggesting naturalness, not mere relevance of response, is the stronger predictor of perceived credibility. UES Reward correlated strongly with Source Credibility ($\rho = .80$).

T2: Personalisation as credibility anchor. The most consistently cited positive feature was CV-specific questioning, including picking up a word mid-answer to generate a follow-up and requesting recall of protocols from prior experience, with one participant noting the system referenced "specific examples that I gave and... tie[d] them into my experience." This was valued even by participants who rated naturalness low; one gave responsiveness 5/5 but naturalness just 2/5, suggesting contextual relevance drives credibility *independently* of conversational fluency. The anchor proved fragile. However, one participant reported the system mischaracterising their degree, and in this single observed instance the credibility damage appeared disproportionate, as the system had implicitly claimed to understand their background.

T3: Conversational contract violations. Three failure types disrupted the experience and together explain the naturalness–credibility correlation above ($\rho = .74$). The first, temporal disruption, arose from pauses in turn-taking that felt 'eerie'; one participant identified the absence of paralinguistic cues, rather than the AI's content, as the actual source of discomfort. The remaining two were tonal incongruity, where over-emphatic praise felt repeated and 'a little bit fake', and voice inconsistency, where changing timbre made it 'feel like I wasn't talking to the same person throughout'. These point to concrete, addressable design variables, namely latency, feedback variation, and voice consistency, rather than general artificiality. Naturalness was not uniformly poor, however: one participant in the clean group described the voice as sounding "very natural, tone also their natural, and use of language very clear and human like," indicating substantial variance in conversational quality across sessions rather than a fixed ceiling.

4.4. Mental Models (RQ1, Reported Briefly for Completeness)

Participants selected the description that best characterised their perception of the AI interviewer from 8 options; with $n = 14$, expected cell sizes are below 2 (Examiner $n = 5$, Tool $n = 3$, Coach $n = 3$, remaining three spread across three other categories), so this item cannot reliably distinguish a genuine pattern from noise and is reported briefly rather than as a full finding.

T4: Functional framing. The more defensible evidence comes from the qualitative data: participants consistently framed the AI in functional terms while remaining aware of its artificial nature (71% reported no change in perception during the session), and even the most positive participant in the dataset, who called the experience 'disconcertingly realistic' and 'better than most in-person interviews', still categorised the AI as a Coach rather than a peer. Illustrating this pragmatically: one participant was "sort of impressed by its ability to ask for clarification" while remaining "still not fond of how robotic it sort of feels"; the most positive participant described the AI "mirror[ing] my answers back... which was reassuring," reporting "confidence increased through the reflective engagement." This functional stance is consistent with Gambino et al.'s (2020) extended CASA framework, which anticipates users developing distinct interaction scripts for media agents rather than applying social heuristics wholesale. UES Perceived Usability was the highest subscale ($M = 3.95$), Focused Attention the lowest ($M = 2.93$).

5. DISCUSSION AND DESIGN IMPLICATIONS

Source Credibility, post-interaction Trust, and UES Reward correlated at $\rho = .77$–$.87$ in this sample; the implications below should be read with this possible construct overlap in mind.

Feedback delivery may be a critical value driver. The T1 pattern, drawn from a small and unevenly split sample, indicates that users approached the tool to learn, and when feedback was absent the experience lost its purpose even when the interaction itself was perceived positively. Hattie and Timperley's (2007) feedback model asks 'where am I going, how am I going, where to next'; the AI partially addressed 'how am I going' but rarely 'where to next'. Even in the clean group, feedback was praise-focused rather than process-focused, among the least effective forms (Kluger and DeNisi 1996); one clean-group participant summarised this directly as "superficial and not intellectually engaging." If this pattern holds under confirmatory testing, feedback reliability and specificity should take priority over other design concerns as users appear to tolerate conversational artificiality, but not absent or shallow feedback.

Personalisation may anchor credibility, but the error case is a single observation. The strong Reward–Credibility correlation ($\rho = .80$, itself part of the overlap noted above) supports personalisation as a mechanism linking engagement

to the tools perceived legitimacy, consistent with T1, and connects to McCroskey and Teven's (1999) competence measurements. Expertise can, in principle, be undermined by a single error, though this comes from one observed instance and should be treated as a hypothesis rather than a finding, building on Shi and Wang's (2025) employability work.

Trust is reduced distrust, not built confidence. The interaction reduced distrust more than it built active trust, supporting Scharowski et al.'s (2025) two-factor treatment in a real interaction context. This reframes the early-session design question from 'how do we make users trust the AI?' to 'how do we avoid giving users reasons to distrust it?': a system that consistently avoids trust-damaging events may accumulate confidence more effectively than one that intermittently impresses but unpredictably fails.

Pragmatic, not anthropomorphic, stance. Participants tolerated considerable artificiality, consistent with the functional framing in T4; however, the contract violations in T3 caused discomfort even within this functional frame. Users do not appear to require anthropomorphism, but do appear to require the AI to respect the basic conversational contract of an interview: consistency, not sophistication, appears to be the acceptance threshold. This contrasts with recruitment contexts where anthropomorphic features influenced applicant experience (Clavel et al. 2024), suggesting the same AI interviewer may need different design priorities depending on whether it is evaluating or educating; notably, 71% of participants reported no change in perception during the session, indicating this functional frame was set early rather than negotiated. One candidate explanation is that anthropomorphism signals legitimacy under real judgement in recruitment, whereas training foregrounds reliability instead; this is speculative and warrants direct comparison. These technical issues are therefore best read as central findings rather than incidental flaws, demonstrating that robustness and stability are foundational requirements for AI-driven training tools.

6. LIMITATIONS AND FUTURE WORK

The sample of 14 matched pairs limits statistical inference and all findings are exploratory. The clean-vs-degraded comparison (10 vs. 4) is underpowered, non-randomised, and confounded with session timing, so the feedback helpfulness effect ($r = .53$, wide confidence interval) is a candidate for confirmatory testing rather than an established finding. Demographic data was not collected, leaving transferability to experienced professionals untested. High intercorrelations among post-measures ($\rho = .77$–.87) suggest possible construct collapse. The design also captures only immediate post-interaction perceptions rather than multi-session or downstream hiring outcomes. A further limitation concerns researcher independence: the qualitative coding was conducted by the same author who built the underlying system, without a second coder or formal inter-rater reliability, and without blinding to which theme each excerpt might support.

Priority future work includes: (a) a larger replication ($n \approx 60$–100) with demographic data and randomised feedback conditions; (b) a controlled comparison of detailed explanatory, generic positive, and absent feedback, testing Hattie and Timperley's (2007) model directly and isolating explanation's effect on trust, defensiveness, and willingness to act on critique; (c) a longitudinal design testing whether trust accumulates or plateaus beyond the initial distrust-reduction observed here; and (d) an experimental test of the personalisation–accuracy trade-off from T1, namely whether inaccurate personalisation produces *lower* credibility than none at all.

7. CONCLUSION

This feasibility study set out to examine what makes users accept an AI conversational interviewer as a credible training partner. Distrust was reduced more than active trust was built across the full sample, the strongest-powered result reported here. In addition, feedback delivery showed preliminary signs of being similarly consequential, and credibility was anchored by CV-grounded personalisation but destabilised by conversational inconsistencies and occasional inaccuracies. A more tentative, secondary observation is that participants approached the AI pragmatically, they saw the AI as a functional tool rather than a relational partner, and tolerated artificiality as long as the system behaved consistently.

As conversational AI moves further into evaluative and developmental roles, including code review, sales coaching, tutoring, and clinical training, these preliminary findings point to a design question worth taking seriously: not 'how realistic can we make it?' but 'how reliably can it help someone improve?'. The race to make AI feel more human may matter less than the effort to make it consistently, demonstrably useful.

REFERENCES

Braun, V. and V. Clarke (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology 3*(2), 77–101.

British Psychological Society (2021). Code of human research ethics. Available at: https://www.bps.org.uk/guideline/bps-code-human-research-ethics. Accessed April 2026.

Clavel, C., S. d'Armagnac, S. Hebrard, T. Hesters, and D. Potdevin (2024). Humanized AI in hiring: An empirical study of a virtual AI job interviewer's social skills on applicants' reactions and experience. *The International Journal of Human Resource Management 36*(2), 206–234.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Ding, H., J. Simmich, A. Vaezipour, N. Andrews, and T. Russell (2024). Evaluation framework for conversational agents with artificial intelligence in health interventions: A systematic scoping review. *Journal of the American Medical Informatics Association 31*(3), 746–761.

Gambino, A., J. Fox, and R. A. Ratan (2020). Building a stronger CASA: Extending the Computers Are Social Actors paradigm. *Human-Machine Communication 1*, 71–86.

Hattie, J. and H. Timperley (2007). The power of feedback. *Review of Educational Research 77*(1), 81–112.

Indeed (2025). Generative AI and job seekers. Available at: https://www.indeed.com. Accessed April 2026.

Jian, J.-Y., A. M. Bisantz, and C. G. Drury (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics 4*(1), 53–71.

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics 4*(4), 287–291.

Kluger, A. N. and A. DeNisi (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin 119*(2), 254–284.

Lavanchy, M., P. Reichert, J. Narayanan, and K. Savani (2023). Applicants' fairness perceptions of algorithm-driven hiring procedures. *Journal of Business Ethics 188*, 125–150.

Lee, J. D. and K. A. See (2004). Trust in automation: Designing for appropriate reliance. *Human Factors 46*(1), 50–80.

Linderbaum, B. A. and P. E. Levy (2010). The development and validation of the Feedback Orientation Scale (FOS). *Journal of Management 36*(6), 1372–1405.

McCroskey, J. C. and J. J. Teven (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs 66*, 90–103.

McGrath, M. J., O. Lack, J. Tisch, and A. Duenser (2025). Measuring trust in artificial intelligence: Validation of an established scale and its short form. *Frontiers in Artificial Intelligence 8*, 1582880.

Nass, C. and Y. Moon (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues 56*(1), 81–103.

O'Brien, H. L., P. Cairns, and M. Hall (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies 112*, 28–39.

Orsmond, G. I. and E. S. Cohn (2015). The distinctive features of a feasibility study: Objectives and guiding questions. *OTJR: Occupation, Participation and Health 35*(3), 169–177.

Scharowski, N., S. A. C. Perrig, N. von Felten, L. F. Aeschbach, K. Opwis, P. Wintersberger, and F. Brühlmann (2025). To trust or distrust AI: A questionnaire validation study. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 361–374. ACM.

Shi, W. and D. Wang (2025). Empirical research on the application of AI mock interviews in enhancing graduate perceived employability: A case study in Hangzhou, China. *Education and Information Technologies 30*, 18461–18484.

Verified Market Research (2025). Interview preparation tool market size, scope, and forecast. Report ID: 458535. USD 2.5B (2023) projected to USD 6.3B by 2031. Available at: https://www.verifiedmarketresearch.com/product/interview-preparation-tool-market/.

Wesche, J. S. and A. Sonderegger (2021). Repelled at first sight? expectations and intentions of job-seekers reading about AI selection in job advertisements. *Computers in Human Behavior 125*, 106931.