

How can digital and AI-enabled behaviour change interventions be evaluated while they are changing? HUMMBER: identifying what changed, for whom, why and with what effect

Maria Alejandra Pinero de Plaza

alejandra.pinerodeplaza@flinders.edu.au

Flinders University <https://orcid.org/0000-0001-5421-9604>

Ian Halkett

University of Adelaide

Damith Ranasinghe

University of Adelaide

Carlos Javier Barrera Causil

Instituto Tecnológico Metropolitano

Joanne Dollard

University of Adelaide

Renuka Visvanathan

University of Adelaide

Research Article

Keywords: artificial intelligence, behaviour change, digital health, human-centred evaluation, implementation science, longitudinal evaluation, micro-randomisation, accountability, simulation

Posted Date: October 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11369973/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Digital behaviour change interventions (DBCIs), including AI-enabled DBCIs, can change during use and scale-up as users and implementation circumstances also change. Within a focused exploratory evidence map, we did not identify a reusable method connecting longitudinal change, affected people, circumstances, explanations, component comparisons, evidence limits and human decision authority.

Methods

We aimed to specify HUMMBER (Human-centred, User-informed, Multimethod Micro-randomised Behaviour Evaluation and Response) as a reusable method for ongoing evaluation and to examine selected proposed functions in a hypothetical synthetic scale-up demonstration. HUMMBER is organised around six procedural questions and a version-controlled Application Specification. Optional software supports recording, allocation and analysis without interpreting evidence or authorising action. In a wholly synthetic 26-week scale-up demonstration using a 600-person mixed cohort and five 1,200-person scenarios, we examined: (1) What changed? through longitudinal trajectories with similar final outcomes; (2) For whom? through support opportunity, burden and missingness across groups; (3) Under what circumstances? through assessment timing, information and burden; (4) What might explain the change? by updating possible explanations as evidence accumulated; (5) Did a component make a difference? through prespecified valid randomised comparisons of next-week engagement; and (6) What can evaluators conclude? through evidence-status rules, sensitivity analysis and Monte Carlo checks.

Results

(1) Known synthetic change patterns remained descriptively distinguishable despite similar final maintenance (68.5%-74.4%). (2) Limited digital access reduced opportunities to receive the eligible support options (34.3% vs 72.7%) and increased missing data (8.4% vs 4.2%), without a programmed difference in support effect. (3) Participant-selected timing produced the greatest useful information per active participant minute in synthetic units. (4) Accumulating evidence assigned the highest probability to the programmed explanation for most synthetic participants, even from misleading starting assumptions. (5) Technical assistance (+ 23.2 percentage points; 95% CI 15.5 to 30.8), goal problem-solving (+ 13.5; 5.9 to 21.1) and flexible timing (+ 14.6; 6.4 to 22.7) improved next-week engagement. (6) The prespecified evidence-status rules limited or withheld conclusions when evidence was insufficient, comparisons unavailable, effects changed over time or missingness threatened interpretation.

Conclusions

HUMMBER, specified with transdisciplinary author-team input including a consumer perspective and designed for co-design in live applications, is a proposed human-governed framework for traceable evaluation of DBCIs that change during use and scale-up. In this wholly synthetic proof of concept, the synthetic implementation demonstrated computational feasibility and bounded inference under the documented demonstration conditions, not real-world effectiveness. Prospective testing should include participatory co-design, adaptive and agentic AI-enabled interventions and live governance processes.

Background

Why ongoing, accountable evaluation is needed

Digital behaviour change interventions (DBCIs) use websites, apps, wearable devices and other technologies to support changes in health behaviour. They can extend support beyond appointments and reach people at scale [1], an increasingly important capability as health services face constrained capacity and a projected global shortage of 11 million health workers by 2030 [2]. Some DBCIs also use artificial intelligence (AI) to predict needs, personalise content or determine when and how support should be offered [3]. Evaluation therefore needs to consider not only whether behavioural or health outcomes are achieved, but also how the intervention operates and how people experience and respond to it during use.

Ongoing use and scale-up create a moving evaluation target. Similar patterns of continued use, disengagement or behavioural change can arise because access failed, a goal difficulty increased, support lost value, competing responsibilities took priority or repeated assessment became burdensome. Software updates, revised content and recommendation rules can also change what the DBCI delivers. A final outcome can therefore hide important differences in what happened, for whom and under which circumstances.

We undertook an exploratory evidence map to determine whether an integrated method for ongoing evaluation of DBCIs, including AI-enabled DBCIs, already existed and, if not, which approaches could inform its development. The map identified complementary capabilities in process evaluation [4], just-in-time adaptive interventions [5,6], micro-randomised trials (MRTs) [7,8], longitudinal realist evaluation [9], and human-centred approaches including PROLIFERATE and PROLIFERATE_AI [10,11]. Additional file 1 reports the review question, exploratory approach, narrative synthesis and evidence tables.

Whilst the mapped approaches address different parts of the evaluation problem, we did not identify a reusable and comprehensive method that connected longitudinal change, human experience, circumstances, possible explanations, valid component comparisons, evidence limits and human decision authority over time. The gap is increasingly important for adaptive and agentic AI-enabled systems, for which current guidance emphasises fairness, traceability, explainability, validation, lifecycle monitoring, ethical oversight and appropriate human responsibility [12–15].

HUMMBER's contribution is the integration of longitudinal observation, contextual inquiry, explanation updating, component comparison and bounded evidence reporting within a governed evaluation architecture. The architecture specifies what evidence can support, what it cannot support, and who retains authority to act.

The HUMMBER proposition, study aim and research questions

To address the identified evaluation gap and together with feedback from co-authors, including a consumer perspective, we propose HUMMBER (Human-centred, User-informed, Multimethod Micro-randomised Behaviour Evaluation and Response) as a human-governed method for transparent, ongoing evaluation of DBCIs, including AI-enabled DBCIs. HUMMBER connects qualitative and quantitative evaluation through six procedural questions, with optional software support for standardised and reproducible quantitative work alongside qualitative inquiry. HUMMBER is an evaluation method, not an intervention, AI model or autonomous decision-maker; evaluators interpret the evidence and designated human roles retain authority over subsequent action [4,7–11].

We further propose that HUMMBER organises ongoing evaluation around six procedural questions that guide the methods, analyses, responses and conclusions:

1. What changed?
2. For whom?
3. Under what circumstances?
4. What might explain the change?
5. Did a particular component make a difference?
6. What can evaluators conclude, and what may authorised decision-makers consider next?

The questions form a sequence rather than six independent tests: observed change is linked to the people affected and their circumstances; possible explanations and mechanisms are examined using appropriate evidence; causal component effects are estimated only when a valid comparison exists; and the final evidence status records what can and cannot reasonably be concluded.

The study aimed to specify HUMMBER as a reusable method for ongoing evaluation and to examine selected proposed functions through a hypothetical synthetic scale-up demonstration. Four research questions addressed that aim:

RQ1. Can HUMMBER distinguish different longitudinal trajectories and opportunities to receive support when final maintenance outcomes are similar (maintenance: meaningful engagement during at least two of the final four weeks of the evaluation period)?

RQ2. Can HUMMBER compare ways of obtaining information about the circumstances surrounding behavioural change while considering both the information obtained and the burden placed on users?

RQ3. Can HUMMBER trace system actions, user experiences and subsequent outcomes while distinguishing possible explanations from causal effects and identifying when the available evidence cannot support a conclusion?

RQ4. What must consumers, practitioners, service decision-makers, technical staff and evaluators co-design and record to apply HUMMBER to another DBCI, including an AI-enabled DBCI, while preserving traceability, safeguards and human decision authority?

Aims

This methodological study had one overarching aim: to specify HUMMBER as a reusable method for ongoing evaluation and examine selected proposed functions under controlled synthetic conditions. The work proceeded in two linked stages.

- First, findings from the exploratory evidence map, together with feedback from co-authors, including a consumer perspective, were translated into requirements for the six procedural questions, safeguards, evidence limits, traceability and decision responsibilities.
- Second, selected methodological functions were implemented in an optional software layer and examined in a computational demonstration. The demonstration followed ADEMP: Aims, Data-generating mechanisms, Estimands, Methods and Performance measures [16].

Methods

Method development and demonstration design

The methodological study followed the two stages specified above. First, the exploratory evidence map and co-author feedback, including a consumer perspective, were translated into requirements for the six procedural questions, safeguards, evidence limits, traceability and decision responsibilities. Second, selected functions were implemented in an optional software layer and examined in a computational demonstration following ADEMP: Aims, Data-generating mechanisms, Estimands, Methods and Performance measures [16].

The specification was informed by transdisciplinary author-team consultation, including a consumer perspective. We do not claim that the present study was a prospective formal co-design study; formal co-design with affected users remains part of live application.

HUMMBER structure and operational terms

HUMMBER has three connected parts: the evaluation method, an application-specific HUMMBER Application Specification, and an optional software layer.

Table 1 defines the operational terms used throughout the manuscript; the same terms are used consistently thereafter.

Table 1. Key operational terms used in HUMMBER

Term	Plain-language definition
Assessment burden	The active participant time and effort required to respond to evaluation requests, such as questionnaires or prompts.
Bayesian updating	Where prespecified, revision of documented probabilities as new evidence accumulates. Starting probabilities represent assumptions or elicited judgements rather than findings. The updating rule used in the demonstration is stated explicitly in Methods and Additional file 2, Table S10.
Causal effect	A difference that can reasonably be attributed to one option rather than another when the design provides a valid comparison and its required assumptions are supported.
Change signal	A prespecified rule that flags a potentially meaningful change over time and prompts closer evaluation. It is not itself an explanation, diagnosis or causal finding.
Component	A distinct feature or support option delivered by, through or alongside the host system.
Context or circumstances	Personal, technological, service or wider conditions surrounding a system action, user response or outcome.
Decision point	A moment when eligibility requirements are met and at least two acceptable, available and permitted options could genuinely be provided.
Event	A dated or time-stamped record that something relevant occurred, such as a login or failed login, recommendation, refusal, error, assessment, support action or software change.
Evidence provenance	A record of where evidence came from, when it was generated and which system version or evaluation conditions applied.
Evidence status	An evaluator-recorded statement of what the current evidence supports and where uncertainty remains.
Explainability (HUMMBER)	Making the relationship between a recorded system action and what subsequently happened visible and reviewable. It does not mean access to an AI model's internal reasoning.
Expert knowledge elicitation (EKE)	A structured way of recording expert or lived-experience knowledge and uncertainty as explicit, reviewable judgements [17].
Host system	The DBCI being evaluated, including an AI-enabled DBCI.
HUMMBER Application Specification	A dated and version-controlled specification defining how HUMMBER will be applied to one host system and setting.
In-situ evaluation	Evaluation that continues while the host system is being used and while the system, its users or implementation circumstances may change.
Micro-randomisation	Use of chance to allocate between eligible options at repeated decision points.
Possible explanation	A documented hypothesis about why an observed pattern may have occurred.

Term	Plain-language definition
Possible mechanism	A specific, testable explanation of how a system action or component may contribute to an outcome under particular circumstances. An observed association does not establish a mechanism.
Realist analysis	Examination of how circumstances and possible mechanisms may combine to produce different outcomes [18].
Software layer	Optional research software (Python modules and Google Colab-compatible notebooks) that supports selected HUMMBER recording, allocation and analysis functions. The software supports the method but does not interpret evidence, authorise action or control the host system.

Part 1: the HUMMBER evaluation method

Before ongoing evaluation begins, consumers, practitioners, service decision-makers, technical staff and evaluators co-design what will be evaluated: the host-system boundary; relevant people and experiences; components, events and outcomes; possible explanations; acceptable measurements and comparisons; accessibility, equity, burden and safeguards; and the roles responsible for interpreting evidence and authorising subsequent action. The agreed requirements are recorded in the Application Specification described in Part 2.

The ongoing method then connects Observe, Detect, Ask, Explain, optional Experiment and Govern. Observe retains actions, experiences and outcomes over time; Detect applies prespecified change signals; Ask examines circumstances and requests additional information only when needed; Explain evaluates documented possible explanations; Experiment can compare eligible components when an approved causal question exists; and Govern records the evidence-supported conclusion and its limits.

Update maintains the evolving evidence state as new information accumulates, retains contradictory evidence and, where prespecified, revises probabilities assigned to possible explanations. Micro-randomisation is optional and is used only when at least two options are acceptable, available, permitted and approved. Refused options leave the comparison, and realised assignment probabilities are recorded.

Causal conclusions are restricted to the people, decision points, options, outcome period and system version represented by a valid comparison. A component effect does not establish the mechanism through which the effect occurred. The comparison matched to the evaluation question is prespecified as primary; other contrasts are diagnostic, and a diagnostic interval excluding zero is retained as a signal requiring confirmation rather than treated as a supported effect.

Table 2 maps the HUMMBER functions to the six procedural questions. Figure 1 shows how the functions connect, including the optional experimental branch, evidence updating, change control, safeguards and human decision authority.

Table 2. HUMMBER setup, evaluation functions and procedural questions

Function	Role in HUMMBER
Co-design and specify	Define what will be evaluated, for whom, under what conditions, with which safeguards, evidence rules and decision responsibilities.
Observe	Q1. What changed? Record relevant actions, states and outcomes over time.
Detect	Q2. For whom? Identify the people, groups or system trajectories associated with a detected change.
Ask	Q3. Under what circumstances? Determine what is already known and obtain additional contextual information only when needed.
Explain	Q4. What might explain the change? Examine longitudinal, contextual, qualitative and other evidence against documented possible explanations.
Experiment where justified	Q5. Did a particular component make a difference? Compare eligible components only when an appropriate and approved comparison is available.
Update	Combine accumulating evidence, retain contradictions and revise the evidence status and, where prespecified, probabilities assigned to possible explanations.
Govern	Q6. What can evaluators conclude, and what may authorised decision-makers consider next? Record what the evidence supports, its limits and uncertainty, while keeping subsequent decisions with the designated human role or governance group.

Part 2: the HUMMBER Application Specification

The Application Specification translates the general method into a dated, version-controlled plan for one host system and setting. Its 24 fields cover the evaluation purpose and boundary; relevant people, components, events and outcomes; evaluation questions, possible explanations and permitted comparisons; measurement and support options; consent, accessibility, equity and burden safeguards; and responsibilities for interpretation, decision-making, audit, transfer and re-evaluation. The completed illustrative specification is provided in Additional file 2, Table S5, and the reusable co-design workbook in Additional file 3, Table S22.

Version control links each body of evidence to the system, population, implementation conditions, permitted comparisons and decision responsibilities under which it was generated. The specification also names who interprets evidence and who is authorised to approve subsequent action.

Any live application of HUMMBER would require local co-design and applicable ethics, safety, service-governance and regulatory review.

Part 3: optional HUMMBER software layer

The optional software layer consists of reusable Python modules and executable Google Colab-compatible notebooks. In the present study it generated and organised synthetic records, applied prespecified change-signal and allocation rules, updated evidence where specified, and reproduced the planned analyses.

The software operates only within the approved Application Specification. It does not decide which options are acceptable, interpret what evidence means, authorise action, modify the host system or become a second AI decision-maker.

Common event record and integration modes

HUMMBER links information over time through a common event record. Each record identifies what occurred and when, relevant user and system circumstances, available or delivered options, the subsequent outcome and evidence provenance; where applicable it also retains permission, acceptance or refusal, system version, assessment burden and the realised randomisation probability. Additional file 2, Table S6 specifies the common record.

Before relevant outcomes are examined, the co-design team also selects how HUMMBER is allowed to interact with the host system. Table 3 summarises the three integration modes and the strongest inference each mode can support.

Table 3. HUMMBER integration modes and inferential boundaries

Integration mode	What the evaluation may change	Strongest conclusion potentially supported
Controlled integration	Approved assessment approaches and, when explicitly permitted, eligible support or host-system components may be randomised.	A causal effect may be estimated for the randomised options within the people, decision points, outcome period and system version represented by the comparison, when required assumptions are defensible.
Evaluation-only integration	Host-system delivery remains unchanged. Only when or how evaluation information is collected may be randomised.	Causal effects may be estimated for randomised assessment strategies. Findings about host-system actions remain descriptive or associational unless another valid causal design exists.
Observational or audit integration	HUMMBER records relevant information but introduces no randomisation and does not alter host-system delivery.	Longitudinal patterns and associations may be described, but observation alone cannot establish that a host-system action caused an outcome.

The approved integration mode, not the amount of data collected, determines whether a causal claim is possible. Material changes to the host system, permitted comparisons or decision responsibilities trigger review and, where necessary, reapproval of the Application Specification.

Computational demonstration

Demonstration case and scale-up rationale

All participant-level records in the demonstration were synthetic and generated under conditions specified before analysis. Known data-generating processes and selected effects therefore provided a reference for examining whether HUMMBER behaved appropriately when an explanation or component effect was present, absent, uncertain or not identifiable. Detailed generation rules are reported below under Synthetic data-generating mechanism and in Additional file 2, Tables S9 and S10.

The six-month AVOID hybrid type 1 trial [19] randomised 60 adults aged 60 years or older who were not frail. Compared with usual care, AVOID produced a greater reduction in frailty index (-0.044 , 95% confidence interval [CI] -0.076 to -0.012) and improved quality of life ($+0.032$, 95% CI $+0.006$ to $+0.058$); adoption was 90.3% and maintenance was 71.0% [19]. Source provenance, including the published quality-of-life CI discrepancy, is documented in Additional file 2, Table S7. This trial was selected to examine selected HUMMBER functions under a hypothetical scale-up because the trial provided a realistic context in which a final scale-up outcome could conceal different preceding trajectories.

Published trial findings served only as aggregate context and broad plausibility anchors. Every participant-level trajectory, circumstance, comparison, support effect and missing outcome in the demonstration was computer-generated; the study cannot identify participant-level mechanisms, subgroup differences or component effects in the trial.

Published implementation findings informed four synthetic processes: technical barriers, goal difficulty, declining perceived value and competing demands [19]. Assessment burden was added as a fifth methodological stress test because the original study reduced parts of its assessment to limit burden.

Figure 2 shows the evidence boundary: published trial findings inform the illustrative specification and prespecified simulation rules, which generate the wholly synthetic event stream evaluated by HUMMBER.

Synthetic data-generating mechanism

A reproducible Python 3.12.12 simulation with master seed 20260824 generated 26 weekly records per synthetic participant. Prespecified models generated participant characteristics, engagement, logins, help-seeking, incomplete tasks, host-system availability, competing demands, perceived value, assessments, support and missing outcomes. Detailed data-generating rules and parameter provenance are reported in Additional file 2, Tables S9 and S10.

The main mixed cohort contained 600 synthetic participants. Five additional scenarios contained 1,200 participants each, with one programmed process per scenario. Scenario-specific intercepts were

calibrated so that final maintenance remained close to the published 71.0% AVOID maintenance value while preceding trajectory markers were free to differ.

Only acceptable, available and permitted options could enter a randomised comparison, and the full eligible option set and realised assignment probability were retained at each decision point. For every detected change, the simulation also retained probabilities across the five possible explanations and updated them as new events or assessments accumulated. A separate stress test made proximal outcomes more likely to be missing after disengagement.

No expert knowledge elicitation (EKE) was conducted. Equal starting probabilities of 0.20 and a deliberately misleading starting distribution assigning 0.58 to technical barriers were used only to test whether accumulating evidence could revise neutral or incorrect starting assumptions.

Additional file 2, Tables S8-S10 provide the complete demonstration specification and assumptions. Additional file 4 contains two executable Google Colab notebooks with the analysis and data-generating code embedded for peer-review reproducibility, together with the wholly synthetic datasets and machine-readable data dictionary used by the notebooks.

Analysis aligned with the six HUMMBER questions

The six HUMMBER questions determined the information and comparisons required before statistical techniques were selected. Estimands, analysis rules and simulation parameters were prespecified, and the Results follow the same six-question order. Additional file 2 provides the detailed computational specification and Additional file 4 the executable analyses.

What changed?

Maintenance was defined as meaningful engagement during at least two of the final four simulated weeks. Weekly event records retained the preceding trajectory. A change signal was triggered by disengagement in two consecutive weeks or by current disengagement accompanied by help-seeking or an incomplete task. The signal threshold was deliberately set to produce repeated evaluation cycles; in a live application, threshold and alert budget are co-design parameters (Additional file 2, Tables S5 and S10).

The analysis compared final maintenance with the preceding weekly trajectories in the mixed cohort and five single-process scenarios.

For whom?

Age, sex, digital access and accessibility need were retained to examine who experienced detected changes and who had the opportunity to receive support or contribute evidence. The main equity stress

test deliberately made technical assistance less available to participants with limited digital access while programming the same logit-scale support-effect parameter for both digital-access groups and no digital-access interaction.

Support availability, actual exposure, assessment burden and missing outcomes were analysed separately from effect. Because a common logit-scale effect can yield different risk differences when baseline probabilities differ, exposure differences were not interpreted as effect modification. The demonstration-specific equity requirements are recorded in Additional file 2, Table S5.

The technical-assistance effect was also estimated within six age-by-sex subgroups and then partially pooled using a random-effects normal model with between-subgroup variance estimated by restricted maximum likelihood (REML) [20]. The exercise examined shrinkage of noisy subgroup estimates, not real sociodemographic heterogeneity.

Intervals from the hierarchical illustration condition on the estimated between-subgroup variance and therefore understate total uncertainty. The data-generating mechanism contained no sociodemographic interaction, so the pooling analysis was not used to infer homogeneity.

Under what circumstances?

After an eligible detected change, the simulation compared passive observation with immediate, delayed and participant-selected contextual assessment. Active strategies were randomised at eligible assessment opportunities. The comparison retained response, active assessment time, useful explanatory information and missing responses.

Information gain was a simulation-specific quantity, not a validated scale. Information per active minute was calculated for the three active strategies as total synthetic information divided by total active assessment minutes; unanswered prompts still contributed 0.25 minutes of burden. Passive observation required no active participant time and was evaluated separately (Additional file 2, Table S10).

What might explain the change?

For each detected change, HUMMBER retained probabilities across five documented possible explanations: technical barriers, goal difficulty, declining perceived value, competing demands and assessment burden. The analysis examined whether those probabilities moved toward the process programmed into the synthetic data as evidence accumulated.

Weekly evidence was combined using a tempered naive-Bayes update. Process-specific event likelihoods from Additional file 2, Table S10 were raised to exponent 0.10 before updating so that correlated weekly indicators did not dominate the accumulated evidence.

Answered or passively inferred assessments contributed source-specific likelihood ratios derived from their programmed accuracies: immediate self-report 14.18, delayed self-report 10.29, participant-selected assessment 21.00 and passive inference 4.00. Contradictory evidence was retained and no explanation was forced to probability zero.

The analysis compared equal starting probabilities with the deliberately misleading start and tracked both the final highest-probability explanation and the probability assigned to the programmed process over time. Additional file 2, Table S10 gives the full update rule.

Exploratory k-means clustering (k=5, 10 initialisations) on six standardised features from a random eight-week window was used only as a negative check on whether sparse data could create apparently meaningful participant groups. Performance was described using mapped accuracy, adjusted Rand index and silhouette; clusters were not treated as phenotypes or mechanisms.

Did a particular component make a difference?

Micro-randomisation was used only for a causal component question when a valid comparison existed. Synthetic participants who had not refused additional support could be allocated among eligible options: usual support, technical assistance, goal problem-solving, interactive voice support or flexible timing. Eligible options were assigned with equal probability, so the realised assignment probability was 1 divided by the number of eligible options at that decision point.

The proximal estimand was the difference in next-week engagement between an eligible component and usual support. Estimation was restricted to reconstructible decision points at which both options were acceptable, available and permitted and the next-week outcome was observed. The common event record retained per-option availability and the realised assignment probability (Additional file 2, Table S6).

Component effects were estimated as normalised inverse-probability-weighted (IPW; Hájek) risk differences using the logged assignment probabilities. We used the Hájek estimator rather than weighted-and-centred least squares (WCLS) [21]. Participant-clustered sandwich standard errors were calculated from the two weighted-mean estimating equations with finite-cluster correction $G/(G - 1)$; full equations are provided in Additional file 2, Table S10.

Risk differences are reported in percentage points; for example, +0.10 corresponds to about 10 additional engaged next-week outcomes per 100 comparable decision points. Assignment probabilities were known from the randomisation record and were not estimated from observed data.

Within each scenario, the component matched to the evaluation question was prespecified as the primary contrast. A primary 95% CI excluding zero was classified as supported within synthetic scope; a valid contrast with a CI including zero was classified as insufficient evidence; and a diagnostic contrast

with a CI excluding zero was retained as a signal requiring confirmation. No causal estimate was produced when the required valid comparison did not exist.

Estimator operating characteristics were assessed separately under a true null using 300 consecutive, unselected replications, with bias, empirical standard deviation, mean model-based standard error, 95% CI coverage and false-positive rate prespecified as performance measures (Additional file 2, Table S19).

What can evaluators conclude, and what may authorised decision-makers consider next?

Evidence-status rules recorded what the available evidence supported without converting uncertainty into a decision. Prespecified statuses included supported within synthetic scope, insufficient evidence, not identifiable, contradicted/unstable and diagnostic signal requiring confirmation.

Four stress tests examined a true null effect, an unavailable target option, an effect that reversed direction over time and informative missingness after disengagement. The first three tested whether HUMMBER withheld or qualified causal conclusions when the design or evidence did not support a stable effect.

For informative missingness, the observed-outcome analysis was followed by a deterministic two-dimensional tipping-point sensitivity analysis [22]. Missing outcomes in the component and usual-support arms were shifted independently from -0.30 to +0.30 on the log-odds scale across a 7 x 7 grid, and the effect and participant-clustered 95% CI were re-estimated in each cell. A tipping point was defined as a cell in which the CI included zero. The -0.30 to +0.30 range was prespecified to represent moderate departures from the observed-data missingness mechanism; to locate the departure at which the conclusion would change, the grid was also extended post hoc beyond this range (Results). The resulting intervals are conditional on each fixed missingness assumption and do not include uncertainty about the missing-data model itself.

Governance output was limited to the evidence status and the type of action a hypothetical authorised decision-maker could consider. HUMMBER did not authorise any action concerning AVOID, its implementation or any participant.

Quality assurance and reproducibility

Computational quality assurance used a fixed master seed, explicit parameter provenance, programmatic output generation and 29 automated integrity checks. The checks covered record uniqueness and week ranges; eligibility of assessments and support; validity of assignment probabilities; non-negative cumulative burden; preservation of primary versus diagnostic contrasts; and absence of real AVOID participant identifiers.

Selected stress-test cohorts were used only to demonstrate prespecified difficult conditions and were not used to assess estimator performance. Reproducibility was evaluated by rerunning the complete workflow and comparing regenerated results with reference values using stated tolerances. Twenty-nine automated integrity checks were run during package preparation. Additional file 4 supplies the wholly synthetic data and two self-contained executable notebooks so editors and reviewers can inspect and reproduce the computational analyses.

Results

Results are reported in the same six-question order used for the analysis. All participant-level findings, explanations and component effects are synthetic. Table 4 summarises the answer to each question before the supporting results are presented. RQ4 is answered by the method-development stage rather than the computational demonstration: it produced the 24-field, version-controlled Application Specification recording what must be co-designed and recorded before evaluation begins (completed AVOID-informed illustration in Additional file 2, Table S5) and the reusable co-design workbook for adaptation to another host system (Additional file 3, Table S22).

Table 4. HUMMBER questions and findings from the synthetic demonstration

HUMMBER question	What the demonstration examined	Synthetic finding and interpretation
What changed?	Compared final maintenance with preceding events in five known synthetic scenarios.	Final maintenance was similar across the five scenarios (68.5%-74.4%), while the preceding event patterns differed. A similar endpoint therefore did not identify what had happened beforehand.
For whom?	Reduced technical-support availability for synthetic participants with limited digital access while specifying the same logit-scale support-effect parameter and no digital-access interaction.	Limited digital access reduced opportunities to receive the eligible support options (34.3% vs 72.7% of decision points) and increased missing evidence (8.4% vs 4.2%). The data-generating mechanism specified the same logit-scale support-effect parameter for both groups and no digital-access interaction; partial pooling is presented only as an illustration of shrinkage toward the pooled estimate, not as evidence of homogeneity. The group difference was in opportunity and evidence availability, not a programmed digital-access interaction in the support effect.
Under what circumstances?	Compared passive, immediate, delayed and participant-selected contextual-assessment strategies.	Participant-selected timing produced the greatest useful information per active participant minute (0.451 vs 0.343 delayed and 0.226 immediate). The comparison illustrates an information-burden trade-off, not a universal preference.
What might explain the change?	Updated probabilities across five known synthetic explanations from neutral and deliberately misleading starting distributions.	<i>The highest-probability explanation matched the programmed synthetic process for 90.0% of participants from a neutral start and 87.5% from a misleading start, with probability recovering gradually over the 26 weeks. Weak clustering on sparse snapshots (adjusted Rand index 0.168) did not support converting patterns into fixed participant types.</i>
Did a particular component make a difference?	Randomised only acceptable and available synthetic support components and compared next-week engagement with usual support using the prespecified weighted estimator.	Matched components produced supported synthetic gains in three scenarios; a deliberately small effect remained insufficiently evidenced; and two diagnostic contrasts whose intervals excluded zero were recorded as signals requiring confirmation, not supported effects.
What can evaluators conclude, and what may authorised decision-makers consider next?	Applied evidence-status rules to a null effect, unavailable option, changing effect and disengagement-related missingness, with a performed tipping-point sensitivity analysis.	The method retained insufficient evidence, not identifiable and contradicted/unstable classifications; the informative-missingness claim was downgraded pending sensitivity analysis and retained, as sensitivity-checked, after the two-dimensional tipping-point analysis found no conclusion change within the examined range.

What changed?

The mixed cohort had 71.5% final maintenance and a mean of 7.62 change signals per participant over 26 weeks. The five single-process scenarios ended with similar maintenance (68.5%-74.4%) but different preceding markers: help-seeking and unavailability for technical barriers, incomplete goals for goal difficulty, falling perceived value for declining value, competing demands for competing demands, and accumulated assessment time for assessment burden. Figure 3 shows the contrasting trajectories and similar endpoints; detailed values are reported in Additional file 2, Table S11.

For whom?

Limited digital access reduced opportunities to receive the eligible support options (34.3% vs 72.7% of decision points), technical-assistance exposure (10.6% vs 19.8%), active assessment time (2.86 vs 4.33 mean minutes) and outcome availability (8.4% vs 4.2% missing). Because the data-generating mechanism specified the same logit-scale support-effect parameter and no digital-access interaction, the differences represented unequal opportunity and evidence availability, not a programmed difference in support effect (Additional file 2, Table S12).

Raw age-by-sex technical-assistance estimates ranged from +7.5 to +41.7 percentage points. REML partial pooling reduced that range to +13.7 to +34.7 around the pooled +23.2-point estimate (between-subgroup variance 0.011). The exercise illustrates shrinkage of noisy subgroup estimates and does not establish homogeneity (Additional file 2, Table S18 and Figure S2).

Under what circumstances?

Participants received a mean of 5.46 active contextual assessments and 3.87 active assessment minutes over 26 weeks. Participant-selected timing had the highest response rate (75.8%) and the greatest useful information per active minute (0.451 synthetic units), compared with 0.343 for delayed and 0.226 for immediate assessment. Passive observation required no active participant time and was evaluated separately. Figure 4 displays the information-burden comparison; full outputs are in Additional file 2, Table S13.

What might explain the change?

From equal starting probabilities, the highest-probability explanation matched the programmed process for 90.0% of participants; from the misleading start, concordance was 87.5%. The probability assigned to the true process recovered gradually rather than instantly, showing that starting assumptions remained visible as evidence accumulated. Because the likelihood ratios were derived from the programmed assessment accuracies and the true process was always among the candidate explanations, this

experiment demonstrates computational recovery under prespecified, partly correctly specified synthetic conditions; it does not establish robustness to misspecified likelihood ratios, alternative tempering exponents or an incomplete candidate set. Additional file 2, Figure S1 shows the trajectories and Table S14 reports the updating outputs.

Clustering on sparse eight-week snapshots performed poorly (mapped accuracy 52.2%, adjusted Rand index 0.168, silhouette 0.174), supporting its use here as a negative check rather than a basis for participant phenotypes or mechanisms (Additional file 2, Table S14).

Did a particular component make a difference?

Prespecified valid comparisons identified higher next-week engagement for technical assistance (+23.2 percentage points, 95% CI 15.5 to 30.8; 595 decision points), goal problem-solving (+13.5, 95% CI 5.9 to 21.1; 647) and flexible timing (+14.6, 95% CI 6.4 to 22.7; 562). The deliberately small interactive-voice effect remained insufficiently evidenced (+2.1, 95% CI -7.0 to 11.2; 494). Across these four prespecified matched contrasts, realised inverse-probability weights ranged from 2.0 to 5.0 and effective sample sizes ranged from 487.9 to 638.9 decision points, close to the corresponding analysed counts of 494 to 647; no weight exceeded 5.0. Figure 5 presents the primary and flagged diagnostic comparisons; all contrasts are reported in Additional file 2, Table S15.

Two of 16 non-prespecified diagnostic contrasts had intervals excluding zero despite no programmed effect: goal difficulty/interactive voice -7.8 points (95% CI -15.4 to -0.2) and assessment burden/interactive voice +10.1 (95% CI 1.3 to 18.9). Both were retained as diagnostic signals requiring confirmation, not supported effects.

What can evaluators conclude, and what may authorised decision-makers consider next?

The true-null comparison remained insufficient (+5.9 percentage points, 95% CI -1.7 to 13.6). When the target option was never available, the effect was not identifiable because the comparison did not exist. When the programmed effect reversed direction, the early estimate was +12.1 points and the late estimate -25.8, while the pooled estimate was -4.7 (95% CI -13.5 to 4.1); the evidence status was contradicted/unstable and required reporting by period (Additional file 2, Table S16).

Informative missingness initially downgraded the claim pending sensitivity analysis. The complete-case technical-assistance estimate was +15.1 points (95% CI 7.6 to 22.6). Across the 49-cell tipping-point grid, adjusted estimates ranged from +13.1 to +17.2 and no CI included zero within the examined -0.30 to +0.30 log-odds departures. The claim was therefore retained as sensitivity-checked within that range, without excluding stronger departures beyond it (Additional file 2, Table S17 and Figure S3). In the post hoc extended grid, the conclusion first changed when component-arm missing outcomes were shifted

downwards and usual-support missing outcomes upwards by 1.33 log-odds; component-arm-only shifts required -3.51 log-odds; and usual-support-only shifts did not change the conclusion even at +6.0 log-odds (extended-grid analysis reproducible in Additional file 4, notebook 1).

The stress tests therefore produced bounded outputs rather than forced yes/no answers: insufficient evidence, not identifiable, contradicted/unstable, or supported only after sensitivity review. Human decision-makers could consider those evidence statements, but HUMMBER did not authorise action.

Across 300 consecutive, unselected true-null replications, the normalised IPW estimator had a bias of +1.5 percentage points (Monte Carlo SE approximately 0.50 points), empirical SD of 8.7 points, mean model-based SE of 9.0 points, 95% CI coverage of 94.3% (Monte Carlo SE approximately 1.34 percentage points) and false-positive rate of 5.7% (Monte Carlo SE approximately 1.34 percentage points) (Additional file 2, Table S19). These values are simulation-based estimates and should be interpreted with the corresponding Monte Carlo uncertainty. All 29 automated integrity checks passed during package preparation. The peer-review notebooks regenerate the reported analyses and figures from the supplied wholly synthetic data and include targeted reproducibility checks against documented manuscript values.

Discussion

Principal findings

The synthetic implementation demonstrated that the evaluation architecture can keep different kinds of evidence connected without treating them as equivalent under the documented demonstration conditions. Similar final outcomes could be linked back to different trajectories; unequal opportunity to receive support could be separated from estimated support effect; assessment burden remained visible; possible explanations could be revised as evidence accumulated; and causal estimates were restricted to valid comparisons.

The demonstration also showed that uncertainty is a legitimate evaluation output. HUMMBER did not force a conclusion when a comparison was unavailable, an effect changed over time, missingness threatened interpretation or a diagnostic contrast appeared noteworthy outside the prespecified primary question.

Together, these findings support the internal logic of the proposed method under known synthetic conditions. They do not establish its feasibility or effectiveness in practice; those questions require prospective real-world evaluation.

Methodological contribution

HUMMBER integrates complementary capabilities identified in the exploratory evidence map [4–11, 17, 18, 20, 21, 23–30] rather than treating any one method as sufficient for the full evaluation problem.

The integration matters because the questions require different evidence. Observation can establish that a pattern changed, but not why. Qualitative, experiential and contextual evidence can strengthen or weaken a possible explanation or mechanism, but cannot by itself establish a component effect. Causal estimation is introduced only when an appropriate comparison exists. Governance then records the strongest conclusion the accumulated evidence supports before any authorised human considers action.

Realist explanation and causal attribution remain deliberately separate. A context-mechanism-outcome account can explain how an outcome may arise under particular circumstances [9, 18], whereas a causal component effect requires a valid comparison. Keeping both forms of evidence in the same evolving record allows them to inform one another without overstating what either can prove.

Co-design is part of that methodological logic. The specification was informed by transdisciplinary author-team consultation, including a consumer perspective, and prospective applications require relevant stakeholders to specify what matters, what can be measured or compared, what burden and safeguards are acceptable, and who holds responsibility for interpretation and action. The Application Specification makes those choices explicit and versioned rather than leaving them implicit in an analysis.

HUMMBER provides a practical route for Qual-Quant integration during ongoing evaluation. Quantitative signals can identify where closer enquiry is warranted; qualitative and experiential evidence can clarify circumstances and plausible mechanisms; and quantitative comparison can follow when the question and design support an effect estimate. The present study mainly tested computational functions, so real-world work is needed to determine whether the full integration is understandable, feasible and useful.

Implications for adaptive and agentic AI

Adaptive and agentic AI are important use cases because the evaluated system may change its actions, tools, prompts, rules or versions during use. Current guidance emphasises fairness, traceability, explainability, validation, lifecycle monitoring, ethics, safety and continuing human oversight [12–15, 31]. This human-control focus is also consistent with emerging AI-safety expectations, such as the Australian Voluntary AI Safety Standard guardrail that AI systems should enable human control or intervention to achieve meaningful human oversight [33]. HUMMBER can retain which system action and version were active, the surrounding circumstances, what the person experienced and what followed, so evidence generated under different system states is not treated as interchangeable.

For agentic AI, the system action is not automatically the mechanism. The same recommendation or automated action may lead to different outcomes depending on understanding, trust, prior experience, resources, competing demands or service context. HUMMBER can therefore retain candidate context-mechanism-outcome explanations and examine how qualitative, experiential, contextual and quantitative

evidence supports or challenges them over time [9, 18]. Explainability remains focused on observable actions and consequences; it does not claim access to hidden model reasoning.

Human governance is equally important. Consumers, clinicians, service leaders, evaluators and technical specialists may each see different parts of the same human-system interaction [34]. The Application Specification provides a place to make those perspectives, safeguards, acceptable comparisons and decision responsibilities explicit. The software can support recording and analysis, but it cannot decide what evidence means or what should happen next [34].

Prospective agentic-AI studies should test whether actions and version changes can be captured with sufficient fidelity, whether plausible mechanisms can be refined without excessive burden, whether causal claims remain bounded by valid comparisons, and whether named human responsibilities remain workable as the system changes [34].

Implications for scale-up and equity

Scale-up can preserve an acceptable overall outcome while changing who can access support, who contributes evidence and why people continue or disengage. The equity stress test showed why representation, opportunity, exposure, refusal, burden, missingness and effect should be analysed separately. Additional file 2, Table S12 reports the opportunity and missingness outputs; Table S18 and Figure S2 show how partial pooling changes noisy subgroup estimates without proving homogeneity.

Prospective applications should therefore identify equity questions during co-design, not only after subgroup differences appear. Consumers can identify accessibility barriers, burdens and outcomes that may not be visible in system logs, while clinical, service, methodological and technical expertise can determine what can validly be measured or compared.

Strengths and limitations

A major strength is the use of known synthetic conditions. Prespecified processes and effects allowed HUMMBER to be examined when evidence was supportive, null, contradictory, unavailable or affected by missingness. The study also separated published AVOID evidence from synthetic participant-level data, specified the estimator and updating rule, performed the planned sensitivity analysis, and used executable notebooks and automated integrity checks during computational verification. Additional files 1–5 provide the evidence map, technical specification, co-design materials, peer-review reproducibility materials and consumer brief.

Transdisciplinary author-team consultation, including a consumer perspective, informed the specification, but the development process was not itself a prospective formal co-design study. The study cannot establish whether a broader or different stakeholder group would choose the same evaluation factors, mechanisms, safeguards or governance arrangements, or whether participation can be reproduced feasibly across settings.

The synthetic demonstration does not predict real-world frequencies, burden, missingness, subgroup effects or component effects, and the 600-person mixed cohort is not a sample-size recommendation. The study did not test real qualitative enquiry, live EKE, routine service workflows, governance processes, regulatory requirements or a live agentic AI system. It therefore cannot establish understandability, acceptability, feasibility, clinical effectiveness, fairness, safety, deployment readiness or valid personalisation for the AVOID intervention or another host system. Explanation updating was examined only with likelihood ratios derived from the programmed assessment accuracies and a single tempering exponent; robustness to misspecified likelihoods, alternative tempering and incomplete candidate explanation sets remains untested.

The exploratory evidence map was not a systematic review and may have missed relevant approaches. The optional software layer also requires independent security, reliability, maintainability and usability evaluation before live implementation.

Next steps

Prospective studies should begin by testing whether consumers, practitioners, service decision-makers, technical teams and evaluators can use the Application Specification to define meaningful outcomes, explanations and mechanisms, acceptable measurements and comparisons, burden and equity safeguards, evidence rules and decision responsibilities without disproportionate complexity.

Real-world studies should then test whether the intended Qual-Quant and realist integration improves understanding of changing outcomes, whether human roles and governance remain workable, and whether the method transfers across different host systems and populations. Co-design should also be evaluated directly, including whose perspectives are represented, how disagreement is handled and whether participation is accessible and meaningful [34].

Adaptive and agentic AI should be a priority for that work because changing system actions and versions make longitudinal traceability and human accountability especially important. Technical validation of the software layer should proceed in parallel with methodological validation. Additional file 3 provides practical co-design and transfer materials, and Additional file 4 provides the reproducible computational starting point.

Conclusions

HUMMBER, specified with transdisciplinary author-team input including a consumer perspective and designed for co-design in live applications, is a proposed human-governed framework for ongoing evaluation of digital and AI-enabled behaviour change interventions as the intervention, its users and the circumstances of use change over time. The synthetic implementation demonstrated that the evaluation architecture can preserve differences hidden by similar endpoints, keep explanation and causation

distinct, make burden and unequal opportunity visible, and retain uncertainty when the evidence cannot support a stronger conclusion under the documented demonstration conditions.

The next stage is prospective real-world testing of the complete method and its optional software support. Studies should examine understandability, feasibility, acceptability, usefulness and transferability; the intended Qual-Quant and realist integration; and the reliability of automated analytical support while human interpretation and decision authority are preserved. Adaptive and agentic AI are priority applications because changing system actions and versions increase the need to trace what happened, for whom, under which circumstances, what may explain the consequences and what can legitimately be attributed to the system.

Abbreviations

ADEMP: Aims, Data-generating mechanisms, Estimands, Methods and Performance measures; AI: artificial intelligence; AVOID: Activity; Vaccinate; Optimise medications; Interact; and Diet and nutrition, including oral health; CI: confidence interval; DBCI: digital behaviour change intervention; EKE: expert knowledge elicitation; HUMMBER: Human-centred, User-informed, Multimethod Micro-randomised Behaviour Evaluation and Response; JITAI: just-in-time adaptive intervention; LRE: longitudinal realist evaluation; IPW: inverse probability weighting; MNAR: missing not at random; MRT: micro-randomised trial; PE: process evaluation; REML: restricted maximum likelihood; WCLS: weighted and centred least squares.

Declarations

Ethics approval and consent to participate

Not applicable. This methodological study used wholly synthetic participant-level data and published aggregate findings. No human participant records, tissue or identifiable data were accessed or reconstructed for the simulation.

Consent for publication

Not applicable.

Availability of data and materials

The confidential peer-review reproducibility package submitted as Additional file 4 contains the wholly synthetic datasets, machine-readable data dictionary and two executable notebooks used for the computational demonstration. The notebooks embed the simulation and analysis code and regenerate

the reported analyses from the supplied synthetic data. No participant-level AVOID trial data were accessed, reconstructed or imputed. Public release of the broader HUMMBER software layer is subject to institutional intellectual-property and commercialisation review.

If public availability of computational materials is required as a condition of publication, the applicable access and licensing arrangements will be determined through the relevant institutional processes before release. Clinical or service deployment would additionally require appropriate methodological and technical validation and applicable approvals.

Competing interests

Potential intellectual-property and commercialisation interests in HUMMBER computational materials are under institutional consideration; any future licensing would be subject to applicable institutional intellectual-property policies and approvals. All other authors declare that they have no competing interests in relation to this research.

Funding

No specific funding was received for the methodological simulation reported in this manuscript.

Authors' contributions

Maria Alejandra Piñero de Plaza (MAPdP): Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualisation and Project administration. Ian Halkett (IH): Methodology review, consumer perspective, Validation and Writing - review and editing. Damith Ranasinghe (DR): Methodology review, technical and software review, Validation and Writing - review and editing. Carlos Javier Barrera Causil (CJBC): Methodology review, Formal analysis review, Validation and Writing - review and editing; independently checked the published AVOID trial quality-of-life estimate and confidence interval. Joanne Dollard (JD): Methodology review, Validation and Writing - review and editing. Renuka Visvanathan (RV): Conceptualisation review, Methodology review, Supervision and Writing - review and editing. All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the AVOID trial participants and citizen scientists whose published findings motivated the work demonstrated, and the research communities that contributed to the methodological lineage represented in the cited work. The authors further acknowledge that the AVOID trial was made possible because of funding from the Hospital Research Foundation.

OpenAI ChatGPT and Kimi AI were used as assistive tools during iterative drafting and code development. The tools did not author the work or exercise autonomous decision-making. The corresponding author directed the conceptual development, supplied and checked source material, reviewed AI-assisted code and generated outputs, and accepts responsibility for the integrity of those contributions. Co-authors remain responsible for the content they reviewed and approved. Neither tool made any autonomous scientific, clinical or governance decisions.

Authors' information

Not applicable.

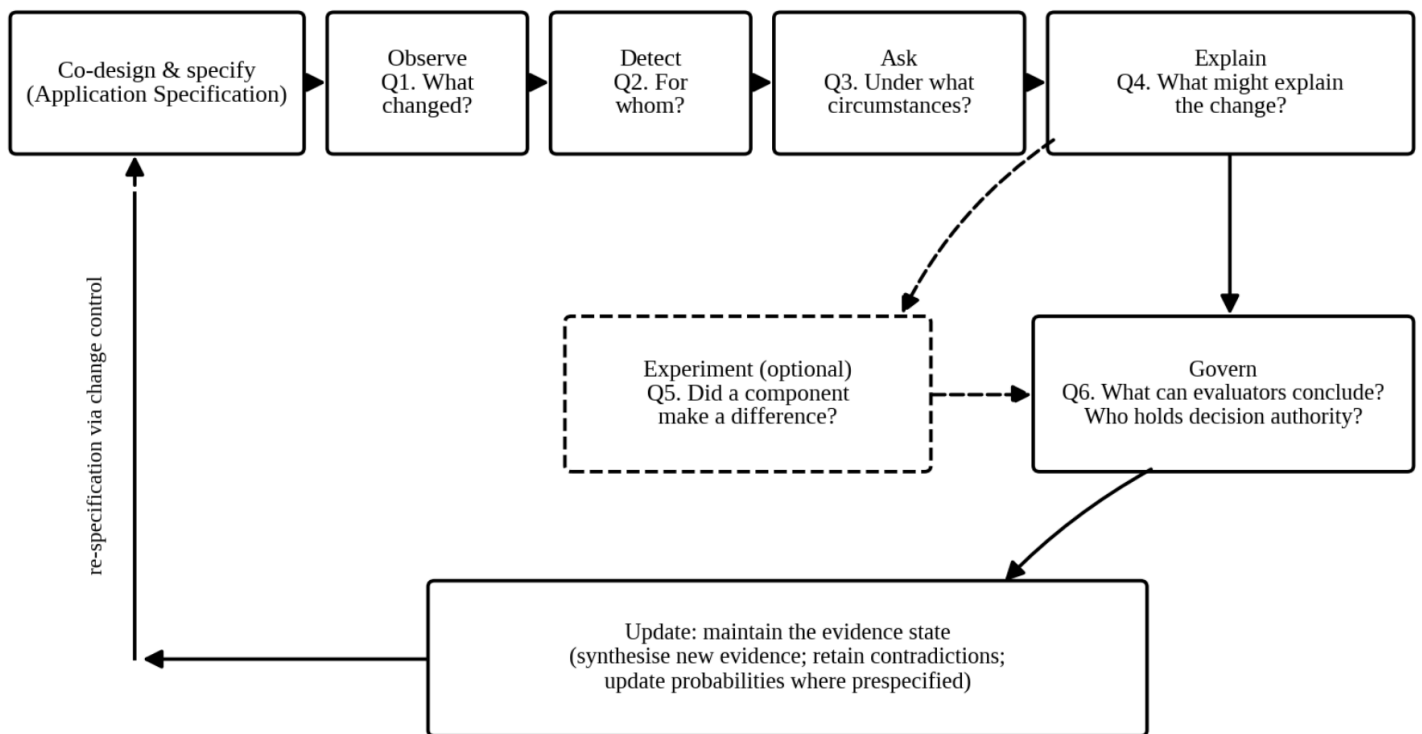
References

1. Ananthakrishnan A, Cong C, Riccalton V, Meinert E (2025) Progressing implementation of behavior change frameworks for digital health interventions: challenges and ways forward. *Transl Behav Med* 15(1):ibaf069. 10.1093/tbm/ibaf069
2. World Health Organization. Health workforce [Internet]. Geneva: World Health Organization (2026) [cited 2026 Aug 20]. Available from: <https://www.who.int/health-topics/health-workforce>
3. Bucher A, Blazek ES, Symons CT (2024) How are machine learning and artificial intelligence used in digital behavior change interventions? A scoping review. *Mayo Clin Proc Digit Health* 2(3):375–404. 10.1016/j.mcpdig.2024.05.007
4. Bracci EL, Collyer S, Oster C, Lynch E, Olasunkanmi-Alimi T, Bulto LN et al (2025) Do process evaluations open up the 'black box' of implementation interventions in health care? A scoping review. *Implement Sci* 20:55. 10.1186/s13012-025-01468-x
5. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A et al (2018) Just-in-time adaptive interventions in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med* 52(6):446–462. 10.1007/s12160-016-9830-8
6. Hsu TCC, Whelan P, Gandrup J, Armitage CJ, Cordingley L, McBeth J (2025) Personalized interventions for behaviour change: a scoping review of just-in-time adaptive interventions. *Br J Health Psychol* 30(1):e12766. 10.1111/bjhp.12766
7. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A et al (2015) Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. *Health Psychol* 34(Suppl):1220–1228. 10.1037/hea0000305
8. Qian T, Walton AE, Collins LM, Klasnja P, Lanza ST, Nahum-Shani I et al (2022) The micro-randomized trial for developing digital interventions: experimental design and data analysis considerations. *Psychol Methods* 27(5):874–894. 10.1037/met0000283
9. Archibald MM, Lawless MT, Piñero de Plaza MA, Kitson AL (2023) How transdisciplinary research teams learn to do knowledge translation, and how knowledge translation in turn impacts

- transdisciplinary research: a realist evaluation and longitudinal case study. *Health Res Policy Syst* 21(1):20. 10.1186/s12961-023-00967-x
10. Pinero de Plaza MA, Yadav L, Kitson A (2023) Co-designing, measuring, and optimizing innovations and solutions within complex adaptive health systems. *Front Health Serv* 3:1154614. 10.3389/frhs.2023.1154614
 11. Pinero de Plaza MA, Lambrakis K, Marmolejo-Ramos F, Beleigoli A, Archibald M, Yadav L et al (2025) Human-centred AI for emergency cardiac care: evaluating RAPIDx AI with PROLIFERATE_AI. *Int J Med Inf* 196:105810. 10.1016/j.ijmedinf.2025.105810
 12. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP et al (2025) FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 388:e081554. 10.1136/bmj-2024-081554
 13. World Health Organization (2026) Artificial intelligence-related health research: ethics review and oversight. World Health Organization, Geneva
 14. Phelps N, Vossler P, Kothari A, Steineman K, Goldman S, Gasga A et al (2026) Patient perspectives about deployment of artificial intelligence decision support tools in a safety-net healthcare system. *JAMIA Open* 9(2):ooag029. 10.1093/jamiaopen/ooag029
 15. Bharija A, Khan A, Pavon JM, Hashmi AZ, Jindal SK, Dredze M et al (2026 Jul) Generative AI in the care of older adults: a position statement from the American Geriatrics Society. *J Am Geriatr Soc* 21. 10.1111/jgs.70566
 16. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. *Stat Med* 38(11):2074–2102. 10.1002/sim.8086
 17. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ et al (2006) Uncertain judgements: eliciting experts' probabilities. Wiley, Chichester
 18. Pawson R, Tilley N (1997) Realistic evaluation. Sage, London
 19. Dollard J, Jadcak AD, Halkett I, Tucker GR, Thompson MQ, Mahadavan S et al (2026) A hybrid type 1 trial of a digital behaviour change intervention for frailty prevention in community-dwelling older adults aged 60 years and older. *Age Ageing* 55(5):afag120. 10.1093/ageing/afag120
 20. Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge
 21. Boruvka A, Almirall D, Witkiewitz K, Murphy SA (2018) Assessing time-varying causal effect moderation in mobile health. *J Am Stat Assoc* 113(523):1112–1121. 10.1080/01621459.2017.1305274
 22. White IR, Horton NJ, Carpenter J, Pocock SJ (2011) Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* 342:d40. 10.1136/bmj.d40
 23. Johansson JF, Lam N, Ozer S, Hall J, Morton S, English C et al (2022) Systematic review of process evaluations of interventions in trials investigating sedentary behaviour in adults. *BMJ Open* 12(1):e053945. 10.1136/bmjopen-2021-053945

24. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B et al (2019) Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. *Ann Behav Med* 53(6):573–582. 10.1093/abm/kay067
25. Nahum-Shani I, Greer ZM, Trella AL, Zhang KW, Carpenter SM, Ruenger D et al (2024) Optimizing an adaptive digital oral health intervention: micro-randomized trial protocol. *Contemp Clin Trials* 139:107464. 10.1016/j.cct.2024.107464
26. Liao P, Klasnja P, Tewari A, Murphy SA (2016) Sample size calculations for micro-randomized trials in mHealth. *Stat Med* 35(12):1944–1971. 10.1002/sim.6847
27. Browne J, Halverson TF, Vilardaga R (2021) Engagement with a digital therapeutic for smoking cessation designed for persons with psychiatric illness fully mediates smoking outcomes in a pilot randomized controlled trial. *Transl Behav Med* 11(9):1717–1725. 10.1093/tbm/ibab100
28. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K et al (2016) Understanding and promoting effective engagement with digital behavior change interventions. *Am J Prev Med* 51(5):833–842. 10.1016/j.amepre.2016.06.015
29. De La Torre SA, El Mistiri M, Hekler E, Klasnja P, Marlin B, Pavel M et al (2024) Modeling engagement with a digital behavior change intervention (HeartSteps II): an exploratory system identification approach. *J Biomed Inf* 158:104721. 10.1016/j.jbi.2024.104721
30. Shiffman S, Stone AA, Hufford MR (2008) Ecological momentary assessment. *Annu Rev Clin Psychol* 4:1–32. 10.1146/annurev.clinpsy.3.022806.091415
31. Mohr DC, Cheung K, Schueller SM, Hendricks Brown C, Duan N (2015) Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. *J Med Internet Res* 17(7):e166. 10.2196/jmir.4391
32. Hrynyschyn R, Prediger C, Stock C, Helmer SM (2022) Evaluation methods applied to digital health interventions: what is being used beyond randomised controlled trials? A scoping review. *Int J Environ Res Public Health* 19(9):5221. 10.3390/ijerph19095221
33. Australian Government Department of Industry, Science and Resources. Voluntary AI Safety Standard: Guardrail 5 – enable human control or intervention in an AI system to achieve meaningful human oversight [Internet]. Canberra: Commonwealth of Australia (2024) [cited 2026 Sep 29]. Available from: <https://www.industry.gov.au/publications/voluntary-ai-safety-standard/10-guardrails>
34. Pinero de Plaza MA (2026) Making human/AI systems an object of evidence. *JBHI, World Evidence-Based Healthcare Day*. 30. 10.70253/FSVN1820

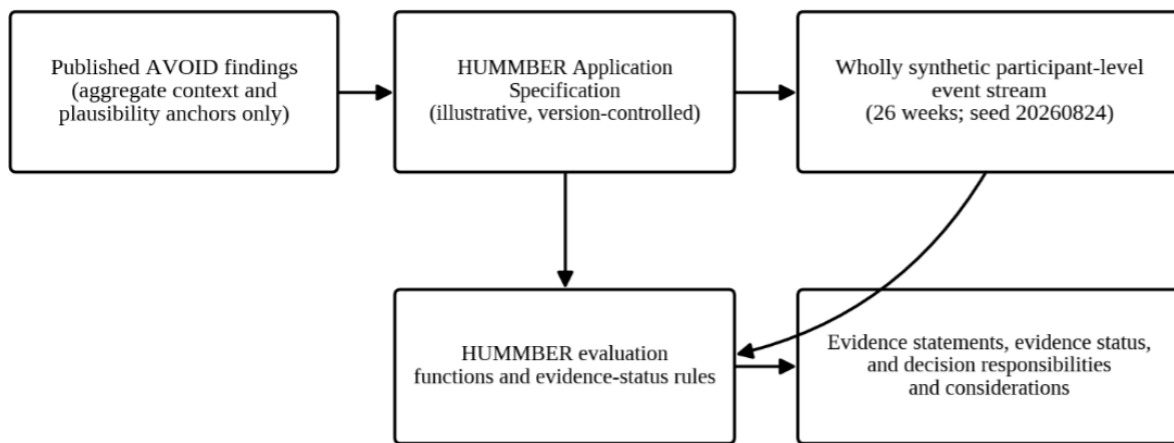
Figures



Solid route: always available. Dashed branch: only where a valid, approved comparison exists. Safeguards and human decision authority apply throughout.

Figure 1

HUMMBER co-design and ongoing evaluation structure. The six procedural questions are operationalised through Observe, Detect, Ask, Explain, optional Experiment, and Govern. Update maintains the evolving evidence state and is not a seventh question; probabilities are updated only where prespecified. The return path to Co-design and specify represents re-specification under change control. Safeguards and human decision authority apply throughout.



The synthetic event stream is generated under the illustrative Application Specification and prespecified simulation rules, then evaluated using the HUMMBER functions. No participant-level AVOID data were accessed, reconstructed or imputed.

Figure 2

HUMMBER applied to the synthetic AVOID trial-informed scale-up demonstration. Published trial findings provide whole-program context only. The illustrative Application Specification defines the evaluation boundary, and prespecified simulation rules generate the wholly synthetic participant-level event stream evaluated by HUMMBER. No participant-level trial data were accessed, reconstructed or imputed.

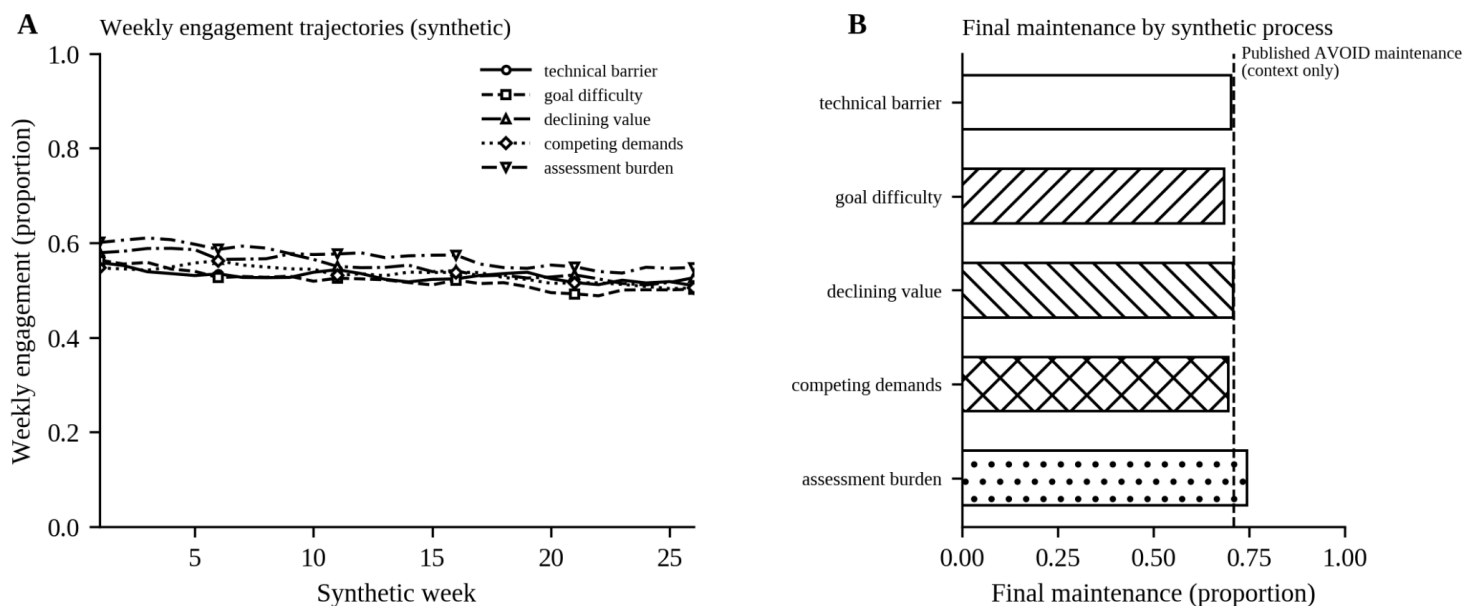


Figure 3

Similar final maintenance under different known synthetic processes. Panel A shows weekly engagement trajectories (three-week centred moving average) for each synthetic process. Panel B

shows final maintenance (at least two engaged weeks in weeks 23-26) by synthetic process; the dashed vertical line marks the published AVOID maintenance percentage (71.0%) as context only. The synthetic scenarios were not constructed to reproduce participant-level AVOID data.

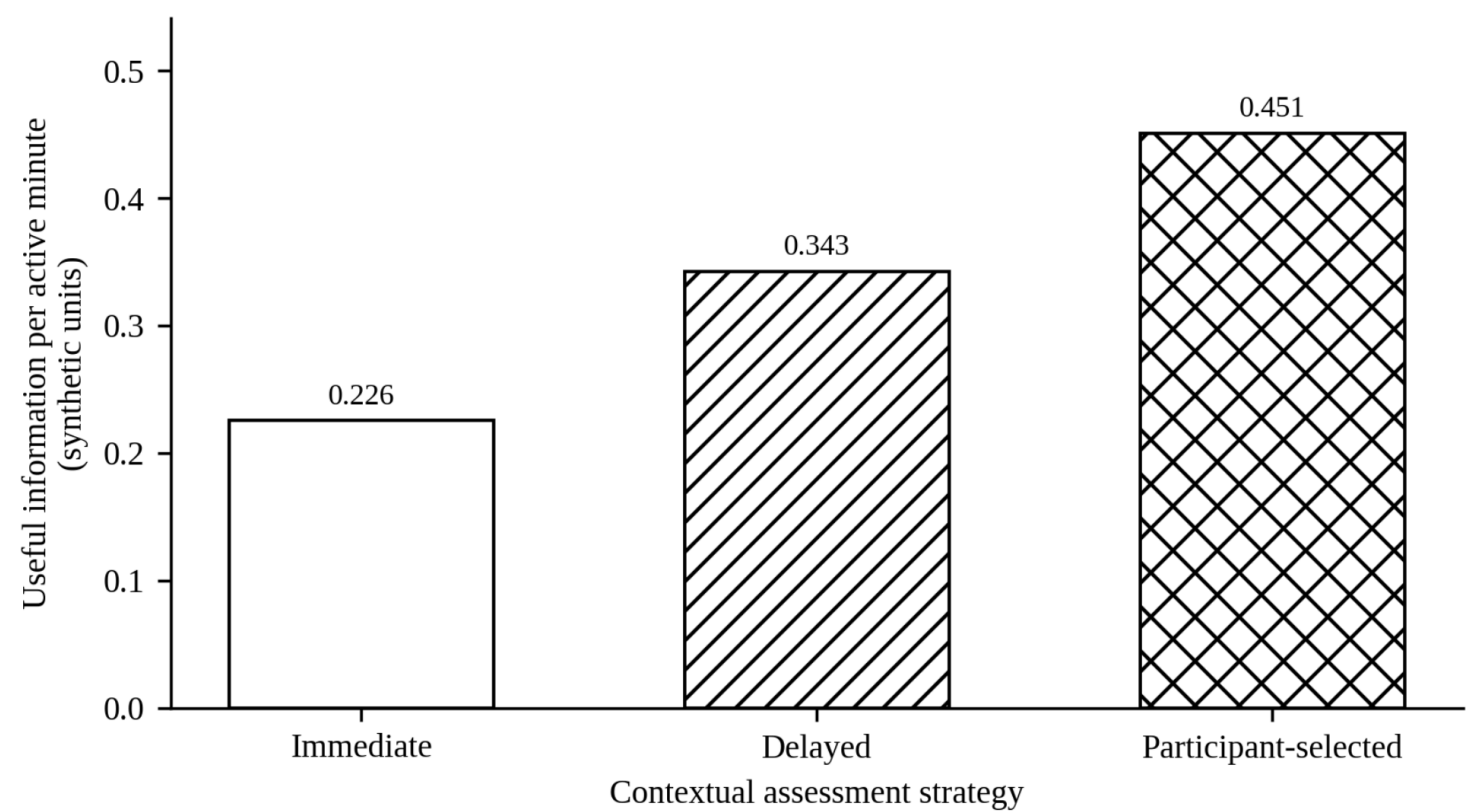


Figure 4

Useful contextual information per active participant assessment minute (synthetic units). Passive observation is not plotted because it requires no active participant assessment time; the comparison illustrates the information-burden trade-off rather than a universally preferred assessment strategy.

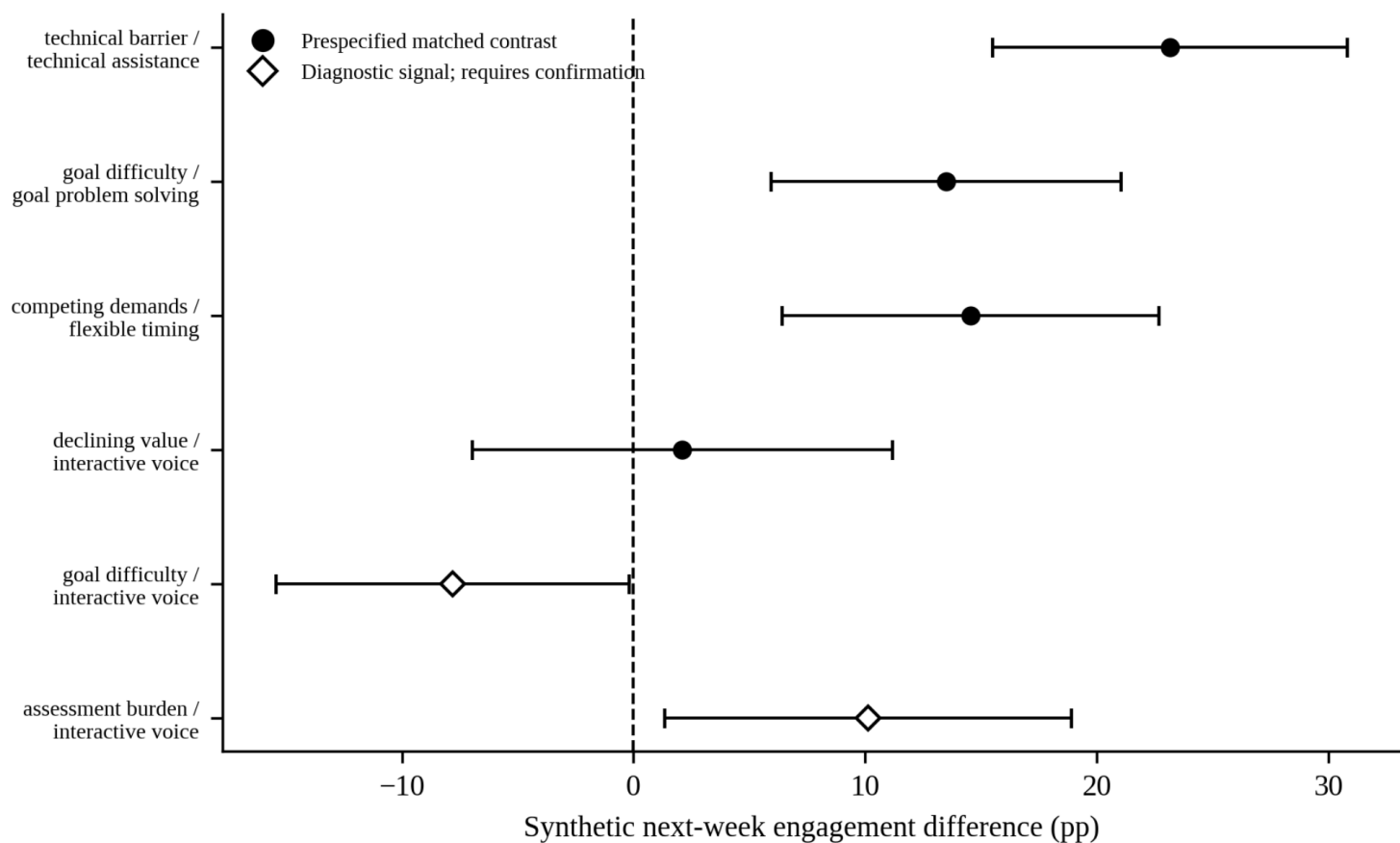


Figure 5

Synthetic next-week engagement differences for the prespecified matched support components compared with usual support, plus the two diagnostic contrasts flagged as signals requiring confirmation. Points show percentage-point differences and horizontal lines show 95% confidence intervals from the weighted estimator with participant-clustered standard errors. The comparisons are synthetic and are not AVOID trial treatment effects.