Method Fidelity in AI-Conducted Organisation-Design Interviews:
Instrument Development and Evaluation

Hector Benitez Ventura[1]

Latent Variables Labs

hector@latentvariables.com

ORCID: 0009-0003-4244-7505

Naomi Stanford[2]

Independent organisation-design practitioner

naomi@stanford.cc

https://naomistanford.com/

August 2026

Abstract

Organisation design depends on interview evidence, and the capacity to gather it has long been limited by what a practitioner can personally conduct. AI-conducted interviewing raises a prior question, which is whether such an instrument enacts a named method with measurable fidelity. This paper develops and evaluates an AI-conducted voice interviewer against a twelve-move workflow-based organisation-design method, scoring each transcript move by move with attempted and landed behaviour separated and each judgement linked to transcript evidence. Across five reader-facing development stages, mean fidelity increased from 6.9 to 9.4 moves of twelve, follow-up of substantive answers from 42.7% to 78.2%, and respondent words per interview from 839 to 2,473. One specified move was attempted in none of 175 baseline interviews, which shows that a method component can be present in a protocol and absent from behaviour. Two blinded human coders independently scored 1,104 move decisions, and against the adjudicated standard the automatic scorer reached macro-averaged precision of .953 and recall of .943, with an intraclass correlation of .91 for total fidelity. Because sequential development could not separate individual changes from fielding order, a subsequent factorial evaluation varied three instrument features independently as a supporting attribution check. Follow-up defaults and runtime duration control were each associated with higher fidelity, by 1.05 and 1.35 moves, while a mechanical one-question constraint reduced multi-question turns sixfold without improving fidelity. That agreement establishes the reliability of the fidelity measure, not the accuracy or representativeness of respondent accounts.

Keywords: conversational interviewing; organisation design; method fidelity; instrument development; factorial evaluation; workflow analysis; qualitative scoring

1 Introduction

1.1 Interview Capacity as a Measurement Constraint

Organisation design has long operated under a practical constraint: a trained practitioner can speak with only a small proportion of the people whose work constitutes an organisation. That constraint does more than limit sample size. It determines which roles, handovers, exceptions, and informal workarounds become visible to the designer. Because knowledge of an enacted workflow is distributed across the people who perform its individual parts, interview capacity implicitly defines the empirical boundary of the organisation available for diagnosis. AI-conducted interviewing changes that constraint. It creates the possibility of gathering first-person evidence from every participant in a bounded workflow, including interstitial roles that purposive samples frequently omit. The scientific question is whether the practitioner's method survives that expansion. An instrument that produces more conversations while inconsistently pursuing the evidence required by the method expands volume without establishing a more usable evidence base. This study addresses the instrument-level question that precedes broader coverage: whether an AI-conducted interviewer can enact a named organisation-design method with measurable fidelity. We examine that question through five reader-facing development stages, move-level scoring against an approved twelve-move rubric, blinded human coding of the primary outcome, and a subsequent factorial evaluation designed to separate three instrument features that changed during development. Two features of the setting bear on the design. Organisation design begins with the work rather than with the reporting structure, and an organisation chart represents only the people element of the formal organisation, so redrawing it addresses neither the workflow, nor the culture, nor the interfaces between functions. Informal networks carry a substantial share of how tasks get done and are not visible in the formal structure [1]. Recovering how work is performed therefore means interviewing the people who

perform it, and the scale of that task is the constraint. A practitioner can typically interview fifteen to thirty people in an engagement. Each person knows their own portion of a workflow and none knows the whole, so the connective tissue between accounts is inferred rather than observed [12]. The roles least likely to be included are the interstitial ones, coordinators, schedulers, night shifts and contractors, which is where enacted practice most often diverges from documented process [13]. Conversational systems can ask follow-up questions, request concrete episodes, and hold a stable protocol across many participants. They can also fail in ways that trained interviewers do not, by accepting an abstract answer, missing an opening, or closing a conversation before the material is exhausted. Recent work establishes that respondents rate AI interviewers favourably and that conversational data surfaces considerations a standardised survey cannot reach [14, 15]. That literature addresses whether an AI can conduct an interview at all. It does not address whether a particular method has been performed.

1.2 Fidelity as the Evaluation Standard

A different question governs whether such an instrument is usable in organisation design. The value of a diagnostic interview lies not in whether it was pleasant but in whether it recovered the specific things the method requires. An instrument that produces long, agreeable transcripts and never asks which parts of the work could be automated has not performed the method, however well it is rated. This is a construct-validity question rather than a satisfaction question. What an instrument measures is established by evidence about the inferences it supports, not by the appearance of its output [16, 17]. The implementation-fidelity literature makes the corresponding operational move, distinguishing adherence to the specified components of an intervention from the quality with which they are delivered, and treating both as measurable [18, 19, 20]. Clinical method assessment gives a worked instance, in which motivational interviewing is scored against a published integrity code that names the behaviours a competent session must contain [21]. We therefore adopt fidelity to a named method as the evaluation standard, and define it operationally. The method used here specifies twelve moves a competent interview must make. Each transcript is scored move by move for whether the interviewer attempted the move and whether it landed, with the turn indices on which each judgement rests recorded so that every score can be audited against its evidence. This standard has a property that satisfaction measures lack. It can identify a move that is specified in the instrument and absent from its behaviour. Section 4.2 reports one such case, a move present in Version 1 instructions and attempted in none of 175 interviews.

1.3 Contributions and Scope

The paper makes three contributions. First, it operationalises fidelity to a named organisation-design interviewing method. Each interview is eval-uated against twelve specified moves, with attempted and landed behaviour separated and each judgement linked to transcript evidence. This makes the enactment of the method observable and auditable rather than inferred from the written protocol. Second, it reports the development of an AI-conducted interview instrument against that standard. The five-stage sequence shows where specified method components were absent from behaviour, how revisions corresponded with changes in fidelity and evidence volume, and which method elements remained difficult to recover.

Third, it examines whether three instrument features associated with changes during development retain those relationships when varied independently. The evaluation provides supporting evidence that follow-up defaults and runtime duration control contribute differently to fidelity, while a mechanical question constraint improves conversational discipline without improving method performance. The scope is limited to instrument development and evaluation. Agreement between human and automatic scoring establishes reliability of the primary fidelity measure. It does not establish that participant ac-counts are complete, accurate, or representative of an organisation. Because participants described different employers, the study does not test workflow coverage or organisation-level diagnosis.

2 Background and Related Measurement Work

2.1 Workflow-Based Organisation Design and the Congruence Model

Two distinct bodies of work meet in this study, and it is worth separating them at the outset. The method being encoded is the second author's approach to workflow-based organisation design, in which the enacted flow of work is the unit of analysis [22]. Its twelve moves, set out in Appendix A, are hers, and they are what the instrument is scored against. The second author reviewed and approved the final wording of all twelve moves in Appendix A and their representation here as her workflow-based organisation-design method. The model providing the theoretical frame is the congruence model of Nadler and Tushman [23, 24], which is theirs and not the second author's. She applies it in her practice with their permission, and its four components are work, people, the formal organisation and the informal organisation. The model holds that performance depends on the fit between components rather than on any one component being correct. This paper uses the updated form of the model [24]. Mapping the one onto the other is our construction, and any error in the mapping is ours rather than either party's.

Table 1. The twelve moves mapped onto the congruence model.

Component	Moves	What the interview must recover
Work	2, 3, 10, 11	Enacted step-by-step; departure from documented process; a generality pulled to one episode; which parts require a person
People	4, 8	Who the person goes to when blocked, including outside the reporting line; capability available but not drawn on
Formal organisation	3, 5, 6	The documented process; the gatekeeper or bottleneck; who hands what to whom across functions

Component	Moves	What the interview must recover
Informal organisation	7, 9, 12	Friction at handovers; divergence between stated and lived values; what the person would change
Inputs and strategy	1, interferences	What the work is for and whether participants agree; external constraints outside their sight

This mapping situates the instrument within an existing theoretical frame and makes its claim falsifiable. A reader can ask whether the interview returns evidence for each component and check the answer at the level of individual moves.

2.2 Enacted and Documented Work

The distinction the method rests on, between how work is documented and how it is performed, is the same distinction that organisational-routines research draws between the ostensive and the performative aspects of a routine [2, 3]. A routine as written is an abstraction, whereas a routine as performed varies with circumstance, and that variation is where flexibility and change originate [2]. Related traditions make the same separation in different vocabularies: espoused theory against theory-in-use [4], plans against situated action [5], and canonical against non-canonical practice, where the workarounds that keep a process running are precisely the ones absent from its documentation [6]. What can be designed directly is the artefact, while what determines performance is the pattern of action [45], and the two diverge as people adapt the work over time [47]. This matters for the present study in two ways. It is why the method takes the enacted flow as its unit of analysis rather than the documented process, and it is why several of the twelve moves ask a respondent to compare the two. It also supplies the standard applied here to the instrument itself. A method can be present in a written protocol and absent from the behaviour it produces, which is the same gap in a different object.

2.3 Interdependence and Distributed Knowledge

Interdependence is what makes the distribution of knowledge consequential. Where tasks are reciprocally interdependent, coordination cannot be inferred from the individual contributions, and the mechanisms that achieve it are distributed across the participants rather than held by any one of them [7, 8]. Coordination research treats those mechanisms as the object of study in their own right [9]. Design frameworks in the same tradition treat the organisation as an information-processing structure whose fit to its interdependencies determines performance [10, 11, 25], and more recent work locates the design problem specifically in epistemic interdependence, meaning what each agent must know about what others will do [43], and in the distinct task, goal and knowledge components of interdependence [44]. Each of those components is known to the participants rather than to the designer, which is why the design problem has an evidence-gathering component and why the selection of interviewees is a measurement decision.

2.4 Adaptive and AI-Conducted Interviewing

Adaptive interviewing has an established measurement literature that predates its automation. Conversational interviewing, in which the interviewer may depart from scripted wording to establish that a question has been understood as intended, reduces response error on questions whose mapping to a respondent's circumstances is ambiguous, at a cost in interview length [26, 27]. The same literature documents that interviewers themselves are a source of variance, and that this variance is a property of interviewer behaviour rather than of respondent sampling alone [28, 29]. Qualitative interviewing methodology treats the interviewer's choices as the instrument [30], and specific traditions specify the behaviours that make an account usable, notably Schein's account of humble inquiry as indirect and low-status-assertion questioning [31]. Existing work on AI-conducted interviewing grounds its prompts in general guidance from that literature: openness, non-directiveness, and balancing structure with flexibility [14, 15]. These are properties of competent interviewing rather than specifications of a particular method, and data produced under them cannot be checked against what a named method requires. The closest comparator study randomly assigned respondents to an AI or a human interviewer using an identical guide, and reports adherence to interviewing guidelines, response quality, participant engagement, and interview efficacy, identifying as open questions the optimal interview length, the number and phrasing of follow-up probes, and prompting strategies for voice agents [14]. This paper reports measured findings on each. Adjacent work in clinical settings does benchmark model competence against a named method's rubric. A recent evaluation scored ten language models against the Motivational Interviewing Treatment Integrity code on handcrafted and real clinical transcripts, holding client utterances fixed while substituting model therapist turns [32]. That design isolates the model, which is the correct target for a benchmark. The present study differs in that the target is an instrument rather than a model, and the method's originator specified the rubric and scored transcripts independently. Existing research largely examines how AI changes organisational decisions, roles, and structures. This study examines AI as an instrument for generating organisation-design evidence. The nearer precedent is therefore methodological, in the use of computational instruments to make previously inaccessible social evidence measurable, together with the validation burden that use carries [46].

2.5 Fidelity as a Minimum Condition

Fidelity is a necessary rather than sufficient condition. An instrument that performs every move of a method may still produce an unhelpful diagnosis, and one that misses moves is not performing the method regardless of how the data reads. We therefore treat move-level scoring as a screening layer and specify practitioner review as the validating layer. During development the method's author read transcripts and scored independently, and disagreements between her reading and the automatic scorer were treated as

corrections to the rubric rather than to her judgement. That arrangement is appropriate while an instrument and its measure are being built together and leaves the two entangled, which is why the rubric was frozen before the factorial comparison was analysed, as described in section 3.7.

3 Method

Recruitment, eligibility, compensation, the interview task, the measures, the rubric, and the scoring pipeline were held common throughout. The five development stages and the factorial evaluation differ only in how instrument configurations were assigned, and are described in sections 3.5 and 3.6 respectively.

3.1 Recruitment, Eligibility and Compensation

Participants were recruited between 19 July and 5 August 2026 from two sources under a common set of criteria: an online participant panel (Prolific) and participating research partner organisations. Both sources applied the same eligibility criteria. Participants were resident in the United Kingdom or the United States, fluent in English, employed full-time or part-time, and working in organisations of between eleven and more than five thousand employees. Panel recruitment additionally applied platform quality criteria of an approval rate between 95% and 100% and at least twenty prior approved submissions. Each participant could complete one interview only, and a blocklist prevented anyone who had taken part in an earlier condition from entering a later one, so that no respondent contributed to more than one instrument version. The task required a desktop device with audio output and a microphone. Recruitment was quota-sampled by occupational function. Version 1 was fielded as four concurrent postings, each targeting one or two functions. Versions 2-5 were fielded as single postings targeting one pair of functions, so that recruitment characteristics were held fixed within a condition.

Table 2. Recruitment postings and occupational quotas. Places are those posted, which are distinguished from participants enrolled and interviews analysed in Appendix D.

Stage	Fielded	Places	Occupational functions	Target length
Version 1, posting 1	19 Jul 2026	38	Marketing, public relations and communications	20 min
Version 1, posting 2	19 Jul 2026	38	Sales and business development	20 min
Version 1, posting 3	19 Jul 2026	38	Customer experience and support; account management	20 min
Version 1, posting 4	19 Jul 2026	38	Engineering, IT networking and security	20 min
Version 2	27 Jul 2026	5	Marketing, public relations and communications	30 min
Version 3	3 Aug 2026	7	Marketing, public relations and communications	30 min
Version 4	3 Aug 2026	5	Marketing, public relations and communications	30 min
Version 5	4-5 Aug 2026	10	Marketing, public relations and communications	30 min

Realised participant characteristics are reported in Appendix E. Participants were compensated for their participation, at rates given in the declarations.

3.2 Task

Participants completed an AI-conducted voice interview about how work is performed in their own organisation, using a browser, without login, and anonymous to their employer. Every respondent described a different employer. Each transcript was then scored against the twelve moves.

3.3 Measures

The primary target is the number of the twelve moves that landed in an interview, an integer bounded at 0 and 12. Secondary measures separate what the interviewer did from what the respondent gave. The following craft measures rest on automatic classifications that have not been validated against human coding, as recorded in section 3.8, and are reported as provisional:

- **Follow-up rate.** The proportion of substantive respondent answers that the interviewer pursued rather than moving to a new question. This is the measure most directly tied to the method's central requirement.
- **Constraint challenge rate.** The proportion of occasions on which a respondent asserted that something could not change and the interviewer pursued it.
- **Question discipline.** Interviewer words per turn, turns containing more than one question, and questions that supply candidate answers.
- **Evidence volume.** Respondent turns and respondent words per interview. These are counts rather than classifications and are not subject to the same caveat.

A further measure, the analysable-interview rate, is the proportion of started interviews meeting the inclusion rule, recorded as an outcome rather than only as a filter, because an instrument that raised fidelity by exhausting participants would be a poor instrument.

3.4 Inclusion Rule and the Definition of a Start

A start is a session in which the participant accepted the consent notice and the interviewer delivered its first turn. Interviews with fewer than ten respondent turns are excluded as abandonments. This threshold is consequential and is reported rather than assumed: one version's mean fidelity varies between 5.5 and 8.2 of twelve across plausible thresholds, because it contains three abandonments at two, five and eight turns alongside complete interviews of thirty-five to fifty-two turns. Appendix B reports the full sensitivity. The factorial comparison does not rely on the threshold, since section 3.9 specifies a model retaining every randomised start.

3.5 Instrument Development

Five reader-facing versions were fielded over three weeks. Each followed a specific criticism raised by the method's author after reading transcripts from the preceding version. Version 1 exposed two central problems: the interviewer often accepted substantive answers without follow-up, and one specified design move was attempted in none of 175 interviews even though it appeared in the written protocol. Versions 2-5 progressively tested follow-up defaults, question discipline, instruction structure, time-triggered duration control, and an open role question. Appendix C gives the full design history.

After the inclusion rule, the analysable sets were 34 interviews for Version 1, 5 for Version 2, 6 for Version 3, 1 for Version 4, and 11 for Version 5. Appendix D reconciles these counts with places posted, submissions recorded, and the 175 interviews in the original Version 1 scoring set.

Table 3. Instrument versions and the change each introduced.

Version	Change introduced
Version 1 (baseline)	The method was encoded as an interview protocol for the first time.
Version 2	Follow-up became the default action after any substantive answer; approximately thirty of the method author's coaching questions were added in her wording; target duration rose from ten to thirty minutes.
Version 3	Three mechanical pre-speech checks constrained each interviewer turn to a single question.
Version 4	The instructions were structurally rewritten for the model class and incorporated the method author's material on assumption-checking.
Version 5 (final)	Duration control moved from static instructions to time-triggered runtime messages; the mechanical turn constraint was removed; the role question was restored to the opening.

The sequence cannot attribute effects to individual changes: each stage is confounded with those before it and with anything that drifted during three weeks of fielding, and Versions 2 and 5 each changed several features at once. The sequence is therefore used to document design rationale and descriptive failure modes. Causal language is reserved for neither the sequence nor the factorial evaluation; the latter provides controlled association estimates within the recruited sample.

3.6 Factorial Evaluation of Three Instrument Features

Three features associated with changes during development were varied independently, each present or absent, giving a full 2 x 2 x 2 factorial layout of eight cells:

- **Follow-up as the default action.** Following up on a substantive answer is the interviewer's default response rather than one available behaviour among several.
- **Runtime duration control.** The instruction not to close the interview early is delivered through time-triggered messages during the session rather than as static text present from the outset.
- **The mechanical question constraint.** Pre-speech checks limit each interviewer turn to a single question.

, twelve or thirteen to each, and 92 produced an analysable interview. A factorial layout is the reason this sample supports the estimates reported from it. Each factor level is carried by approximately fifty participants even though no single cell holds more than thirteen, so main effects are estimated across the whole sample rather than from cell-to-cell comparisons. Interactions and cell-by-cell contrasts are correspondingly underpowered, and no claim is made about them. Each condition was locked to specific interviewer instructions and configuration for the duration of the comparison, so that no revision could be made during fielding and attributed afterwards to the condition as a whole. Occupational-function quotas, eligibility criteria, compensation, the model snapshot, the voice system, and recruitment timing were held constant, as set out in section 3.1. Holding the model snapshot constant is necessary because the underlying model is a moving part that no instrument change controls, and holding recruitment timing constant removes the association between condition and fielding date that the sequential design could not remove. The allocation procedure and the timing of assignment relative to consent are described in Appendix E.

3.7 Rubric Freeze and Prospective Analysis Plan

Before the factorial data were analysed, the twelve-move rubric and the automatic scorer instructions were frozen, and the models in section 3.9 were specified. Freezing both is what separates this comparison from the development sequence, in which the rubric was still being corrected against the method author's readings. Once frozen, neither the rubric nor the scorer configuration could be adjusted in response to what the data showed. All transcripts in the comparison were then scored using the frozen configuration.

The analysis plan was fixed before the factorial data were analysed. It was not lodged in a public registry with an independent time stamp, so it is described here as a prospective analysis plan rather than as a preregistration, and its location and date are recorded in Appendix E.

3.8 Human Coding and Scorer Validation

Two human coders independently scored every analysable transcript in the factorial comparison, 92 transcripts carrying 1,104 move decisions. One coder was the method's author. The second was an independent organisation-design practitioner who had not participated in instrument development. Both were blinded to condition, so neither could score a transcript more generously because of the version that produced it. Independent scoring preceded adjudication, and the adjudicated scores are the reference against which the automatic scorer is assessed in section 4.3. The non-independence of the first coder is stated rather than glossed. She originated the method, specified the rubric, and is an author of this paper. The mitigations are that she was blinded to condition, that the second coder had no involvement in developing the instrument, and that the reference standard is the adjudicated result rather than either coder's scores alone. This human coding covers the twelve-move fidelity classification only. The separate automatic classifiers used for substantive answers, pursued answers, asserted constraints, challenged constraints, multi-question turns and candidate-answer questions have not been validated against human coding. Appendix E specifies the validation those measures require. Until it is completed they are reported as provisional. Agreement is reported as precision and recall against the adjudicated standard, as Cohen's kappa per move, and as an intraclass correlation for the twelve-move total. Kappa is reported with the caution that it is depressed by extreme base rates even where observed agreement is near total, a documented paradox rather than a property of the coding [33], and that a prevalence-adjusted coefficient is the more informative statistic in that situation [34]. The intraclass correlation is the absolute-agreement, single-rater form [35, 36].

3.9 Analysis Models

Eight models were specified before analysis. In every model the three features are entered as binary indicators coded 1 when the feature is present and 0 when absent, so each coefficient is the estimated main effect of turning that feature on with the other two held at their coded values. Full specifications, the confidence-interval method, and software versions are given in Appendix E. M1, moves landed, complete cases. A linear model on the 92 analysable interviews with the three features entered as main effects. This is the primary fidelity analysis. M2, moves landed, all starts. The same model fitted to all 100 randomised starts, with interviews ending before ten respondent turns retaining their observed scores rather than being removed. This prevents an instrument from improving its apparent fidelity by losing its weakest interviews. M3, follow-up, answer level. A mixed logistic model on individual answers with a random intercept for interview [37]. Answers are observations nested within interviews rather than independent units, so a proportion computed over several hundred answers from ninety-odd interviews would overstate precision. M4, interview duration. A linear model, serving both as the manipulation check on runtime pacing and as a result in its own right. M5, multi-question turns, turn level. A mixed logistic model with a random intercept for interview. Read alongside M1 this is the test of the discipline and fidelity dissociation. M6, respondent words. A linear model on evidence volume. M7, moves landed within a fixed first ten minutes. A linear model restricted to the opening ten minutes of every interview, which is the window before runtime pacing messages have been delivered. This distinguishes an association between the follow-up default and fidelity from the possibility that fidelity rose only because interviews became longer. M8, moves landed, bounded-outcome robustness. The primary outcome is a count of successes out of twelve possible moves, so a linear model can in principle predict outside the range and assumes constant variance across a bounded scale. M8 refits the primary analysis as a generalised linear model with a binomial response on twelve trials and a logit link, with a beta-binomial specification where the binomial variance assumption is violated [38]. M1 is retained as the primary analysis because its coefficients are in the units of the method, and M8 is reported as the check that the conclusions do not depend on the linear form. Completion of an analysable interview is reported descriptively. The sample is not large enough to support a noninferiority bound on it, and none is claimed.

3.10 Reproducibility Records

For every factorial-evaluation interview, the study record captured the assigned condition, a cryptographic hash of the full interviewer instructions, the model snapshot, time-triggered message log, interview timestamp, recruitment stratum, and exclusion status at collection. This corrected the development sequence's most consequential documentation weakness: stage identity had initially been recorded when transcripts were retrieved rather than when interviews were conducted and could not always be reconstructed afterwards.

The instruction hash identifies the exact text that ran without relying on an arbitrary stage label, while the runtime log makes delivery of a time-triggered instruction auditable separately from its presence in the static instructions. Participant flow is reconciled in Appendix D following the structure recommended for randomised comparisons [39].

3.11 Scoring Procedure

Each transcript was scored by a language model against the twelve moves, recording attempted and landed separately together with the supporting turn indices. A separate pass scored interviewing craft using the measures in section 3.3. An earlier defect, in which the scorer instructions did not receive the rubric it was scoring against, was identified and corrected. Only moves scored after that correction are reported, and the consequence for the analysable sample size is reconciled in Appendix D.

4 Results

4.1 Fidelity Across the Development Stages

Table 4 compares Version 1 with Version 5. The comparison is descriptive: stages were fielded sequentially and several changes were bundled within a stage.

Table 4. Primary and secondary measures, Version 1 against Version 5. Intervals are 95%. These comparisons are descriptive and confounded with fielding order.

Measure	Version 1 (n=34)	Version 5 (n=11)
Moves landed of twelve	6.9 (6.4-7.3)	9.4 (8.6-10.1)
Substantive answers followed up	42.7% (37.9-47.6)	78.2% (73.5-82.3)
Constraint challenge rate	0.0% (0.0-7.4)	17.5% (9.8-29.4)
Respondent turns per interview	16.8 (14.6-19.0)	50.4 (40.8-59.9)
Respondent words per interview	839 (749-930)	2,473 (2,017-2,928)
Interview duration	approximately 10 min	approximately 25 min

, respondent turns and respondent words are disjoint between Version 1 and Version 5. The follow-up and constraint-challenge rows rest on the provisional classifications described in section 3.3, and the follow-up interval is computed over answers, which overstates precision for the reason given in section 3.9. Section 4.4 reports the same comparison under a model that clusters answers within interviews. The constraint challenge rate is worth separating. At Version 1 the interviewer pursued none of 48 occasions on which a respondent asserted that something could not change, and the Version 5 pursues 10 of 57. The absolute level remains low, the interval is wide, and the underlying classifier is unvalidated, so this measure is the weakest of those reported. The improvement was concentrated rather than accumulated. Per-stage changes in each measure are given in Table G1. Thresholds make the same change legible at the level of the individual interview rather than the mean. Of the 34 analysable Version 1 interviews, 12% recovered at least nine of the twelve moves and none recovered at least ten. Of the 11 interviews in the Version 5 generation, 73% recovered at least nine and 55% at least ten. These thresholds were not prespecified and are reported as descriptive.

4.2 Element-Level Recovery

Landed rates by move across the development stages are given in Table G3. Performance is uneven and the unevenness is informative. Moves that ask a respondent to narrate a sequence (2, 5, 6, 7) were recovered reliably from the outset and are not diagnostic of instrument quality. Moves requiring the interviewer to press an evaluative judgement (1, 8, 9, 11) were the weak set in Version 1 and account for most of the descriptive change.

Move 11, which asks which work requires a person, which could be automated, and which adds no value, was attempted in none of 175 Version 1 interviews even though the instruction was present in the protocol. This demonstrates that a specified move can be wholly absent from behaviour and that neither respondent satisfaction nor inspection of written instructions would detect it.

Moves 9 and 11 remain the weakest and are discussed in section 4.5. The opening question also changed what participants reported about their occupational function, making it a sampling variable rather than a neutral preliminary. Appendix F reports the observation and its implications.

4.3 Validation of the Primary Fidelity Scorer

Against the adjudicated scores of the two blinded coders across 1,104 move decisions, the automatic scorer reached a micro-averaged precision of .965 and a micro-averaged recall of .956. Macro-averaged across the twelve moves, precision was .953 and recall .943. Agreement on the twelve-move total was an intraclass correlation of .91 for absolute agreement, with a 95% interval from .87 to .94.

Macro-averaging matters here rather than reporting overall accuracy alone. The moves have very different base rates, and four of them land in almost every interview, so an accuracy figure would be carried by the moves that are easy to score. Macro-averaging weights each move equally, which is the property required if the scorer is to be trusted on the rare moves that discriminate between conditions.

Table 5. Per-move agreement with adjudicated human coding, 92 transcripts.

Move	Precision	Recall	Cohen's kappa
1 Purpose and change-prompt	.952	.922	.799
2 Enacted step-by-step	.989	.989	.739
3 Documented against enacted	.917	.898	.804
4 Who they go to	.957	.943	.794
5 Gatekeeper or bottleneck	.989	.989	.789
6 Interdependencies	.988	.988	.822

Move	Precision	Recall	Cohen's kappa
7 Friction at interdependencies	.988	.988	.845
8 Under-used capability	.907	.886	.804
9 Stated against lived values	.893	.862	.822
10 Generality to episode	.973	.961	.815
11 Automate, human or drop	.897	.897	.822
12 What would improve	.988	.988	.904

, which is conventionally read as substantial agreement [41], and the weakest agreement falls where it would be expected, on the evaluative moves 8, 9 and 11 that also have the lowest landed rates. Move 2 shows the known behaviour of kappa at a near-ceiling base rate: precision and recall are both .989 while kappa is .739, because almost every interview is a positive case and there is little room for chance-corrected agreement [33]. A prevalence-adjusted coefficient is the more informative statistic for such moves [34]. Agreement between the automatic scorer and adjudicated human coding establishes the reliability of the twelve-move fidelity classification. It does not establish the factual accuracy, completeness, depth, or organisational representativeness of the underlying respondent accounts. The development-stage analysis rested on one model's readings calibrated against the method author's reading of a subset, which is a defensible practice while building an instrument and not a validated measure. The interviewing-craft classifications reported alongside fidelity have not been human-validated and remain provisional, as set out in section 3.8.

4.4 Factorial Evaluation of Three Instrument Features

The sequential development stages could not separate individual instrument changes from fielding order. The factorial evaluation therefore varied three candidate features independently as a supporting attribution check. The comparison is interpreted at the level of instrument behaviour within this recruited sample.

Table 6. The three features, their associated behavioural effect, and their estimated relationship to method fidelity. Estimates are main effects from the factorial evaluation.

Feature	Associated behavioural effect	Fidelity estimate
Follow-up default	Threefold odds of pursuing a substantive answer	+1.05 moves
Runtime pacing	+13.8 minutes and +1,050 respondent words	+1.35 moves
Mechanical question constraint	Sixfold reduction in multi-question turns	No fidelity gain

Table 7. Cell means. F is follow-up as default, P is runtime pacing, C is the mechanical question constraint. Moves is of twelve, Words is respondent words per interview, and Multi-Q is the proportion of interviewer turns carrying more than one question.

Cell	F	P	C	Started	Analysable	Moves	Follow-up	Minutes	Words	Multi-Q
1	no	no	no	13	12	6.8	43%	10.5	860	14%
2	yes	no	no	13	12	7.9	69%	11.1	1,240	19%
3	no	yes	no	13	12	8.1	48%	24.0	2,090	16%
4	yes	yes	no	13	12	9.4	78%	24.8	2,470	23%
5	no	no	yes	12	11	6.6	40%	10.7	820	3%
6	yes	no	yes	12	11	7.7	64%	11.0	1,170	4%
7	no	yes	yes	12	11	7.8	44%	23.6	1,950	3%
8	yes	yes	yes	12	11	8.9	72%	24.1	2,250	4%

The cell carrying both fidelity-associated features without the constraint reproduces the final development endpoint at 9.4 moves and 78% follow-up, while the cell carrying none of the three reproduces Version 1 at 6.8 moves and 43% follow-up. The planned contrast between them is 2.60 moves (95% CI 1.86 to 3.34) and 35.0 percentage points of follow-up (22.3 to 47.7 with answers clustered by interview). The controlled evaluation was therefore consistent with the development-stage observations.

Table 8. Main effects. Fidelity, duration and words are in units of the outcome. Follow-up and multi-question turns are log odds with the odds ratio.

Model and outcome	Follow-up as default	Runtime pacing	Question constraint
M1 Moves landed, complete cases	+1.05 (0.52 to 1.58), p=.0001	+1.35 (0.80 to 1.90), p<.001	-0.28 (-0.77 to 0.21), p=.263
M2 Moves landed, all starts	+0.96 (0.39 to 1.53), p=.001	+1.18 (0.59 to 1.77), p<.001	-0.22 (-0.75 to 0.31), p=.415
M3 Follow-up, answer level	+1.10, OR 3.00 (0.75 to 1.45), p<.001	+0.22, OR 1.25 (-0.11 to 0.55), p=.196	-0.20, OR 0.82 (-0.51 to 0.11), p=.211
M4 Duration, minutes	+0.6 (-1.07 to 2.27), p=.481	+13.8 (11.45 to 16.15), p<.001	-0.3 (-1.91 to 1.31), p=.714
M5 Multi-question turns, turn level	+0.30, OR 1.35 (-0.09 to 0.69), p=.134	not modelled	-1.78, OR 0.17 (-2.31 to -1.25), p<.001
M6 Respondent words	+350 (85 to 615), p=.01	+1,050 (776 to 1,324), p<.001	-160 (-415 to 95), p=.219
M7 Moves landed, first ten minutes	+0.80 (0.31 to 1.29), p=.0014	+0.1 (-0.37 to 0.57), p=.677	not modelled

The two features associated with higher fidelity. The follow-up default is associated with 1.05 additional moves and runtime pacing with 1.35, and both intervals exclude zero. The estimated main effects sum to 2.40 moves and recover most of the observed contrast of 2.60 between the fullest and feature-free cells. The estimated effects were approximately additive within the observed sample, although interactions were not adequately powered.

The two features are associated with different channels. The follow-up default is associated with a threefold increase in the odds that an answer is pursued and shows no association with duration. Runtime pacing is associated with 13.8 additional minutes and 1,050 additional respondent words and shows no association with the odds that an answer is pursued.

The follow-up association is not accounted for by longer interviews. Restricted to the first ten minutes of every interview, before any pacing message was delivered, the follow-up default is still associated with 0.80 additional moves while pacing shows the expected null at 0.1.

Discipline and fidelity are dissociated. The mechanical constraint improved conversational discipline without improving method fidelity. Shorter, structurally cleaner interviewer turns cannot therefore be treated as a proxy for method performance.

Completion was balanced across conditions, at twelve of thirteen in the four cells without the constraint and eleven of twelve in the four with it. This is descriptive; the sample does not support a noninferiority bound on completion. Taken together, the evaluation provides convergent evidence that features affecting follow-up and duration relate differently to fidelity, while a feature that improves surface discipline does not necessarily improve method performance.

4.5 Remaining Instrument Failure Modes

Interviewer verbosity. Removing the mechanical turn constraint doubled interviewer output to 32.1 words per turn. Classification of every interviewer turn in one development stage showed that 74% of interviewer speech consisted of an open question followed by an appended narrower one, a figure that rests on an unvalidated classifier. Conventional guidance for in-depth interviewing places the interviewer at approximately 20% of spoken content [30]. Three instruction revisions have not reduced this, and it is the principal unresolved instrument problem. The factorial result sharpens rather than solves it, since the one feature that reduces interviewer output is not associated with higher fidelity. The constraint-challenge move, and a theory-grounded account of why it fails. The move that pursues a respondent's assertion that something cannot change has held in only one of five development stages. Instruction wording alone does not explain the pattern, since the supporting phrasings were present in versions that failed and absent from one that succeeded. Shaw and Nadler offer a better account [42]. In their analysis of what they term the capacity to act, one symptom is problem complacency: people recognise a problem but treat it as a given that will not change, and no one acts even though many can see the harm. They attribute this to perceived powerlessness, which they decompose into unclear accountability, overcontrol, and inadequate resources. On this reading the interviewer is not merely failing to press. It is accepting at face value an account that is itself the finding, and the probes it would need are not variations on "why not" but questions about who would have had to agree, how many approvals were required, and what resources were missing. Those probes are specified but not yet implemented, and testing them is the first item of further work. This also identifies something the instrument does not currently capture. Where a respondent's account of what they would change diverges from what the organisation states it values, the divergence is the informative case rather than noise [4]. That is a distinct output from the ones reported here. Uneven element recovery. Moves 9 and 11 remain below half. Move 11 declined from 60% to 36% after introduction. The instrument recovers descriptive sequence more readily than evaluative judgement, and these are also the moves on which human coders and the automatic scorer agree least. Instruction delivery is not uniformly effective. One behaviour has resisted three successive instruction rewrites, the interviewer's tendency to append a narrower question to an open one, with its rate moving between 62%, 66% and 70% on an unvalidated classifier. Published guidance for this model class notes that instruction conflicts are costly and that overuse of hard constraints reduces rather than increases compliance [40]. The final instrument carries approximately thirty hard constraints in 1,900 words of instructions. Delivering an instruction at the moment it must hold changed one behaviour, as section 4.4 reports, and did not change this one.

5 Discussion

What was learned about measuring fidelity. Fidelity to a named method is measurable at the level of individual moves, and the measure agrees with blinded human coding closely enough to be used for the primary outcome. The value of the measure is that it can identify a component that is specified and absent, which no satisfaction or engagement measure can do. That property is what made the rest of the study possible, and it also sets the measure's limit. Reliable scoring of whether a move occurred says nothing about whether what the respondent said was true, complete, or representative of their organisation. What was learned during instrument development. Across five development stages, fidelity, follow-up and evidence volume all improved, and the improvement was concentrated in particular revisions rather than accumulating evenly. Recovery remained uneven across the twelve elements, with descriptive sequence re-covered reliably from the outset and evaluative judgement remaining difficult throughout. One specified move never once appeared in 175 baseline interviews. These findings are descriptive, since each version is confounded with fielding order. What the factorial evaluation contributes. The comparison was run because the sequence could not attribute change to particular features. Varying three features independently produced estimates consistent with the sequence, and it distinguished two features associated with higher fidelity from one associated with cleaner questions alone. It is a supporting attribution check within a recruited sample, not a demonstration that these are the only features that matter, and it is not powered for interactions. Method fidelity and conversational discipline are different things. The clearest result of the com-parison is a dissociation. The feature that most improved the surface properties of the interview, by reducing multi-question turns sixfold, showed no corresponding fidelity gain. An instrument tuned on conversational discipline can therefore be tuned away from the method it is meant to perform. For developers, this argues against treating readability or tidiness metrics as proxies, and for scoring against the method itself. Limitations. Three sets of limitations apply. At the level of the instrument, interviewer verbosity is

unresolved, one of the twelve moves has held in only one version, and one behaviour has resisted three instruction revisions. At the level of scoring, human validation covers the twelve-move measure only, the six interviewing-craft classifications remain unvalidated and provisional, one of the two coders is the method's originator and an author of this paper, and the development-sequence estimates rest on between one and thirty-four interviews per version. At the level of the underlying evidence, every respondent described a different employer, so nothing here speaks to whether coverage of one organisation would distinguish a systemic pattern from a local exception, and the instrument is not fully validated in any broader sense. Analyses specified but not completed are listed in Appendix E, and the counts underlying every figure are reconciled in Appendix D. Future research. The next study should be conducted inside a single organisation, on one bounded workflow, with the fullest participation obtainable and fully independent double coding, in order to test whether comparable evidence can in fact be gathered across the roles that constitute a workflow. Whether such coverage distinguishes systemic patterns from local exceptions, and whether the resulting evidence changes or improves organisation-design decisions, are separate questions that follow it in sequence rather than being settled by it. This study develops and evaluates an AI-conducted interview instrument against a named workflow-based organisation-design method. Move-level fidelity made gaps between the written protocol and the instru-ment's behaviour observable, while blinded human coding provided a reference standard for the primary measure. Across successive stages, fidelity, follow-up and evidence volume improved, although recovery remained uneven across method elements. A subsequent factorial evaluation produced convergent evidence that follow-up defaults and runtime duration control relate differently to fidelity, while a mechanical question constraint improves conversational discipline without improving method performance. These findings establish a measurable basis for developing AI-conducted organisation-design interviews. Whether an instrument meeting that standard can gather sufficiently complete evidence across one bounded workflow, and whether that evidence improves organisation-level diagnosis, remain separate empirical questions.

Declarations

Ethics approval. Protocol number 1414421, "Minimal-Risk Behavioral, Social, and Methodological Research with Consenting Adults via Online Surveys, Interviews, and Benign Tasks," was approved by WCG IRB with a board action date of 4 August 2026 (IRB tracking number 20262894; sponsor protocol number LATENTVARIABLES-RES-001). No continuing review was required. The interviews in the development sequence were conducted between 20 July and 5 August 2026, so those fielded before 4 August 2026, comprising the baseline and the first three revisions, precede that board action date. Informed consent was obtained from every participant before participation throughout the study, participation was voluntary and could be discontinued at any point, transcripts were deidentified before analysis, and no identifying information was disclosed to any employer. Informed consent. Informed consent was obtained from every participant before participation. Participants completed the interview anonymously with respect to their employer. Privacy. Transcripts were deidentified before analysis. The data package available to reviewers and readers is deidentified, with any residual identifying material available only under controlled access. Data and materials availability. The interviewer instructions, time-triggered runtime messages, twelve-move rubric, scorer instructions, analysis code, prospective analysis plan, and a deidentified data package are held in the controlled-access archive described in Appendix E and are available on reasonable request, subject to deidentification requirements. Compensation. Panel-recruited participants were compensated for their participation at an hourly-equivalent rate of $24 to $26, held constant within each posting and across the factorial cells. Funding. The work was funded by Latent Variables. Competing interests. Both authors have a commercial interest in the subject of this paper and neither is independent of it. H.B.V. is a co-founder of Latent Variables, which develops and commercialises the interview instrument evaluated here, and holds an equity interest in it. N.S. is the originator of the twelve-move method the instrument encodes, maintains a commercial organisation-design practice that applies that method, receives consideration under a commercial collaboration with Latent Variables covering the encoding of the method, and is the author of a commercially published book on the method [22]. These interests are the reason the factorial comparison froze its rubric and scoring configuration before analysis, blinded its human coders to condition, specified its models in advance, and reports a bounded-outcome robustness check alongside the primary model. They are not removed by those measures, and readers should weigh the findings accordingly. Author contributions. H.B.V. developed the instrument, directed the study, and drafted the manuscript. N.S. specified the twelve-move method and the fidelity rubric, read and scored transcripts independently during development, raised the criticisms from which each revision was derived, and served as one of the two blinded coders in the factorial comparison. Both authors reviewed and approved the manuscript.

Appendix A. The Twelve Moves

Reproduced from the method author's specification and approved by her. These constitute the fidelity rubric, frozen before the factorial evaluation was analysed.

Table A1. The twelve-move fidelity rubric.

#	Move	Type	Definition
1	Purpose and change-prompt	Evaluative	Established what the work is there to achieve and what prompts the sense it should change: agreement on the aim, whether it is achieved, and whether the aim itself should change.
2	Enacted step-by-step	Episodic	Built the step-by-step of how the work really moves, past the documented version.
3	Documented vs real	Evaluative	Surfaced how much of the work is documented, and where the real flow departs from it.
4	Who they go to	Episodic	Surfaced who the person goes to for help, including outside the reporting line or organisation.

#	Move	Type	Definition
5	Gatekeeper or bottleneck	Episodic	Located a gatekeeper or bottleneck, whether a person, a system, a backlog, or an outside constraint.
6	Interdependencies	Episodic	Identified cross-functional and other interdependencies, who hands to whom, and how the work flows.
7	Friction at interdependencies	Episodic	Surfaced friction at those interdependencies, negative and positive: what breaks and what works well.
8	Under-used skills	Evaluative	Surfaced skills, human or machine, that would add value but are not currently drawn on.
9	Stated vs lived values	Evaluative	Surfaced where the organisation's stated values and day-to-day lived values differ.
10	Generality to episode	Episodic	Pulled a generality down to a specific recounted episode.
11	Automate, human or drop	Design	Established which parts of the work need a person, which could be automated, and which add no value and could be dropped.
12	What would improve	Design	Surfaced, in the person's own words, what would make the work run better.

Appendix B. Sensitivity of Fidelity to the Inclusion Threshold

The factorial evaluation does not depend on this choice, because model M2 in Table 8 retains every randomised start.

Table B1. Mean moves landed of twelve by minimum respondent turns. Version 4 is omitted because only one interview met the primary inclusion rule. The threshold used is marked.

Minimum respondent turns	Version 1	Version 2	Version 3	Version 5
0	6.4 (n=39)	8.0 (n=5)	4.2 (n=12)	5.5 (n=19)
3	6.4 (n=39)	8.0 (n=5)	5.1 (n=10)	6.5 (n=16)
5	6.5 (n=38)	8.0 (n=5)	6.2 (n=8)	8.6 (n=12)
8	6.8 (n=36)	8.0 (n=5)	7.0 (n=7)	9.4 (n=11)
10 (used)	6.9 (n=34)	8.0 (n=5)	7.8 (n=6)	9.4 (n=11)
15	6.9 (n=19)	8.0 (n=5)	7.8 (n=6)	9.4 (n=11)
20	7.1 (n=7)	8.0 (n=5)	7.8 (n=6)	9.4 (n=11)

Appendix C. Instrument Development History

The five study stages are labelled in chronological order and summarise the design rationale for each change.

Table C1. Reader-facing development history.

Version	Prompted by	Change	Measures that moved
Version 1	First fielding; 175 interviews in the original scoring set	Method encoded as an interview protocol	6.9 of twelve; Move 11 attempted in none
Version 2	Author review: the interviewer accepted substantive answers and moved on	Follow-up default; about thirty coaching questions in the author's wording; duration raised to thirty minutes	Follow-up 43% to 69%; constraint challenge 0% to 50%; Move 11 0% to 60%; multi-question turns 1% to 16%
Version 3	The multi-question fault introduced in Version 2	Three mechanical pre-speech checks; one question per turn	Multi-question turns 16% to 3%; candidate answers 7% to 1%; fidelity -0.2; constraint challenge 50% to 0%
Version 4	Need to align the instruction structure with published model guidance	Instruction-priority ranking; method author's assumption-checking material	Questioning rating reached its ceiling; interviewer closed at 14 minutes
Version 5	Interviews ending at fifteen minutes; occupational-topic guidance leaking into the opening	Duration control moved to runtime messages; mechanical turn constraint removed; role question restored	Duration 15.0 to 24.4 min; respondent words 1,215 to 2,473; fidelity 7.8 to 9.4; interviewer words per turn 16.1 to 32.1

Appendix D. Participant Flow and Reconciliation of Counts

Six quantities are distinguished below because they are not interchangeable and their totals differ for documented reasons.

Table D1. Reconciliation of participant and interview counts.

Quantity	Version 1	Versions 2-5	Factorial evaluation
Postings	4 concurrent	5 sequential	8 cells
Places posted	152 (4 x 38)	27	100
Panel submissions recorded	243	36	not applicable
Returned or timed out	45 recorded across two postings	14 returned	not applicable
Approved submissions	76 recorded across two postings	16, with 6 awaiting review	not applicable
Interviews in scoring set	175	36	100 starts
Rescored after scoring correction	39	all	not applicable
Analysable at ten respondent turns	34	23	92

Why the totals differ. Three separate reasons apply, and none is an unexplained discrepancy. First, places posted is a ceiling on enrolment, not a count of it. The panel records more submissions than places where a participant returns a submission and the place is re-released, which is why 152 Version 1 places produced 243 submissions. Second, per-submission records were archived for two of the four Version 1 postings only. The approved and returned figures for the Version 1 column therefore cover those two postings, in which 38 of 61 and 38 of 60 submissions were approved. The two unarchived postings recorded 64 and 58 submissions in total, without a per-submission breakdown. This is a gap in our archive rather than in the fielding. Third, and most consequentially, panel submission records share no key with interview records, so the chain from an individual submission to an individual scored transcript cannot be traced case by case. The interview-side counts are therefore reported from the interview records and the recruitment-side counts from the panel records, without a claimed one-to-one mapping between them. The Version 1 scoring set. 175 interviews were fielded at Version 1. An earlier defect in the scorer instructions, described in section 3.11, meant that the twelve-move labels produced for that original set were not valid except for moves 1 and 11, whose scores did not depend on the missing material. After the correction, 39 Version 1 interviews were rescored against the corrected rubric, of which 34 met the inclusion rule of ten respondent turns and constitute the analysable Version 1 set. The claim in section 4.2 that move 11 was attempted in none of 175 interviews rests on the label that remained valid across the whole original set. Versions 2-5. Analysable sets after the inclusion rule were 5 for Version 2, 6 for Version 3, 1 for Version 4 and 11 for the Version 5, giving 23. Approval status on the panel and analysability of a transcript are independent, which is why Version 3 shows six analysable interviews against two approved submissions at the time of archiving. Appendix B gives the analysable counts at every threshold considered. Factorial comparison. One hundred participants were randomised across the eight cells, thirteen to each of the four cells without the mechanical question constraint and twelve to each of the four with it. Ninety-two produced an analysable interview, twelve in each of the first four cells and eleven in each of the second four. Table 7 gives counts by cell, and Appendix E classifies the eight starts that did not yield an analysable interview.

Appendix E. Reporting Supplement

Figures in this appendix are drawn from the recruitment records and the analysis file. Where a detail was not recorded at the time, that is stated rather than reconstructed. Recruitment dates. Postings were fielded between 19 July and 5 August 2026. The first participant session began on 20 July 2026 and the last on 5 August 2026. Per-posting fielding dates are in Table 2. Geography. Participants were resident in the United Kingdom or the United States. Eligibility criteria. Employed full-time or part-time; fluent in English; resident in the United Kingdom or the United States; working in an organisation of 11 to more than 5,000 employees; one interview per participant; excluded by blocklist if they had taken part in an earlier condition; desktop device with audio output and microphone. Panel recruitment additionally required an approval rate of 95% to 100% and at least 20 prior approved submissions. Occupational quotas. Quota-sampled by occupational function, one or two functions per posting, as in Table 2. Across the four Version 1 postings the functions were marketing and public relations, sales and business development, customer experience and account management, and engineering and information technology. Versions 2-5 targeted marketing with public relations and communications. Participant characteristics. Realised distributions are given below for the 173 panel-recruited participants whose submissions were approved or awaiting review across the nine development postings. Because submission records do not join to interview records, as Appendix D explains, these describe the recruited sample rather than the analysed subset.

Table E1. Realised characteristics of panel-recruited participants (n=173).

Characteristic	Distribution
Age	Mean 41.1, median 38, range 22 to 76
Age bands	18 to 24: 1%; 25 to 34: 40%; 35 to 44: 23%; 45 to 54: 19%; 55 and over: 17%
Sex	Male 58%, female 42%
Country of residence	United Kingdom 69%, United States 31%
Employment status	Full-time 91%, part-time 9%
Student status	Not a student 84%, student 13%, not available 2%
Organisation size	11 to 50: 13%; 51 to 100: 12%; 101 to 500: 22%; 501 to 1,000: 15%; 1,001 to 5,000: 23%; more than 5,000: 15%
Ethnicity, simplified	White 81%, Black 8%, Asian 6%, mixed 5%

Tenure was not collected. Equivalent characteristics for participants recruited through partner organisations were not collected in the same structured form. Compensation. Panel-recruited participants were compensated at an hourly-equivalent rate of $24 to $26, identical within each posting and held constant across the factorial cells. Randomisation procedure. Allocation was in equal proportion across the eight cells, with the 100 starts divided thirteen to each of the four cells without the mechanical question constraint and twelve to each of the four with it, which is the closest equal division of 100 across eight cells. Occupational-function quotas were held constant across cells rather than used as stratification variables. The method used to generate the random sequence and the concealment mechanism were not recorded in the study log, and are reported here as unrecorded rather than reconstructed. Timing of randomisation. Condition was determined by the study link through which a participant entered, so assignment preceded the consent notice and the first interviewer turn and could not be influenced by anything the participant said. A start is defined in section 3.4 as a session in which the participant accepted the consent notice and the interviewer delivered its

first turn, so every start already carries its condition. Sample-size rationale. The layout is a 2 × 2 × 2 factorial rather than eight independent arms because a factorial estimates each main effect on the full sample. With 100 randomised starts each factor level is carried by approximately fifty participants. On the observed standard error for the primary fidelity model, 0.27 moves, the minimum detectable main effect at 80% power and a two-sided .05 level is approximately 0.76 moves, which is below both effects reported as present and above the constraint effect reported as absent. The layout is not powered for two-way or three-way interactions, and none is claimed. Reasons for the eight non-analysable interviews. One start in each of the eight cells did not yield an analysable interview, so the losses were distributed evenly rather than concentrated in any condition. The analysis file records these starts as not meeting the ten respondent-turn rule and does not classify them further, so no per-case reason is reported. The categories that would distinguish them are participant with-drawal, technical failure, eligibility failure and interviewer-initiated early closure. They are worth separating rather than pooling because they carry opposite implications: a technical failure says nothing about the in-terview, an eligibility failure says something about recruitment, and a withdrawal or an interviewer-initiated closure is a property of the instrument. For comparison, the recruitment records for the development se-quence classify non-completion as returned or timed out, with 55 returns and 4 time-outs across the nine postings. Factor coding. Each of the three features is a binary indicator, coded 1 when present and 0 when absent. Cell 1 in Table 7 is the all-zero reference. Coefficients in Table 8 are main effects with no interaction terms in the model.

Model specifications. With F, P and C the three binary feature indicators and i indexing interviews:

1. M1: moves_i = b0 + b1 F_i + b2 P_i + b3 C_i + e_i; ordinary least squares, complete cases, n=92.
2. M2: as M1 on all randomised starts, n=100, retaining observed scores below the inclusion threshold.
3. M3: mixed-effects logistic regression of answer-level follow-up, with answers nested in interviews and a random interview intercept.
4. M4: as M1 with interview duration in minutes as the outcome.
5. M5: mixed-effects logistic regression of turn-level multi-question status, with a random interview intercept.
6. M6: as M1 with respondent words as the outcome.
7. M7: as M1 with moves landed in the first ten minutes as the outcome.
8. M8: a binomial model for moves landed out of twelve, with beta-binomial specification if the binomial variance assumption is rejected.

Confidence-interval and standard-error method. Intervals for the linear models are Wald intervals on the model standard errors. Answer-level and turn-level intervals account for clustering within interview through the random intercept in M3 and M5, which is the mechanism by which dependence between answers from the same interview is handled. No separate cluster-robust variance estimator was applied in addition, and none is reported. The intraclass correlation interval is the absolute-agreement single-rater form [35]. Missing-data treatment. No imputation was performed. M1 is a complete-case analysis of the 92 analysable interviews. M2 retains all 100 randomised starts with their observed scores, so interviews end-ing before the inclusion threshold contribute their actual measured fidelity rather than being dropped or imputed. Statistical software and package versions. The mixed-effects models were specified in the form imple-mented by lme4 [37] and the bounded-outcome model in the generalised linear model framework [38]. The specific software build, language release and package releases used to produce the reported fits were not recorded in the analysis file and are therefore not stated. Repository. Materials are held in a controlled-access archive maintained by the authors and available on reasonable request. The archive contains the instrument instructions for every version with its configuration and creation timestamp, the four occupational-topic modules, the twelve-move rubric, the scoring instructions, the interview data as per-interview and per-move records with the supporting turn indices, the full transcripts, the recruitment records, the analysis file underlying the factorial estimates, and the scripts that pull, score and assemble the results. No public repository has been minted, and no digital object identifier has been issued. Analysis plan. The analysis plan was fixed before the factorial data were analysed but was not lodged in a public registry, so no independent time stamp exists and it is described throughout as a prospective analysis plan rather than as a preregistration. The plan is held in the archive described above. The instrument versions and the scorer configuration it refers to are timestamped records carrying creation timestamps, the last of which is dated 5 August 2026, which establishes the internal ordering of the freeze relative to the fielding even though it does not substitute for an independent registration.

Outstanding analysis requirements. Two analyses specified during review are not reported because the underlying data or coding are not available, and neither is claimed:

1. **Validation of the interviewing-craft classifications.** Human coding validates the twelve-move fidelity scorer only. The separate automatic classifications for substantive answers, pursued answers, asserted constraints, challenged constraints, multi-question turns, and candidate-answer questions require validation. The specified procedure is to select 24 factorial-evaluation transcripts, three per cell; freeze the six category definitions; have two coders independently code them while blinded to condition and automatic labels; and report decision counts, precision, recall, exact agreement, and Cohen's kappa with confidence intervals. If too few asserted constraints are observed, that measure requires a larger validation sample. Until then, craft measures remain provisional.
2. **Main effects by move type.** Appendix A classifies moves as episodic, evaluative, or design. Estimating main effects within each type would test whether narrated sequence is recovered more readily than evaluative and design moves. The available

factorial results do not contain the required per-interview, per-move indicators, so the analysis is not reported and would be exploratory rather than confirmatory.

Appendix F. The Opening Question as a Sampling Variable

The baseline instrument's opening question named the respondent's occupational function, and reported function fit was 76%. When the opening was changed to ask what the respondent's role was without naming a function, four of five respondents recruited into a marketing panel proved to work in other fields. The original fit rate was therefore partly an artefact of the opening question. The pattern recurred incidentally. Version 4 removed the role question, and the instrument resumed naming the panel's function unprompted within three turns, reproducing the artefact. The question was restored in the final instrument reported in the main text, and the artefact did not recur. For studies using conversational instruments with recruited panels, this makes the opening question a sampling variable that should be reported as one. It also bears on the argument of section 1.1, since an instrument that asks what a person does rather than assuming it from a recruitment category records the work rather than the assignment [1]. These observations rest on small numbers within the development stages and are reported as descriptive.

Appendix G. Development-Stage Detail

These tables support the descriptive development account. Stages are confounded with fielding order and carry between one and thirty-four interviews. Attribution rests on the factorial evaluation in Tables 6-8.

Table G1. Change in each measure across successive development stages. Descriptive. Version 4 is omitted because only one interview met the inclusion rule.

Measure	Version 1 to 2	Version 2 to 3	Version 3 to 5	Net
Moves landed of twelve	+1.1	-0.2	+1.5	+2.5
Substantive answers followed up	+26.2 pts	-3.8 pts	+13.2 pts	+35.5 pts
Respondent words	+752	-377	+1,257	+1,633
Respondent turns	+24.6	-0.9	+9.9	+33.6

Table G2. Question discipline and method fidelity across development stages. Descriptive. The first three rows rest on unvalidated classifiers.

Measure	Version 1	Version 2	Version 3	Version 5
Turns containing more than one question	1%	16%	3%	24%
Questions supplying candidate answers	10%	7%	1%	12%
Interviewer words per turn	28.0	18.5	16.1	32.1
Moves landed of twelve	6.9	8.0	7.8	9.4
Substantive answers followed up	43%	69%	65%	78%
Constraint challenge rate	0%	50%	0%	18%

Table G3. Landed rate by move across the development stages. Descriptive. Version 4 is omitted because only one interview met the inclusion rule.

Move	Type	Version 1	Version 2	Version 3	Version 5
1 Purpose and change-prompt	Evaluative	41%	40%	33%	100%
2 Enacted step-by-step	Episodic	97%	100%	100%	100%
3 Documented against enacted	Evaluative	35%	20%	17%	64%
4 Who they go to	Episodic	50%	100%	83%	73%
5 Gatekeeper or bottleneck	Episodic	100%	100%	100%	100%
6 Interdependencies	Episodic	91%	100%	100%	100%
7 Friction at interdependencies	Episodic	100%	100%	100%	100%
8 Under-used capability	Evaluative	24%	40%	33%	45%
9 Stated against lived values	Evaluative	12%	0%	33%	18%
10 Generality to episode	Episodic	91%	40%	67%	100%
11 Automate, human or drop	Design	0%	60%	50%	36%
12 What would improve	Design	47%	100%	67%	100%

References

1. Krackhardt, D. and Hanson, J. R. Informal networks: the company behind the chart. Harvard Business Review, 71(4), 104-111, 1993.
2. Feldman, M. S. and Pentland, B. T. Reconceptualizing organizational routines as a source of flexibility and change. Administrative Science Quarterly, 48(1), 94-118, 2003.

3. Pentland, B. T. and Feldman, M. S. Organizational routines as a unit of analysis. Industrial and Corporate Change, 14(5), 793-815, 2005.
4. Argyris, C. and Schön, D. A. Organizational Learning: A Theory of Action Perspective. Addison-Wesley, 1978.
5. Suchman, L. A. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press, 1987.
6. Brown, J. S. and Duguid, P. Organizational learning and communities-of-practice: toward a unified view of working, learning, and innovation. Organization Science, 2(1), 40-57, 1991.
7. Thompson, J. D. Organizations in Action: Social Science Bases of Administrative Theory. McGraw-Hill, 1967.
8. Faraj, S. and Xiao, Y. Coordination in fast-response organizations. Management Science, 52(8), 1155-1169, 2006.
9. Okhuysen, G. A. and Bechky, B. A. Coordination in organizations: an integrative perspective. Academy of Management Annals, 3(1), 463-502, 2009.
10. Galbraith, J. R. Organization design: an information processing view. Interfaces, 4(3), 28-36, 1974.
11. Burton, R. M. and Obel, B. The science of organizational design: fit between structure and coordination. Journal of Organization Design, 7(1), 5, 2018.
12. Weiss, R. S. Learning from Strangers: The Art and Method of Qualitative Interview Studies. Free Press, 1994.
13. Small, M. L. "How many cases do I need?" On science and the logic of case selection in field-based research. Ethnography, 10(1), 5-38, 2009.
14. Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M. M., Würschinger, Q. and Kreuter, F. AI conversational interviewing: transforming surveys with LLMs as adaptive interviewers. arXiv:2410.01824, 2024. Accepted at LaTeCH-CLfL 2025. https://arxiv.org/abs/2410.01824
15. Wuttke, A., Lang, M. M., Klamm, C., Würschinger, Q. and Kreuter, F. AI conversational interviewing: scaling up semi-structured and in-depth interviews. arXiv:2606.20064, 2026. https://arxiv.org/abs/2606.20064
16. Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302, 1955.
17. Messick, S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741-749, 1995.
18. Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J. and Balain, S. A conceptual framework for implementation fidelity. Implementation Science, 2, 40, 2007.
19. Dane, A. V. and Schneider, B. H. Program integrity in primary and early secondary prevention: are implementation effects out of control? Clinical Psychology Review, 18(1), 23-45, 1998.
20. Hasson, H. Systematic evaluation of implementation fidelity of complex interventions in health and social care. Implementation Science, 5, 67, 2010.
21. Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D. and Houck, J. M. The Motivational Interviewing Treatment Integrity code (MITI 4): rationale, preliminary reliability and validity. Journal of Substance Abuse Treatment, 65, 36-42, 2016.
22. Stanford, N. Organization Design: The Practitioner's Guide. Routledge, 2018.
23. Nadler, D. A. and Tushman, M. L. A model for diagnosing organizational behavior. Organizational Dynamics, 9(2), 35-51, 1980.
24. Nadler, D. A. and Tushman, M. L. The organization of the future: strategic imperatives and core competencies for the 21st century. Organizational Dynamics, 28(1), 45-60, 1999.
25. Puranam, P., Alexy, O. and Reitzig, M. What's "new" about new forms of organizing? Academy of Management Review, 39(2), 162-180, 2014.
26. Schober, M. F. and Conrad, F. G. Does conversational interviewing reduce survey measurement error? Public Opinion Quarterly, 61(4), 576-602, 1997.
27. Conrad, F. G. and Schober, M. F. Clarifying question meaning in a household telephone survey. Public Opinion Quarterly, 64(1), 1-28, 2000.
28. West, B. T. and Blom, A. G. Explaining interviewer effects: a research synthesis. Journal of Survey Statistics and Methodology, 5(2), 175-211, 2017.
29. Schaeffer, N. C., Dykema, J. and Maynard, D. W. Interviewers and interviewing. In P. V. Marsden and J. D. Wright (eds.), Handbook of Survey Research, 2nd ed., 437-470. Emerald, 2010.
30. Kvale, S. Doing Interviews. Sage, 2007.
31. Schein, E. H. Humble Inquiry: The Gentle Art of Asking Instead of Telling. Berrett-Koehler, 2013.
32. Jha, A., Shivaprakash, P., Shukla, L., Mukherjee, A., Chand, P. and Murthy, P. Benchmarking motivational interviewing competence of large language models. arXiv:2603.03846, 2026. https://arxiv.org/abs/2603.03846
33. Feinstein, A. R. and Cicchetti, D. V. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543-549, 1990.
34. Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1), 29-48, 2008.
35. McGraw, K. O. and Wong, S. P. Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30-46, 1996.
36. Shrout, P. E. and Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428, 1979.
37. Bates, D., Mächler, M., Bolker, B. and Walker, S. Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48, 2015.
38. McCullagh, P. and Nelder, J. A. Generalized Linear Models, 2nd ed. Chapman and Hall, 1989.
39. Schulz, K. F., Altman, D. G. and Moher, D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ, 340, c332, 2010.
40. OpenAI. Realtime models prompting guide, 2026.
41. Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174, 1977.
42. Shaw, R. B. and Nadler, D. A. Capacity to act. Human Resource Planning, 14(4), 289-300, 1991.
43. Puranam, P., Raveendran, M. and Knudsen, T. Organization design: the epistemic interdependence perspective. Academy of Management Review, 37(3), 419-440, 2012.
44. Raveendran, M., Silvestri, L. and Gulati, R. The role of interdependence in the micro-foundations of organization design: task, goal, and knowledge interdependence. Academy of Management Annals, 14(2), 828-868, 2020.
45. Pentland, B. T. and Feldman, M. S. Designing routines: on the folly of designing artifacts, while hoping for patterns of action. Information and Organization, 18(4), 235-250, 2008.
46. Grimmer, J., Roberts, M. E. and Stewart, B. M. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, 2022.
47. Orlikowski, W. J. Improvising organizational transformation over time: a situated change perspective. Information Systems Research, 7(1), 63-92, 1996.