Recovering ADKAR Change-Readiness Barriers from AI-Conducted Voice Interviews

Hector Benitez Ventura
hector@latentvariables.com

Noah Alexander
noah@latentvariables.com

Yashraj Patel
yashraj@latentvariables.com

Latent Variables Labs

June 2026

Abstract

Organizational change research often detects weak adoption long before it can identify the barrier that would make the next intervention actionable. We ask whether an AI-conducted voice interview preserves enough diagnostic signal for blinded post-hoc recovery of which ADKAR readiness barrier (awareness, desire, knowledge, ability, or reinforcement) is behind a participant's account. In a controlled benchmark where five of six matched workplace accounts each weakened one ADKAR element while the sixth supported all, blinded scorers who saw neither condition nor ground truth recovered the engineered barrier in 20 of 40 deficit cases: 50.0% accuracy (Wilson 95% CI 35.2–64.8), well above the 20% five-label chance baseline ($p < .001$). Participants read one of the six accounts of a workplace move to a tool called Flowboard and completed a voice interview about it.

Three independent scoring passes labeled blinded transcript packets against a prespecified codebook with high inter-scorer reliability (nominal $\alpha = 0.76$ for the six-way barrier label, above the prespecified 0.67 floor). Signal was strongest for knowledge and ability; reinforcement was hardest, and control accounts showed some residual overdiagnosis of knowledge gaps. These findings indicate that controlled, AI-conducted conversational data can carry recoverable readiness-barrier signal above chance. They do not yet establish field diagnosis: human-coder validation and prediction of real adoption outcomes remain necessary, and we frame these as the next steps for this research program.

Keywords: conversational interviewing; change readiness; ADKAR; diagnostic validity; qualitative scoring; organizational change

1 Extended Introduction

1.1 The Diagnostic Problem

Organizational change fails in ways that are often obvious after the fact but poorly identified while the work is unfolding. A team adopts a new workflow and then slowly returns to the old one. A manager interprets the retreat as resistance, while workers describe it as unclear rationale, missing instruction, or absent follow-up. These explanations imply different remedies. More communication may help when workers do not understand why the change exists, but it can be irrelevant when they understand the rationale and cannot execute the new process during normal work. Training, by contrast, helps when workers do not know the steps but is wasteful when the real failure is that no one rewards or reinforces the new behavior. The scientific and practical problem is therefore not merely to measure whether readiness is high or low. It is to recover the first dominant weak link in the change chain at a level of resolution that can guide the next action.

ADKAR provides one influential language for this diagnostic problem. It organizes individual change readiness into Awareness, Desire, Knowledge, Ability, and Reinforcement [1, 2]. Readiness scholarship makes a related but broader claim: change depends on beliefs about need, commitment, efficacy, resources, task demands, and organizational context [3, 4, 5]. Both literatures imply that the same observed non-adoption can be generated by different latent barriers. A worker who ignores a new tool might lack awareness of the problem it solves. Another worker might know the reason and reject the burden. A third might agree with the goal and lack instruction. A fourth might know the procedure but face friction in the environment. A fifth might succeed initially and relapse when the organization stops paying attention. These cases are behaviorally similar but interventionally distinct.

The standard measurement instruments available to change researchers each compress this distinction in different ways. Surveys scale and can support psychometric reliability, but they force participants to translate situated

experience into fixed response options. Focus groups surface local language and shared narratives, but they introduce group dynamics and reduce privacy. Human interviews can probe episodes, contradictions, and examples, but expert time is scarce and hard to deploy across many parallel change efforts. Conversational systems occupy a methodologically interesting position between these instruments. They can ask follow-up questions, request concrete episodes, and maintain a stable protocol across many participants. They can also fail in ways that standard qualitative methods do not: they may miss openings, accept abstract answers, or produce transcript artifacts that make downstream coding brittle.

1.2 Benchmarking Conversational Evidence

We treat an AI-conducted voice interview as a measurement surface, not as a claim about automation. The testable question is narrow: when a known readiness barrier is embedded in a controlled workplace account, does the resulting conversation contain enough evidence for blinded scorers to recover that barrier above chance? This is deliberately smaller than field validation; it asks whether the signal that would make later adoption and intervention studies possible can be recovered at all.

A controlled design addresses an evaluation problem in barrier diagnosis: in real settings the "true" barrier is contested, and outcome data alone do not identify the mechanism. When a worker abandons a new process, the behavior does not reveal whether rationale, instruction, or follow-up was missing. Engineering known weak links fixes a target label before the interview and hides it from interviewer and scorers, so we can test whether conversational evidence carries recoverable diagnostic signal rather than whether post-hoc interpretation merely sounds plausible.

We built such a benchmark around a fictional workplace change: a team replacing a shared spreadsheet with a web tool called Flowboard, in six matched accounts that each weakened one ADKAR element while supporting the rest, plus a full-support control. Participants completed a voice interview with Riley, an AI interviewing system blind to the condition; the full materials and protocol appear in Section 4.

Scoring tested the conversation evidence itself: blinded scorers saw transcript evidence and extracted claims, but not the engineered condition, recruitment source, or ground truth.

1.3 Contribution and Scope

The empirical result is promising but bounded. In the primary scored set, blinded consensus recovered the engineered deficit barrier in 20 of 40 deficit cases. This is substantially above the 20% benchmark implied by random choice among the five deficit labels, but it is far from deployment-grade diagnosis. The control condition produced no-dominant-barrier calls in 10 of 14 cases, showing some resistance to overdiagnosis, while false positives were concentrated in knowledge. Element-level results were uneven. Knowledge and ability were recovered more often than reinforcement. Ability was sometimes confused with knowledge, suggesting that both interview and scoring procedures need sharper separation between not knowing what to do and knowing what to do but being unable to do it in practice. Reinforcement was the weakest element, suggesting that conversations need stronger time-line probes about what happened after initial use.

The contribution of this paper is therefore methodological. First, we define ADKAR barrier recovery as a supervised benchmark for conversational change-readiness measurement. Second, we show that blinded post-hoc scoring of voice-interview outputs can recover engineered barriers above chance in a controlled setting. Third, we document the failure modes that must be solved before stronger claims are justified, chiefly reinforcement probing, knowledge-ability disambiguation, and human-coder validation. The result should be read as a serious first benchmark, not as evidence that conversational interviews can yet diagnose real organizations on their own.

2 Background and Related Measurement Work

2.1 Readiness Theory and ADKAR

ADKAR is widely used because it maps change work onto individual-level readiness conditions that are easy to remember and interventionally interpretable [1, 2]. Awareness concerns whether the person understands why the change is happening. Desire concerns whether the person wants to participate. Knowledge concerns whether the person knows how to change. Ability concerns whether the person can perform the new behavior in practice. Reinforcement concerns whether mechanisms exist to sustain the change. The sequence is not a causal law in the strict experimental sense, but it is a useful diagnostic heuristic because earlier failures can explain later symptoms.

If a participant does not understand the reason for a change, downstream questions about training or reinforcement may be premature.

Organizational readiness theory provides a complementary frame. Weiner defines readiness as a shared psychological state involving commitment to change and efficacy to implement change [3]. Holt and colleagues developed a readiness scale that captures appropriateness, management support, change efficacy, and personal valence [4]. Armenakis and colleagues emphasize discrepancy, appropriateness, efficacy, principal support, and valence as beliefs that shape readiness [5]. These frameworks differ from ADKAR in language and level of analysis, but they share a central implication: readiness is multidimensional, and aggregate readiness scores can mask the mechanism that matters for action.

2.2 Behavioral Measurement Instruments

Qualitative interviewing is a natural response to that mechanism problem. Skilled interviewers can ask for examples, distinguish first-order complaints from deeper causes, and notice when participants answer abstractly [6, 7]. Thematic analysis then turns accounts into codes and patterns [8]. Yet qualitative depth is expensive, especially when an organization wants many small diagnostic reads across roles, locations, or phases of a change. Conversational systems promise a different scaling law, but they need evaluation standards that do not reduce interviewing quality to user satisfaction or transcript length. A conversational interview should be judged by whether it elicits evidence that supports valid downstream inference.

This benchmark borrows from both psychometric and qualitative traditions. From psychometrics, it borrows the idea of a fixed label space, prespecified scoring rules, reliability thresholds, and known ground truth. From qualitative research, it borrows the idea that labels must be grounded in participant speech rather than inferred from metadata. The scoring codebook required direct evidence or grounded extracted claims, and it prohibited use of condition, form, listing, URL, completion code, or ground truth. This design is intentionally conservative. It asks whether the transcript itself carries the barrier signal.

2.3 Reliability as a Minimum Condition

The evaluation also draws on content-analysis reliability. We report nominal agreement across three independent scoring passes and use Krippendorff-style alpha as the reliability policy [9]. Agreement does not prove validity, but low agreement would undermine the claim that the transcript contains a stable diagnostic signal. Conversely, agreement above a floor does not eliminate bias if all scorers share the same blind spots. For that reason, this paper treats model-based scoring as a rapid benchmark layer and specifies human qualitative coding as the next validation layer.

3 The ADKAR Barrier Recovery Task

3.1 Task Setup

Each participant saw one of six hidden conditions. Five conditions embedded a single ADKAR weak link: awareness, desire, knowledge, ability, or reinforcement. A sixth condition described a fully supported change and served as the no-dominant-barrier control. After reading the account, the participant completed an AI-conducted voice interview. The evidence for scoring was the resulting transcript plus extracted claims grounded in participant speech.

Scorers assigned one label to each blinded packet: awareness, desire, knowledge, ability, reinforcement, or none. They also recorded confidence, readiness ratings, evidence quotes, ambiguity notes, and whether the label was supported. The benchmark asks whether the conversation evidence lets scorers recover the hidden weak link without seeing condition or ground truth.

3.2 Primary Target and Benchmark

The primary target is the first dominant weak link in the ADKAR chain. For deficit cases, a response is correct when the consensus label matches the engineered weak link. The chance baseline is 20% because there are five possible deficit labels. We use this five-label baseline rather than a six-label baseline because the control label is not a plausible answer when the account intentionally weakens one ADKAR element. Control cases are analyzed separately as a specificity check: the desired label is none.

Table 1: Experimental conditions. Each participant saw one account. The ground-truth label was withheld from the interviewer and scorers.

Condition	Target	Manipulated weak link
Awareness-minus	Awareness	The change appeared without a clear explanation of what was wrong with the old spreadsheet or why the team was changing.
Desire-minus	Desire	The rationale was clear, but the participant did not want to switch and experienced the change as extra work with little personal value.
Knowledge-minus	Knowledge	The rationale and buy-in were present, but no one showed the participant the steps or workflow rules.
Ability-minus	Ability	The participant knew what to do, but practical execution failed because screens were clunky, timeouts occurred, and the task took too long.
Reinforcement-minus	Reinforcement	Initial rationale, buy-in, knowledge, and ability were present, but no follow-up, reminders, or accountability sustained the behavior.
Full-support control	None	The account supported rationale, buy-in, instruction, execution, and follow-up.

3.3 Operational Definitions

The label definitions were operationalized as follows. Awareness meant the participant did not understand why Flowboard was happening or what problem it solved. Desire meant the rationale was understood but the participant did not want the change. Knowledge meant the participant was willing enough, but lacked instructions, examples, or workflow rules. Ability meant the participant knew what to do but could not execute in practice because of local constraints such as access, workload, errors, or tool friction. Reinforcement meant the participant could do the change, but nothing made it stick after launch. None was used only when no dominant weak link was supported or the transcript was too thin for a defensible specific label.

The first-weak-link rule is important because downstream complaints can be generated by upstream failures. If a participant says they never received training and therefore could not use the tool, the target is knowledge rather than ability. If a participant says they knew the steps but the system timed out during normal work, the target is ability rather than knowledge. If a participant says the rollout worked at first but everyone drifted back once attention disappeared, the target is reinforcement. The rule does not eliminate all ambiguity, but it gives scorers a disciplined way to resolve common boundary cases.

4 Experimental Design

4.1 Conditions and Materials

The study used one shared setup: "Your team replaced a shared tracking spreadsheet with a new web tool called Flowboard." The six versions then varied one diagnostic element while holding the rest of the account as stable as possible. Table 1 summarizes the manipulation. The accounts were not intended to simulate the full richness of an actual organizational rollout. They were designed as controlled signal carriers. Their purpose was to create known weak links that participants could describe in their own words during a voice conversation.

4.2 Participant Flow and Conversation

The planned sample was 240 completed submissions, 40 per condition. The study recorded 235 completed submissions: 39 each in awareness-minus, desire-minus, knowledge-minus, ability-minus, and reinforcement-minus, and 40 in the control condition. Fifty-four responses entered the primary scored set after attention and scoring rules, including 40 deficit-condition rows. Appendix A reports the retention audit and sensitivity checks.

The conversation used a standard conversational skeleton plus an ADKAR-oriented topic pack. Participants were asked to answer as if the Flowboard account had happened to them. Riley was instructed to interview the participant about the change, listen for the five readiness elements, and avoid naming experimental conditions. The control prompt allowed Riley to conclude that there was no dominant barrier if the account showed support across the change. The intended structured live output did not reliably emit the barrier label, so the study analyzed

transcripts and extracted claims after the fact.

5 Scoring and Analysis

5.1 Blinded Scoring Packets

The scoring pass used three independent model-based scorers over the same blinded packets. Each packet contained a response identifier, transcript evidence, and extracted claims when grounded in participant speech. Scorers were explicitly prohibited from using condition, recruitment source, or ground truth. Each scorer returned a barrier label, confidence, element-readiness ratings, evidence quote, ambiguity notes, and a label-supported indicator. The allowed labels were awareness, desire, knowledge, ability, reinforcement, and none.

5.2 Consensus and Reliability

Consensus labels were formed by two-of-three or three-of-three agreement. Three-way splits were flagged. Median confidence at or below 2 was flagged. Knowledge-ability disagreements were flagged because that boundary was theoretically important and empirically likely. Confident non-none labels in control cases were also flagged because a control false positive is a distinct kind of overdiagnosis. The scoring layer therefore served two purposes: it produced a label for analyzable rows and generated diagnostics about where the instrument or codebook was unstable.

The primary outcome was deficit-condition accuracy: the proportion of analyzable deficit rows where the consensus label matched the engineered weak link. We report Wilson confidence intervals and exact binomial tests against a 20% benchmark. Control rows were analyzed separately because their target was no dominant weak link rather than one of the five deficit labels. Inter-scorer reliability was summarized using nominal alpha for the six-way barrier label and for label support. The prespecified reliability floor was 0.67. This floor is not a claim that scoring is complete; it is a minimum condition for treating the label set as sufficiently stable for a pilot benchmark.

6 Results

6.1 Primary Accuracy

In the primary scored set, 54 rows were analyzable: 11 desire-minus, 10 knowledge-minus, 11 ability-minus, 8 reinforcement-minus, and 14 control rows. No awareness-minus rows entered the primary set, so awareness recovery is not estimable in this launch.

Across the 40 analyzable deficit rows, the consensus label matched the engineered weak link in 20 cases: 50.0% accuracy (Wilson 95% CI 35.2–64.8%), well above the 20% five-label chance baseline (exact binomial $p < .001$).

6.2 Control Specificity

The control condition was useful as a specificity check. In the primary set, 10 of 14 control rows were labeled none. The four false positives were knowledge labels. This pattern suggests that the scoring procedure did not simply assign a deficit label to every transcript, but it also shows a bias toward interpreting minor uncertainty or incomplete procedural detail as a knowledge barrier. In future studies, the control condition should include explicit positive evidence of instruction to reduce ambiguity between real absence of knowledge and sparse participant elaboration.

6.3 Element-Level Errors

Element-level performance was uneven. Desire was recovered in 5 of 11 cases, knowledge in 6 of 10, ability in 7 of 11, and reinforcement in 2 of 8. Knowledge and ability together were correct in 13 of 21 cases. Four of the 21 knowledge or ability cases fell on the knowledge-ability off-diagonal, and all four were ability cases called knowledge. This asymmetry matters. The conversation and scoring process identified missing instruction more easily than practical execution friction. When participants described clunky screens, time pressure, or timeouts, scorers sometimes treated those accounts as not knowing enough rather than as inability to execute under working conditions.

Reinforcement was the weakest target. Only 2 of 8 reinforcement-minus cases were recovered in the primary set. The confusion pattern spread across desire, knowledge, ability, and none. This is theoretically plausible.

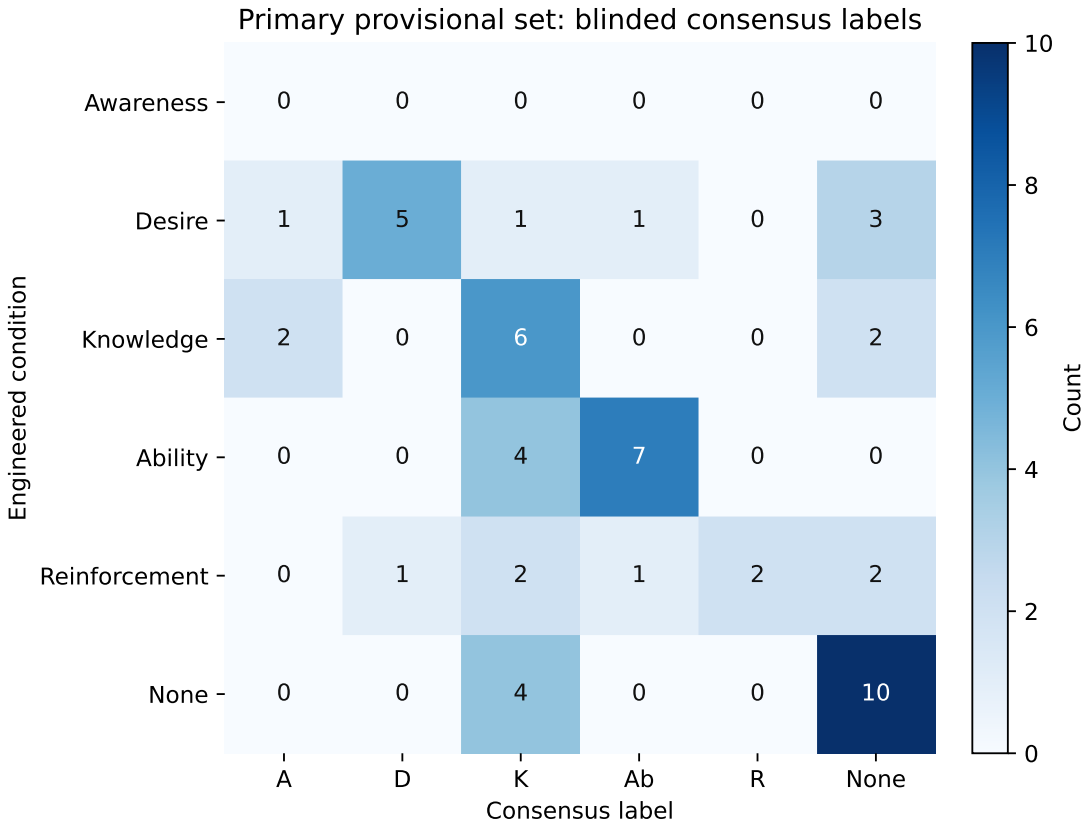

Figure 1: Primary provisional confusion matrix. Rows show engineered condition and columns show blinded consensus label. The awareness row is zero because no awareness-minus row entered the primary scored set.

Reinforcement is temporally downstream. It requires evidence that awareness, desire, knowledge, and ability were sufficiently present and that the failure occurred after initial use because follow-up, norms, or accountability did not sustain the change. A short conversation can easily capture the first part of that chain and still fail to establish the time-line needed for reinforcement. Future prompts should include explicit probes about what happened after the first successful use, what reminders or feedback existed, and when the participant or team drifted back.

6.4 Inter-Scorer Reliability

Scorers agreed well above our prespecified 0.67 floor: nominal α was 0.757 for the six-way barrier label and 0.718 for label support. In other words, the labels reflect a stable signal in the transcripts rather than one coder's reading. The consensus distribution included 41 three-of-three and 13 two-of-three agreements among the 56 scored responses, with two plurality-only cases. High reliability shows the retained transcripts contained a labelable signal; it does not prove the labels are human-valid or field-valid, a distinction central to our interpretation.

7 Error Analysis

7.1 Live Structuring Failure

The first failure was incomplete live structuring. The intended live barrier label did not reliably emit, which forced a post-hoc blinded scoring pass. This was a defensible rescue strategy because scoring was blind and independently replicated, but it is not the desired endpoint. A mature benchmark should capture transcript, extracted claims, structured barrier hypotheses, confidence, and provenance at the moment of conversation. Post-hoc scoring should then validate or audit live outputs rather than substitute for them.

Table 2: Scoring diagnostics from the three blinded passes. Counts are from the available validation summaries where applicable.

Failure mode	Count	Pattern	Implication
K/Ability disagreement	4	Scorers disagreed on instruction absence versus execution friction.	Add sharper interview probes and codebook examples.
Low median confidence	5	Scorers produced weak or ambiguous labels.	Escalate to human review or require richer transcript evidence.
Confident control non-none	2	Supported accounts were confidently overdiagnosed.	Strengthen control accounts and require positive evidence for deficit calls.

7.2 Reinforcement Ambiguity

The second failure was reinforcement ambiguity. Reinforcement is downstream and temporal. A participant can say that the rollout had a rationale, they wanted to try, they knew how, and it worked, but the conversation must then establish that the behavior decayed because no one sustained it. If Riley does not ask about what happened after launch, the transcript may contain enough information to rule out awareness, desire, knowledge, and ability, but not enough to support reinforcement. That can lead to none labels, knowledge labels, or scattered guesses. This suggests that reinforcement requires a distinct probe family focused on follow-up, reminders, manager attention, peer norms, feedback, and relapse.

7.3 Knowledge-Ability Boundary

The third failure was the knowledge-ability boundary. In the codebook, knowledge means "I do not know how" and ability means "I know how, but I cannot do it reliably here." The primary errors show that practical friction can be interpreted as insufficient knowledge unless the transcript makes the distinction explicit. The next version should force the distinction with paired questions: What were you told to do? Could you explain the steps? What happened when you tried during real work? If you knew the steps, what prevented execution? These questions are not cosmetic. They determine whether the resulting label implies training or environmental repair.

7.4 Control Overdiagnosis

The fourth failure was control overdiagnosis. Ten of 14 control rows were correctly labeled none, but four were labeled knowledge. This likely reflects sparse positive evidence. If a participant gives short answers, a scorer may see missing detail about training and infer a knowledge gap, even when the account said instructions were clear. Future controls should ask participants to recount the instruction they received, not merely state that it existed. The scoring rule should also distinguish absence of evidence in a thin transcript from evidence of absence in a participant account.

8 Discussion, Limitations, and Next Steps

8.1 Interpretation

The study provides evidence for a narrow but important proposition: AI-conducted voice interviews can carry recoverable signal about engineered ADKAR readiness barriers. The evidence is not strong enough to claim field diagnosis, automated consulting, or adoption prediction. It is strong enough to justify a research program that treats conversational elicitation as a candidate measurement method and evaluates it with known-answer tasks, blinded scoring, human validation, and eventually behavioral outcomes.

The scientific value of the benchmark is that it decomposes the problem. Elicitation is the process by which Riley turns a participant's account into conversational evidence. Scoring is the process by which blinded coders or models map that evidence onto a label space. Outcome validation is a later step in which labels predict behavior or intervention response. The present study tested elicitation plus post-hoc scoring under controlled ground truth, and that decomposition makes the next studies clearer.

8.2 Next Studies

The first next study should be a clean replication of the barrier-recovery benchmark, balanced across all six conditions with the awareness cell retained. The interview should include additional probes for reinforcement and paired probes that separate knowledge from ability. The scoring plan should include model scoring, two human qualitative coders on overlapping transcripts, and senior adjudication of disagreements.

The second next study should test predictive validity. After the conversation, participants should complete a behavioral task in which they choose whether to use the new workflow, persist through friction, or revert to the old workflow. The hypothesis would shift from barrier recovery to outcome prediction: do interview-derived readiness labels predict adoption, persistence, or reversion? This is a different claim and should not be collapsed into the present result.

The third next study should test intervention. If the conversational interview recovers the barrier, then a subsequent message can be generic, surface-personalized, or barrier-grounded. The behavioral endpoint would show whether acting on the recovered barrier changes behavior. This is the point at which the research moves from measurement to causal intervention. The present study does not make that leap; it prepares the ground for it.

8.3 Limitations

Several limitations remain. The accounts were simulated and intentionally clean. Real change accounts are messier and can contain multiple simultaneous weak links. Participants answered as if the scenario had happened to them, which may differ from the emotion and memory structure of lived workplace change. The retained sample was much smaller than the launched sample, and the awareness condition was not estimable in the primary set. The scoring layer was model-based rather than human-coded, although it used independent blinded passes and reliability checks. Recruitment provenance cannot be fully disclosed in public materials because of confidentiality obligations, which limits external audit. These limitations do not invalidate the benchmark, but they constrain the claim.

8.4 Conclusion

The main conclusion is therefore calibrated. A controlled AI-conducted voice-interview benchmark recovered ADKAR barrier signal above chance under blinded post-hoc scoring. The result is a useful first measurement result, not a final validation. The next step is to test whether the same interview-derived labels remain stable under human coding, predict real adoption behavior, and support barrier-matched interventions.

Ethics and Data Availability

This work was conducted as independent applied methods research outside a university-sponsored protocol. No institutional IRB determination was obtained. Participants saw a consent statement before the voice conversation, were told the conversation would be recorded for research analysis, and were instructed not to provide employer names. Public reporting withholds recruitment-source details where required by confidentiality obligations. Aggregate tables, scoring codebooks, and analysis summaries can be shared. Raw audio, raw transcripts, source provenance, and internal analysis code are not public because they may contain participant speech, confidential operational details, and proprietary research infrastructure. De-identified excerpts and audit artifacts can be reviewed under suitable confidentiality terms.

References

[1] Hiatt, J. M. *ADKAR: A Model for Change in Business, Government, and Our Community*. Prosci Learning Center, 2006.

[2] Prosci. "The Prosci ADKAR Model." Prosci methodology documentation, accessed 2026.

[3] Weiner, B. J. "A theory of organizational readiness for change." *Implementation Science* 4, 67, 2009.

[4] Holt, D. T., Armenakis, A. A., Feild, H. S., and Harris, S. G. "Readiness for organizational change: The systematic development of a scale." *Journal of Applied Behavioral Science* 43(2):232 to 255, 2007.

[5] Armenakis, A. A., Harris, S. G., and Mossholder, K. W. "Creating readiness for organizational change." *Human Relations* 46(6):681 to 703, 1993.

[6] Kvale, S. and Brinkmann, S. *Interviews: Learning the Craft of Qualitative Research Interviewing*. Sage, 2009.

[7] Brinkmann, S. *Qualitative Interviewing*. Oxford University Press, 2013.

[8] Braun, V. and Clarke, V. "Using thematic analysis in psychology." *Qualitative Research in Psychology* 3(2):77 to 101, 2006.

[9] Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*. Sage, 2018.

[10] Tong, A., Sainsbury, P., and Craig, J. "Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups." *International Journal for Quality in Health Care* 19(6):349 to 357, 2007.

[11] Belschak, F. D. and Den Hartog, D. N. "Consequences of positive and negative feedback: The impact on emotions and extra-role behaviors." *Applied Psychology* 58(2):274 to 303, 2009.

[12] Kotter, J. P. "Leading change: Why transformation efforts fail." *Harvard Business Review*, 1995.

[13] Ashwin, J., Chhabra, A., and Rao, V. "Using Large Language Models for Qualitative Analysis can Introduce Serious Bias." arXiv:2309.17147, 2023.

[14] Bommasani, R. et al. "On the opportunities and risks of foundation models." arXiv:2108.07258, 2021.

[15] Jacobs, S. and Wallach, H. "Measurement and fairness." *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2019.

A Exact Test Statistics

The primary deficit-accuracy test is an exact one-sided binomial test of 20 correct calls in 40 deficit rows against the 20% chance rate implied by a uniform choice among the five deficit labels. This yields $p = 2.17 \times 10^{-5}$. The small p indicates that 50% accuracy is very unlikely to arise by chance under the null; it is not a sign of a fragile or overfit effect, and the single primary test was prespecified. Per-tier figures are below.

Table 3: Per-tier deficit accuracy with exact statistics. "Strict deterministic join" denotes rows whose participant–condition link was confirmed by exact identifier match; "timestamp recovery" denotes rows additionally matched by a unique pre-start timestamp.

Tier	Deficit rows	Correct	Accuracy (Wilson 95% CI)
Primary provisional ($n = 54$)	40	20	50.0% (35.2–64.8)
Strict deterministic join ($n = 48$)	36	18	50.0% (34.5–65.5)
Conservative timestamp recovery ($n = 65$)	52	29	55.8% (42.3–68.4)
15-second timestamp recovery ($n = 67$)	54	30	55.6% (42.4–68.0)

B Recruitment Provenance and Public Reporting

Participants were drawn from ongoing lab recruitment and partner-mediated research flows. Some source details are withheld from public materials because upstream relationships are governed by confidentiality obligations. The public manuscript reports the information needed to evaluate the experimental logic: assignment counts, completion counts, scoring rules, and aggregate outcomes. Fuller provenance can be reviewed under appropriate confidentiality terms by an ethics or editorial reviewer.

C Scoring Codebook Summary

Scorers labeled the first dominant weak link in the ordered chain awareness, desire, knowledge, ability, and reinforcement. The none label was reserved for cases where no dominant weak link was supported or where the transcript was too thin for a defensible specific label. Evidence had to come from participant words, Riley-extracted claims grounded in participant speech, or Riley questions only insofar as they clarified the answer. Scorers were instructed not to use condition, recruitment source, completion code, or ground truth. Confidence ranged from 1 to 5, where 5 required a direct quote cleanly supporting one label and 1 indicated unusable, contradictory, or mostly inferential evidence.

D Representative Blinded Evidence Patterns

The study materials preserve raw participant speech and are not reproduced in full. The following paraphrased evidence patterns illustrate the coding logic without exposing identifiable speech. Awareness labels were supported when a participant emphasized that the change appeared without explanation and could not name the problem the new process solved. Desire labels were supported when the rationale was understood but the participant described resentment, low personal value, or unwillingness to switch. Knowledge labels were supported when the participant wanted to comply but described guessing at steps or lacking examples. Ability labels were supported when the participant knew the intended procedure but described timeouts, access problems, workload pressure, or inability to complete the task during normal work. Reinforcement labels were supported when the participant described initial successful use followed by drift after reminders, follow-up, or accountability disappeared.