Research Article

Epistemic Governance in Social Media Radio Learning

Sherif Rashwan[1]

1. Faculty of Mass Communication, Ahram Canadian University, Egypt

Social media radio platforms combine audio streaming, chat, reactions, credibility cues, and recommendation systems in learning environments shaped by social interaction and algorithmic visibility. This protocol proposes a two-week randomized active-sham-controlled field trial testing whether epistemic governance features improve students' confidence calibration, source diversity, reflective evaluation, and media-literacy judgment. The intervention arm receives credibility indicators, source-diversity feedback, algorithmic cooling of homogeneous exposure, ecological momentary assessment prompts, and recommendation explanations. The sham arm receives the same content, interface burden, support, notifications, and data-collection intensity without active epistemic mechanisms. The manuscript specifies theory, operational definitions, measures, implementation procedures, fidelity checks, ethical safeguards, and a mixed-effects analysis plan. Its contribution is a media-literacy-focused protocol for evaluating whether platform governance can shift social audio learning from engagement-centred interaction toward evidence-sensitive judgment.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

Introduction

Social media radio platforms are emerging digital environments in which learners listen to live or recorded audio streams while interacting through comments, reactions, polls, peer cues, and recommendation systems. Their educational promise lies in their ability to combine the accessibility of audio learning with the social presence of networked platforms. However, the same affordances can also encourage attention to popularity, speed, identity alignment, and repeated exposure rather than evidence quality. For educational use, the key empirical question is not whether students enjoy a social audio

platform, but whether platform governance can improve how students evaluate claims, diversify sources, regulate confidence, and reflect on their own understanding.

From a mass communication perspective, social media radio can be understood as a hybrid of interpersonal communication, platformed broadcasting, and algorithmic gatekeeping. It resembles radio in its liveness, voice-centred intimacy, and parasocial potential, but it also resembles social media because audience reactions, network ties, and recommendation systems shape what becomes visible. This makes the platform relevant to classic communication questions about agenda-setting, selective exposure, two-step flow, media credibility, uses and gratifications, and diffusion of innovations[1][2][3][4]. The proposed intervention therefore treats the platform not merely as an instructional tool but as a mediated communication environment in which cues, sources, audiences, and algorithms jointly influence learning.

This study proposes a valid empirical evaluation of a generic social media radio platform. The study is designed as a randomized, sham-controlled field trial rather than a usability demonstration. The intervention is not the livestream itself. Both study arms receive the same stream, content, timing, support, and interface burden. The causal contrast is the presence versus absence of epistemic governance features: valid credibility signals, source-diversity feedback, feed cooling when a participant's stream becomes homogeneous, reflective EMA prompts, and transparent recommendation explanations.

The design follows established principles of randomized evaluation, active control conditions, pilot and feasibility research, and construct validation[5][6][7][8]. It is intentionally generic so that it can be implemented in different universities, colleges, online programmes, community-learning settings, or professional-development contexts without changing the causal structure.

The article is written as a pre-field empirical protocol. It makes the causal contrast, data structures, operational procedures, and analysis decisions explicit before recruitment. The absence of outcome data is therefore a feature of the article type: the study specifies what should be tested, how the platform should be implemented, which variables should be captured, and how analysis should proceed before participants are exposed to the system.

Critical Literature Review and Theoretical Rationale

The theoretical case for this study begins with a tension in mass communication research. Classic broadcast theories treated media as institutional systems that could structure public attention, influence the salience of issues, and distribute interpretive frames across large audiences. Agenda-setting research showed that media may be powerful not simply because they tell audiences what to think, but because they influence what audiences think about[1]. Gatekeeping theory further demonstrated that communication systems are shaped by decisions about what enters, leaves, and becomes prioritized within a channel[9]. Social media radio transforms these older questions because selection is no longer performed only by editors, producers, or teachers. It is also performed by algorithms, reaction metrics, peer signals, recommendation histories, and platform defaults. Therefore, the study treats platform design itself as a communicative intervention.

A second strand of literature concerns audience activity. Uses and gratifications research challenged passive-audience assumptions by asking what people seek from media, such as information, identity, companionship, entertainment, and social integration[2]. Social media radio is strongly aligned with this tradition because learners may use it for lecture access, peer belonging, background listening, social recognition, and informal interpretation. Yet the same audience activity can weaken learning if users select only congenial sources, reward charismatic speakers regardless of evidence, or mistake repeated exposure for validity. The active sham design responds to this problem: it holds motivation, novelty, and social presence constant while testing whether epistemic governance adds value beyond general platform engagement.

A third tradition is the two-step flow and diffusion of innovations. The two-step flow model emphasized that interpersonal networks and opinion leaders mediate the influence of mass communication[3]. Diffusion theory similarly explains how innovations move through social systems through perceived advantage, compatibility, complexity, trialability, and observability[4]. In social media radio, students may treat peer reactions, badges, follower counts, or confident voices as credibility shortcuts. This creates a methodological challenge: if the platform increases engagement, it may be unclear whether learning improves because of better evidence evaluation or because of social contagion. The proposed trial therefore separates social interaction from epistemic validity by making the sham platform socially plausible but epistemically inactive.

A fourth body of research focuses on source credibility and digital credibility heuristics. Hovland and Weiss demonstrated that the same message can be interpreted differently depending on perceived source credibility[10]. In contemporary digital settings, credibility is often inferred from interface cues, social endorsement, reputation metrics, and surface fluency rather than from careful evaluation of evidence[11][12]. This literature directly supports the intervention logic. The platform should not merely display popularity; it should help learners distinguish between attention, authority, evidence, transparency, and uncertainty. Credibility indicators in the intervention arm are therefore operationalized as explicit epistemic cues rather than generic reputation scores.

A fifth strand concerns misinformation, filter bubbles, and selective exposure. Digital journalism research has shown that false or misleading content is difficult to define and detect because misinformation can involve fabrication, misleading context, satire, manipulation, or selective emphasis[13]. Selective exposure research suggests that audiences often prefer information that confirms prior attitudes, especially when identity, emotion, or political relevance is high[14][15]. In social media environments, peer endorsement can also override traditional source cues when people select news, which makes reaction metrics and peer signals theoretically important rather than merely decorative interface features[16]. In educational settings, the equivalent problem is not only ideological polarization; it is also epistemic narrowing, where learners repeatedly encounter the same source type, explanation style, or peer consensus. The proposed source-diversity feedback and algorithmic cooling mechanisms are designed to interrupt homogeneous exposure without censoring required instructional material.

A sixth relevant literature is media literacy. Media literacy is not reducible to access, frequency of use, or technical skill. It involves the ability to access, analyze, evaluate, create, and reflect on media messages in context[17][18]. News media literacy research adds that learners need to understand how news is produced, how evidence is evaluated, and how media systems shape information flows[19]. For that reason, this study does not treat time-on-platform as a sufficient outcome. It distinguishes raw engagement from evidence-seeking, uncertainty monitoring, source comparison, and calibrated confidence.

A seventh literature concerns educational technology, community inquiry, and multimedia learning. Social audio can reduce access barriers and support presence, but it can also increase cognitive load if learners must listen, chat, react, judge credibility, and respond to prompts simultaneously. Community-of-inquiry research emphasizes cognitive, social, and teaching presence in online learning environments, making it relevant to a platform that combines instructional audio with peer interaction[20]. Multimedia

learning theory and cognitive-load measurement suggest that educational technologies should be evaluated not only by satisfaction but also by how they allocate attention and support meaningful processing[21][22]. The intervention therefore uses short, timed EMA prompts and concise credibility explanations rather than long reflective tasks that could overwhelm learners.

A final theoretical contribution is the distinction between communicative affordance and epistemic affordance. A communicative affordance allows users to speak, listen, react, share, or affiliate. An epistemic affordance helps users ask whether a claim is warranted, whether a source is credible, whether a viewpoint is missing, and whether confidence is justified. Many platforms provide rich communicative affordances but weak epistemic affordances. The intervention tests whether adding epistemic affordances to a social audio environment changes the quality of learning-related judgment without removing the social features that motivate participation.

Taken together, the literature suggests that a rigorous study of social media radio must avoid two weak claims. The first weak claim is technological optimism: assuming that a lively platform automatically improves learning. The second is technological pessimism: assuming that social platforms necessarily degrade evidence evaluation. The stronger empirical position is to test whether specific governance mechanisms can redirect attention from popularity and repetition toward evidence, diversity, and reflection while preserving the accessibility and social presence that make audio platforms attractive.

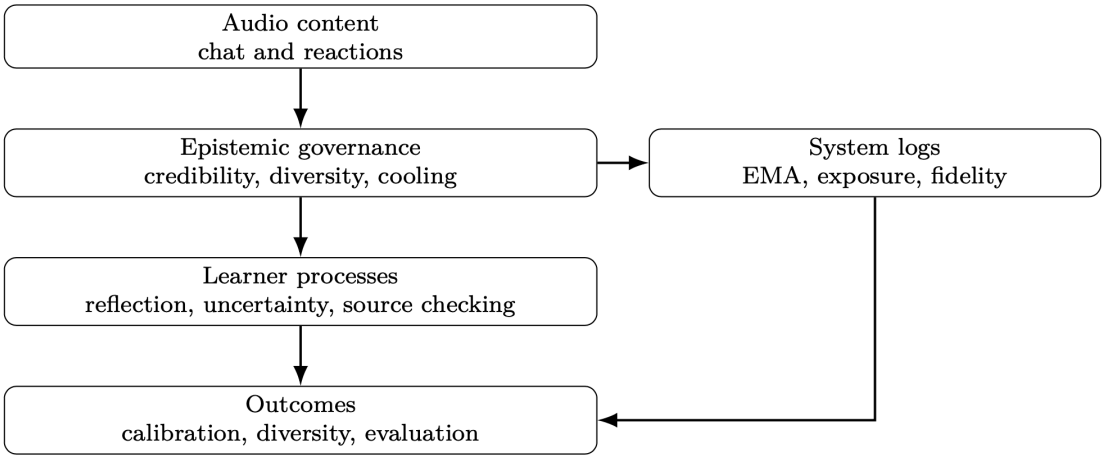


Figure 1. Conceptual pathway linking platform governance, learner processes, behavioural traces, and outcomes.

Research Questions and Hypotheses

RQ1: Does an epistemically governed social media radio platform improve students' epistemic confidence calibration over two weeks compared with an active sham platform?

H1: Participants in the intervention arm will show a larger positive change in epistemic confidence calibration from baseline to Day 14 than participants in the sham arm.

RQ2: Does source-diversity feedback and algorithmic cooling reduce homogeneous exposure and increase awareness of filter bubbles?

H2: Participants in the intervention arm will show higher source entropy, more diverse interaction patterns, and stronger filter-bubble awareness than participants in the sham arm.

RQ3: Are the EMA measures sufficiently reliable and valid for use in a larger confirmatory trial?

H3: EMA scales for perceived credibility, epistemic monitoring, knowledge reflection, and platform trust will demonstrate acceptable internal consistency, interpretable factor structure, and meaningful relations with behavioural indicators.

RQ4: Do survey, behavioural, and system-log indicators converge into a coherent validity pattern?

H4: Improvements in confidence calibration, reflective knowledge appraisal, source diversity, and credibility-sensitive interaction will be positively associated in the intervention arm.

Conceptual and Operational Definitions

A social media radio platform is defined conceptually as an audio-centred digital communication environment in which learners consume live or recorded streams while also interacting through social and algorithmic features such as chat, reactions, peer cues, recommendation feeds, and reputation indicators. Operationally, the platform is the software environment that records stream exposure, content access, prompts, visible cues, reactions, comments, moderation events, and technical conditions under a study-specific feature-flag system.

Epistemic governance refers to platform rules that make evidence quality, source diversity, uncertainty, and reflection more visible than popularity alone. Operationally, epistemic governance is present when the participant interface displays valid credibility indicators, source-diversity feedback, algorithmic cooling events, epistemic EMA prompts, and brief recommendation explanations generated from a locked rubric. It is absent in the sham arm, where the interface remains visually plausible but does not provide valid epistemic cues.

Credibility indicators are not general reputation points. Conceptually, they represent the degree to which a source or item is transparent, evidence-based, accountable, and appropriate for the learning context. Operationally, a credibility indicator may combine documented authorship, evidence citation, correction history, content review status, source expertise, and alignment with verified instructional material. The rubric should be frozen before recruitment and should be auditable after the study.

For replication, the credibility indicator is calculated on a 0–100 scale before exposure and then displayed as a categorical cue rather than as a raw score. The locked scoring rule is: evidence citation and traceability = 30 points, source expertise or institutional accountability = 20 points, transparency of authorship, funding, method, or uncertainty = 20 points, correction history and responsiveness to error = 15 points, and alignment with verified course or domain material = 15 points. Items scoring 75–100 are labelled “strongly supported,” 50–74 “partly supported,” 25–49 “weakly supported,” and 0–24 “unsupported or unclear.” Popularity metrics, likes, follower counts, and recency do not contribute to this score. Two trained raters should code a calibration subset before deployment; disagreements greater than 15 points should be resolved before the rubric is locked.

Component	Weight
Evidence citation and traceability	30%
Source expertise or institutional accountability	20%
Transparency of authorship, funding, method, or uncertainty	20%
Correction history and responsiveness to error	15%
Alignment with verified course or domain material	15%

Table 1. Prespecified credibility-scoring rubric for intervention cues.

Source diversity refers to the breadth of information exposure across source types, viewpoints, topics, formats, or evidence traditions. Operationally, it is measured using entropy or concentration indices over a predefined exposure window. For example, if a participant’s recent listening and interaction history is

dominated by one source cluster, the diversity score decreases and the cooling rule may introduce credible alternatives.

Algorithmic cooling is the intervention mechanism that interrupts homogeneous exposure without deleting required learning material. Conceptually, it is a corrective recommendation process. Operationally, cooling occurs when a prespecified diversity threshold is crossed and the system introduces alternative sources, contrasting explanations, or reflection prompts. Every cooling event must be logged with its trigger value, content ID, timestamp, and participant study ID.

Epistemic confidence calibration is the primary learner outcome. Conceptually, it is the alignment between how confident a learner feels and how well justified that confidence is by evidence. Operationally, it is measured through repeated self-ratings, short claim-evaluation tasks, and comparison between confidence ratings and evidence-quality judgments. Improvement means not simply becoming more confident, but becoming more appropriately confident under strong evidence and more cautious under weak or incomplete evidence.

The primary calibration score is defined mathematically for participant i at occasion t as $C_{it} = 1 - |Conf_{it} - Evid_{it}|$, where $Conf_{it}$ is the participant's confidence rating rescaled from 0 to 1 and $Evid_{it}$ is the independently scored evidence-quality value for the same claim-evaluation task, also rescaled from 0 to 1. Higher values indicate better calibration, with 1 representing perfect alignment and 0 representing maximum mismatch. If multiple claim tasks are administered on the same occasion, C_{it} is the mean of all valid item-level calibration scores for that occasion. Secondary calibration indices may examine overconfidence and underconfidence separately using signed error, $Conf_{it} - Evid_{it}$, but the unsigned alignment score is the confirmatory primary endpoint.

Engagement quality is defined differently from engagement quantity. Quantity refers to minutes, clicks, reactions, and posts. Quality refers to behaviours that plausibly support learning, such as asking for evidence, comparing sources, revising a judgment, acknowledging uncertainty, or engaging respectfully with alternative explanations. This distinction prevents the study from mistaking platform activity for educational value.

Methods

Study Design

The study uses a two-week, two-arm, individually randomized, sham-controlled field design. Participants are allocated to either an epistemically governed platform or a visually matched sham platform. The individual learner is the unit of randomization and primary analysis. If multiple sites are used, randomization is blocked within site so that site characteristics do not become confounded with study arm[23][24].

The trial is best described as a pilot or early efficacy field study if conducted with approximately 60–120 participants. At this stage, the design should estimate feasibility, adherence, measurement quality, and effect direction rather than claim definitive population-level effectiveness[25]. A later multi-site trial can use the same protocol with larger samples and stronger power for site moderation.

The proposed range of 60–120 participants is justified as a pilot precision and feasibility target rather than as a definitive effectiveness sample. With approximately 30–60 participants per arm and repeated measurements across 14 days, the study should be able to estimate recruitment rate, attrition, EMA completion, variance components, and the direction and uncertainty of the arm-by-day effect. Planning assumptions for the subsequent confirmatory trial are: a small-to-moderate standardized arm-by-time effect of $d = 0.30–0.40$, within-person intraclass correlation of 0.30–0.50, 15%–20% attrition, and 80%–90% power for the primary endpoint. The pilot will not be interpreted as underpowered proof of no effect; instead, its variance estimates and feasibility outcomes will determine the final sample size for a larger trial.

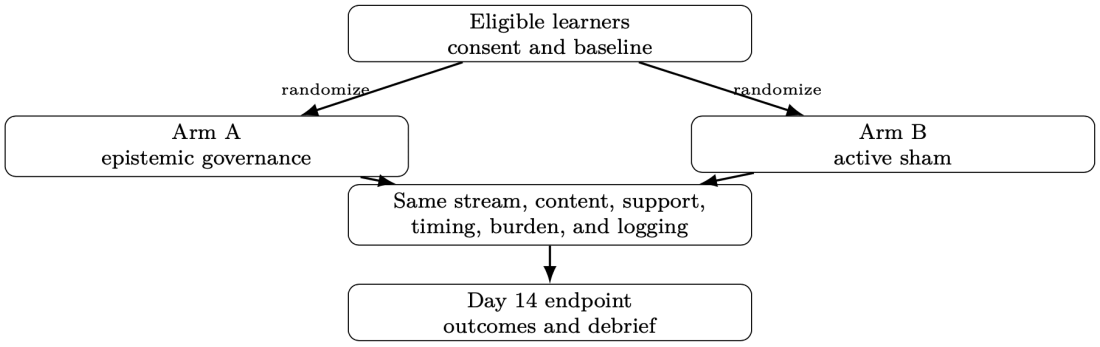


Figure 2. Randomized sham-controlled structure with equal exposure and arm-specific platform mechanisms.

Setting

The protocol is setting-neutral. It can be implemented in face-to-face classes, blended courses, fully remote programmes, continuing professional education, or community-learning workshops. A single instructor or presenter may deliver the same audio/video stream to all participants. Study-arm differences occur only inside the platform layer through feature flags, not through different teachers, different content, or different schedules.

Participants

Eligible participants are learners aged 18 years or older who are enrolled in a course, training programme, or structured learning activity and have access to a smartphone, tablet, or computer with internet connectivity. Exclusion criteria are inability to provide informed consent, inability to access the platform after technical support, or prior involvement in building or evaluating the platform. Participation must not affect course grades, employment evaluation, or access to normal learning materials.

Implementation Roles and Accessibility

The field experiment should define clear implementation roles before recruitment. The presenter delivers the same stream to both arms, the technical operator monitors platform stability and feature flags, the data manager protects randomization and export files, and the support contact uses neutral scripts for access problems. Role separation reduces the risk that staff behaviour becomes an unmeasured

intervention. Accessibility checks should include mobile access, low-bandwidth use, caption or transcript availability where feasible, and language adaptation for bilingual or Arabic-medium settings.

Interface and Readiness Requirements

The learner interface should include a stream or replay player, chat and reactions, content summaries, EMA prompts, credibility or sham cues, source-diversity or neutral feedback, help/report functions, and a brief debriefing page. Before field deployment, the team should conduct a dry run that verifies consent routing, study-ID login, randomization concealment, feature flags, stream stability, EMA delivery, data export, moderation rules, and the locked analysis script. Later changes must be entered into a protocol-deviation log.

Sample Size

For a pilot implementation, a sample of 60–120 participants is appropriate for estimating recruitment, adherence, missingness, technical feasibility, and preliminary effect sizes. The study should not be interpreted as a definitive efficacy trial unless powered in advance for the primary mixed-effects contrast. A subsequent confirmatory trial should use pilot-estimated variance, intraclass correlation, repeated-measure correlation, and attrition to determine sample size.

Randomization and Allocation Concealment

A statistician or data manager not involved in participant support should generate the randomization sequence before allocation. Randomization should be blocked by site and, where relevant, by course section or delivery mode. Allocation is implemented through private study IDs. Participants should not see arm labels. Staff who provide technical support should use neutral scripts and should not manually change scores, prompts, or feed order.

Intervention Arm: Epistemically Governed Platform

Participants in the intervention arm use the social media radio platform with active epistemic governance. The interface includes: (1) valid credibility indicators based on transparent scoring rules; (2) source-diversity feedback showing whether a learner's exposure is becoming narrow; (3) algorithmic cooling that introduces credible alternative sources when exposure becomes homogeneous; (4) EMA prompts that ask learners to reflect on confidence, evidence, uncertainty, and source quality; and (5) concise explanations of why selected content or prompts are shown.

Active Sham Arm

Participants in the sham arm use a visually similar version of the platform. The sham interface preserves time on platform, interface complexity, notification burden, livestream access, general activity indicators, and perceived plausibility. However, it does not provide valid credibility scoring, source-diversity governance, algorithmic cooling, epistemic EMA, or transparent recommendation explanations. This active sham design reduces the risk that any observed effect is due merely to novelty, attention, interface attractiveness, or additional contact[26].

Common Procedures Across Arms

Both arms receive the same onboarding, consent process, livestream schedule, content library, reminder schedule, moderation policy, technical support, and debriefing. Participants may use the platform during scheduled learning sessions and independently outside scheduled sessions. This hybrid structure increases ecological validity while preserving standardized measurement windows. EMA is appropriate because it captures repeated within-person states close to the moment of learning and interaction[27][28].

Remote Researcher Conduction

The study can be conducted by a researcher who is physically remote from participants. In this configuration, the researcher and supervisors operate from laptops using a fixed webcam, microphone, slide deck, run-of-show script, and dashboard. Participants may join from classrooms, homes, training rooms, or mixed locations. The remote researcher delivers one master stream to both arms, while the platform routes arm-specific cues, prompts, dashboards, and chat spaces through study-ID authentication. The stream is a standardized delivery channel, not the treatment.

Remote conduction requires additional safeguards. The researcher should not manually praise, message, or support one arm differently from the other. Supervisors may observe, monitor fidelity, or assist with technical checks, but unscripted intervention must be logged. A remote session checklist should include microphone test, camera framing, recording status, slide version, stream key, dashboard access, randomization lock, support-channel readiness, and post-session data export. If a live stream fails, the fallback recording must be identical across arms, and the outage must be included in fidelity and sensitivity analyses.

Pre-Implementation Readiness

Before field deployment, the research team should complete a dry run that is independent of the main sample. The dry run should verify informed-consent routing, study-ID login, randomization concealment, arm-specific feature flags, livestream stability, EMA delivery, data export, server monitoring, and debriefing. The platform version, prompt bank, credibility-scoring rubric, moderation rules, content schedule, survey instruments, randomization file, and analysis script should be frozen before the first participant is enrolled. Any later change must be entered into a protocol-deviation log with date, reason, affected participants, and expected risk to inference.

Recruitment materials should describe the study as an evaluation of social audio learning environments, not as a comparison of a superior and inferior platform. Onboarding should use the same script for both arms. Participants should be told how to use the player, chat, reactions, reporting tools, EMA prompts, and helpdesk channel, but the specific epistemic-governance hypothesis should be withheld until debriefing if local ethics approval permits. Technical support should solve access problems without explaining arm-specific mechanisms.

Operational Workflow

The study should run through a daily operations checklist. Before each scheduled stream, the operator verifies server health, active feature flags, stream key, content manifest, prompt schedule, database write status, and dashboard access. During each stream, the operator monitors concurrent users, reconnects, latency, failed prompt delivery, moderation flags, and helpdesk requests. After each stream, the operator exports a session-health report, confirms that raw events are append-only, records deviations, and checks whether each arm received comparable exposure. Unscheduled independent use remains open during the two-week period, but all automated prompts, reminders, and data logging must follow the locked schedule.

The minimum implementation team consists of a principal investigator, an instructor or presenter, a data manager, a technical operator, and a participant-support contact. In small deployments one person may hold more than one role, but the person who changes server or feature settings should not manually alter participant outcomes. In larger deployments, each site should have a local coordinator who follows the same checklist and escalates issues to the central operator.

Before implementation, the team should complete three rehearsals. The first is a technical rehearsal with test accounts to confirm login, streaming, telemetry, EMA delivery, server load, and backup export. The second is a procedural rehearsal in which staff read the onboarding script, simulate participant questions, practice neutral helpdesk responses, and verify debriefing language. The third is a data rehearsal in which dummy data are exported, cleaned, merged, and passed through the planned analysis script. A study should not begin until the team can show that a participant's path from consent to debriefing is fully traceable without relying on undocumented manual decisions.

Execution Routines

The daily execution routine should be written as a reproducible operations script. At T minus 60 minutes, the operator checks server health, stream readiness, database write status, queue length, storage, backups, and dashboard access. At T minus 30 minutes, the presenter checks microphone, camera, slides, countdown clock, and recording. At T minus 10 minutes, the helpdesk opens the support channel and confirms neutral scripts. During the session, the operator marks stream start, tracks attendance, logs incidents, and avoids changing arm logic. At session end, the operator marks stream stop, confirms file recording, exports a session-health summary, checks EMA delivery, and records unresolved issues.

The weekly execution routine should include data reconciliation, missingness review, protocol-deviation review, moderation audit, backup restoration test, and a blinded fidelity summary. The end-of-study routine should lock the database, export raw data, generate a hash report, create the analysis dataset, run the prespecified scripts, create a participant-flow report, and archive all app-version, prompt-version, and rubric-version files. These routines make the field experiment executable by a future team rather than dependent on the memory of the original researcher.

Platform and Data Requirements

The platform must be auditable enough to support a randomized media-literacy trial without becoming a black-box application. Minimum modules include study-ID authentication, consent status, feature flags, live or recorded audio delivery, chat and reactions, EMA scheduling, recommendation logic, moderation tools, telemetry, and export. The intervention feature bundle activates credibility indicators, source-diversity feedback, cooling rules, reflection prompts, and recommendation explanations. The sham feature bundle preserves the same interface complexity and notification burden but replaces epistemic cues with neutral activity or usability cues.

The credibility score is generated before exposure from the locked rubric in Table 1. Source diversity is calculated over a prespecified recent exposure window using entropy or concentration indices across source clusters. Cooling occurs only when the diversity threshold is crossed; it introduces credible alternatives or reflection prompts without deleting required instructional material. Every credibility cue, diversity score, cooling event, prompt delivery, and recommendation explanation must be logged with study ID, timestamp, content ID, feature flag, and server receipt time.

The minimum replication dataset includes participant, assignment, consent, event, stream, content, EMA, moderation, technical-log, helpdesk, and protocol-deviation tables. Names and contact details should be stored separately from analysis data. Missing values should distinguish non-delivery, delivery without response, participant skip, technical failure, withdrawal, and administrative exclusion. Data exports should include a codebook, version number, analysis script, and audit trail so that reviewers can distinguish intervention failure from implementation failure.

Study Timeline

Preparation includes ethics approval or explicit ethics-review status, protocol registration or time-stamping, platform testing, finalization of the content schedule, freezing of the EMA prompt bank, and generation of the randomization sequence. Day 0 includes consent, eligibility confirmation, baseline measures, device checks, and orientation. Days 1–5 establish exposure and adherence. Days 6–10 emphasize mechanism exposure through credibility cues, diversity feedback, cooling, and matched sham prompts. Days 11–14 complete final EMA prompts, posttest measures, manipulation checks, usability items, trust measures, debriefing, and data export.

Measures, EMA Design, and Example Instruments

Primary Outcome

The primary outcome is change in epistemic confidence calibration from baseline to Day 14. This construct refers to whether learners become more appropriately confident when evidence is strong and more appropriately cautious when evidence is weak or incomplete. The confirmatory score is $C_{it} = 1 - |Conf_{it} - Evid_{it}|$, where confidence and independently scored evidence quality are each rescaled from 0 to 1. Higher values indicate better alignment between confidence and evidence.

A baseline calibration task should present three to five short claims related to the learning domain. For each claim, participants rate confidence, perceived evidence strength, and willingness to share the claim with peers. Parallel items are administered at the endpoint using equivalent but not identical content. The goal is not to test memorization of a specific broadcast, but to evaluate whether learners improve in judging claims under uncertainty.

Secondary Outcomes

Secondary outcomes include source diversity, filter-bubble awareness, reflective evaluation, perceived credibility, knowledge reflection, news and media literacy, and engagement quality. Source diversity is operationalized using entropy of content categories, source types, or viewpoint clusters. Media literacy is included because communication research distinguishes access and activity from the ability to analyze, evaluate, and create messages in context[17][18][19]. Engagement quality is distinguished from raw activity volume: constructive comments, evidence-seeking behaviour, revisiting alternative sources, and reflective responses are treated differently from clicks or reactions alone.

Secondary outcomes should be interpreted as mechanism evidence rather than as substitutes for the primary outcome. For example, a participant may become more active without becoming more epistemically reflective. Conversely, a participant may post less frequently but show better source comparison and more cautious confidence. The analysis should therefore report both activity volume and engagement quality.

Ecological Momentary Assessment

EMA is a repeated, low-burden measurement strategy used to capture participant states close to the moment in which they occur. In this study, EMA is used because social media radio experiences are dynamic: learners may shift confidence, trust, attention, and perceived diversity during or shortly after exposure. EMA and experience-sampling methods are appropriate when researchers need repeated observations of naturalistic states, contexts, and behaviours rather than a single retrospective summary[27][29][30][31]. EMA reduces recall bias compared with asking participants at the end of two weeks to summarize all experiences from memory[28].

EMA prompts should be brief and scheduled to avoid fatigue. A feasible design is two prompts per day during active study days: one after a scheduled stream or major listening session, and one evening reflection prompt. Each EMA should contain no more than four to six items and should take less than one

minute to complete. The system should record scheduled time, delivery time, response time, non-response, skip, device type, and whether the prompt followed an intervention or sham cue.

Example EMA items use a five-point scale from strongly disagree to strongly agree unless otherwise stated. Credibility monitoring: “The content I just heard was supported by clear evidence.” Source awareness: “I noticed whether the speaker or source had relevant expertise.” Diversity awareness: “The platform exposed me to more than one perspective on the topic.” Confidence calibration: “My confidence in my current judgment matches the strength of the evidence I saw or heard.” Uncertainty regulation: “I can identify what I still do not know about this topic.” Cognitive load: “The stream and platform features felt mentally overwhelming.” Trust: “The platform helped me evaluate information rather than simply react to it.”

The sham arm receives equal-burden EMA prompts that do not activate the epistemic mechanism. Example sham prompts include: “The audio was easy to hear,” “The interface was easy to navigate,” “The notification arrived at a convenient time,” and “The session length felt manageable.” These items preserve measurement burden and attention without prompting credibility evaluation, source comparison, or uncertainty monitoring.

Additional intervention EMA items can be organized by construct. Evidence evaluation items include: “The speaker explained why the claim should be trusted,” “I could identify the evidence behind the main claim,” and “I would need more information before sharing this claim.” Source-diversity items include: “I encountered a source or viewpoint that differed from my first impression,” “The platform made me aware of what kinds of sources I had not heard,” and “I compared at least two sources before forming a judgment.” Confidence-calibration items include: “I feel confident, but I can explain the reason for my confidence,” “My confidence decreased when I noticed weaker evidence,” and “I changed my judgment after seeing a stronger source.” Social influence items include: “Peer reactions affected how credible the content felt,” “A popular comment made me more likely to trust the content,” and “I noticed a difference between popularity and evidence quality.”

EMA validation should occur before the main field period. First, content validity should be checked by experts in communication, educational technology, and measurement. Second, cognitive interviews with five to ten learners should test whether items are understood as intended. Third, a micro-pilot should estimate completion time, skipped items, response variance, and fatigue. Fourth, item analysis should remove items with very low variance, ambiguous wording, or high missingness. Fifth, the final pilot should test reliability, within-person variability, between-person variability, and convergence with

behavioural logs. This validation sequence prevents EMA from becoming a set of attractive but unvalidated pop-up questions.

Pretest and Posttest Survey Examples

The pretest should measure media habits, baseline media literacy, confidence in evaluating online claims, familiarity with social audio, and trust in algorithmic recommendations. Example items include: “I can usually tell when an online source is using evidence rather than opinion,” “I check who produced a media message before trusting it,” “I am comfortable changing my view when stronger evidence appears,” and “Recommendation systems usually show me a balanced range of information.”

The posttest repeats core constructs and adds manipulation checks. Example manipulation-check items include: “The platform showed me cues about source credibility,” “The platform encouraged me to compare different sources,” “The platform mainly showed general activity or usability information,” and “I understood why some content was recommended.” The expected pattern is that intervention participants report stronger exposure to epistemic cues, while sham participants perceive a plausible platform experience without the same epistemic guidance.

Psychometric and Validity Evidence

EMA items assess perceived credibility, confidence, uncertainty, evidence quality, perceived diversity, cognitive load, trust, and intention to continue learning. The measurement model will be evaluated using reliability coefficients, factor-structure evidence, item missingness, response-time patterns, and convergence with behavioural logs. Internal consistency alone is not sufficient; validity evidence must include construct definition, response processes, internal structure, and relations with external variables[32][33][34].

Implementation and Safety Measures

Implementation measures include login frequency, exposure minutes, EMA completion, prompt delivery, failed notifications, stream quality, device type, latency, helpdesk contact, adverse events, moderation flags, and contamination checks. These variables are not secondary outcomes by themselves; they indicate whether the causal comparison was delivered fairly.

A complete measurement packet should contain baseline covariates, repeated mechanism measures, endpoint outcomes, and operational fidelity indicators. Baseline covariates include age band, study level,

prior experience with social audio, prior use of reputation systems, baseline media literacy, and self-rated confidence in evaluating online claims. Repeated mechanism measures capture immediate confidence, perceived credibility, perceived diversity, uncertainty, and cognitive effort after platform exposure. Endpoint measures summarize confidence calibration, source evaluation, trust, acceptability, and perceived learning value. Operational indicators are retained so that an apparently weak treatment effect can be interpreted in light of exposure, fidelity, missingness, and technical disruption rather than treated as a simple failure of the theory.

Data Management and Quality Assurance

All data should be identified by study IDs rather than names. Raw logs should be append-only, time-stamped, and preserved separately from cleaned analysis files. The analysis dataset should include variable definitions, missing-data codes, feature-flag status, and version information. A data dictionary, protocol-deviation log, and analysis script should be locked before outcome analysis. FAIR principles should guide de-identified data sharing where ethics approval and privacy protections permit it[35].

Implementation fidelity is assessed through exposure balance, feature delivery, prompt timing, support equivalence, moderation equivalence, and technical stability. Fidelity is treated as measurable evidence, not a vague assurance[36][37].

Pilot success criteria should be prespecified before recruitment. The protocol will be considered feasible for a larger trial if recruitment reaches at least 80% of the target window, attrition is below 20%, median EMA completion is at least 80%, scheduled streams or recordings are available for at least 95% of planned exposure windows, arm-specific feature delivery is at least 90% concordant with assignment, and serious privacy or safety incidents are absent. Criteria that are not met should trigger protocol revision rather than selective reinterpretation of outcomes.

Statistical Analysis Plan

The primary analysis will use a linear mixed-effects model with repeated observations nested within participants. The model includes fixed effects for day, arm, and the day-by-arm interaction, with participant random intercepts. If multiple sites are used, site indicators or random effects will be included depending on sample size and design. Mixed models are appropriate because they account for within-

person dependence and can use incomplete repeated measures under defensible missing-data assumptions[38][39].

For the primary outcome, the key estimate is the arm-by-day interaction, interpreted as the difference in longitudinal change between the epistemic-governance and sham arms. Secondary models will test source diversity, filter-bubble awareness, reflective evaluation, and engagement quality. Mechanism analyses will test whether exposure to credibility indicators, cooling events, and diversity feedback is associated with changes in epistemic outcomes within the intervention arm.

Psychometric analyses will report coefficient omega or related reliability estimates, exploratory or confirmatory factor evidence where sample size permits, item response patterns, and item-level missingness[40][41]. Missing data will be described by arm, site, day, and technical context. Sensitivity analyses will compare maximum likelihood, multiple imputation, and complete-case results where appropriate[42][43].

Because H1–H4 represent related but distinct confirmatory tests, the analysis will control familywise or false-discovery error using a prespecified correction strategy[44][45]. Exploratory analyses will be clearly labelled and not presented as confirmatory evidence.

The preregistered analysis file should specify variable names, exclusion rules, coding of missingness, model formulas, convergence remedies, multiplicity correction, and the exact order in which sensitivity analyses will be run. The frozen materials should be deposited in a time-stamped repository such as OSF, an institutional repository, or a journal-approved registry, and the manuscript should report the protocol version and registration date once available.

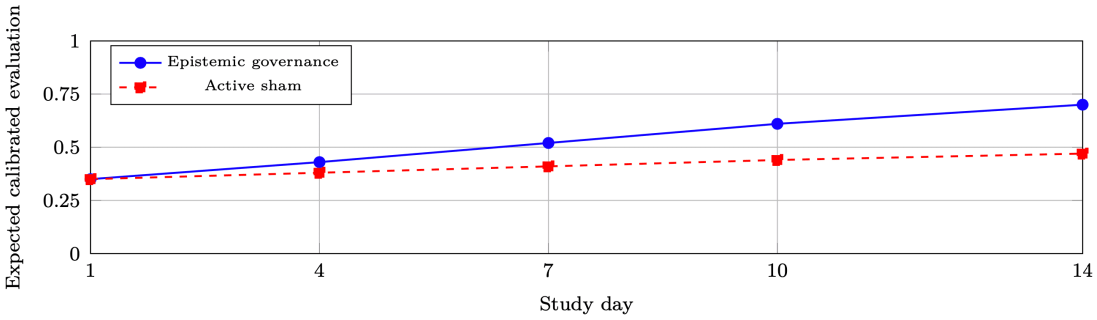


Figure 3. Hypothesized trajectory expected under the primary arm-by-day mixed-effects contrast. Values are illustrative only and do not represent empirical participant data.

Threat	Mitigation
Contamination between arms	Study-ID authentication, feature flags, separate arm-specific cues, and contamination checks
Hawthorne and novelty effects	Active sham with matched interface burden, notifications, support, and exposure
Attrition and missing EMA	Low-burden prompts, missingness codes, mixed models, and sensitivity analyses
Social desirability	Repeated EMA, behavioural logs, and neutral item wording
Technical failure	Dry runs, backup recordings, stream-health logs, and deviation reporting
Instructor or support bias	Common scripts, role separation, helpdesk logs, and fidelity monitoring

Table 2. Major threats to validity and planned mitigations.

Ethical Considerations

Ethics approval is required before recruitment. For submission, this section should state the actual review status rather than remain ambiguous: either ethics approval has been granted, with committee name, approval number, and approval date, or the manuscript is being submitted explicitly as a protocol awaiting review and no recruitment will begin until approval is obtained. Consent materials must explain randomization, passive logging, EMA prompts, algorithmic scoring, livestream participation, privacy protections, withdrawal rights, and debriefing. Participants should be told that they may receive different platform features but should not be told the specific hypothesis until debriefing if local ethics approval permits. Data should be minimized, encrypted where feasible, and reported only in aggregated or de-identified form. Participants in the sham arm may be offered access to the epistemically governed version after the study if the intervention appears beneficial and safe.

The protocol and analysis plan should be preregistered or time-stamped before data analysis. Preregistration helps distinguish confirmatory hypotheses from exploratory findings and supports transparent reporting[46][47].

Reporting Standards

Results should follow CONSORT guidance for pilot, feasibility, and pragmatic trials[7][48]. Reports should include participant flow, baseline comparability, exposure balance, missingness, manipulation checks, fidelity, adverse events, outcomes, psychometric evidence, and sensitivity analyses. Technical success and engagement should be reported as implementation evidence, not proof of learning effectiveness.

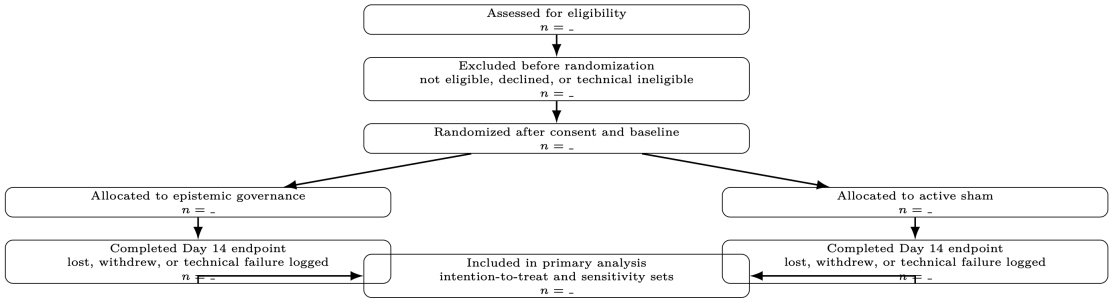


Figure 4. CONSORT-style participant flow template for the completed pilot report. Counts should be filled only after recruitment and follow-up are complete.

Discussion

This study is designed to answer a focused empirical question: whether epistemic governance features in a social media radio platform improve learning-relevant information evaluation beyond a plausible sham interface. The active sham design is essential because social platforms often produce novelty, attention, and social-presence effects that can be mistaken for deeper learning. By keeping content, presenter exposure, interface burden, reminders, support, and measurement equal across arms, the study isolates the epistemic layer as the intended causal mechanism.

The design also recognizes that educational technology trials require both internal and ecological validity. A purely laboratory design may control exposure but miss natural use. A purely observational deployment may capture authentic behaviour but cannot support strong causal inference. The proposed hybrid field trial balances these needs by standardizing onboarding and assessment while allowing natural platform use during the two-week period.

Potential limitations include self-selection into the study, short follow-up, incomplete blinding, EMA fatigue, technical variability, and limited power in a pilot sample. These limitations should be handled

through transparent reporting, manipulation checks, missing-data diagnostics, and cautious interpretation. A positive pilot would justify a larger multi-site trial; a null or mixed pilot would still be informative if it identifies which mechanisms, measures, or implementation conditions require refinement.

Conclusion

A social media radio platform can be studied rigorously only if the platform's educational claims are separated from its novelty, usability, and infrastructure performance. This study provides a generic empirical design for testing whether epistemic governance features, including valid credibility cues, source-diversity feedback, algorithmic cooling, reflective EMA, and transparent recommendations, improve students' confidence calibration, source diversity, and reflective evaluation. The proposed randomized sham-controlled field trial is adaptable across institutions and learning contexts while preserving the causal contrast needed for scientific validity. If implemented with concealed randomization, equal exposure, strong fidelity monitoring, preregistered analysis, and transparent reporting, the study can produce credible evidence about when and how social media radio supports better learning and information evaluation.

Protocol Status and Data Availability

This manuscript is a prospective protocol and does not report participant-level outcome data because recruitment has not yet begun. No simulated findings are presented. After implementation, de-identified data, codebooks, and analysis scripts should be shared only where permitted by ethics approval, consent, institutional policy, and privacy law. For blind review, author-identifying acknowledgements, funding details, and institutional information have been removed and should be restored only after peer review if the journal requests them.

References

1. [a], [b] *McCombs ME, Shaw DL (1972). "The Agenda-Setting Function of Mass Media." Public Opinion Q. **36**(2):176–187.*
2. [a], [b] *Katz E, Blumler JG, Gurevitch M (1973). "Uses and Gratifications Research." Public Opinion Q. **37**(4):509–523.*

3. ^a, ^b*Lazarsfeld PF, Berelson B, Gaudet H (1948). The People's Choice: How the Voter Makes Up His Mind in a P residential Campaign. 2nd ed. Columbia University Press.*
4. ^a, ^b*Rogers EM (2003). Diffusion of Innovations. 5th ed. Free Press.*
5. ^*Campbell DT, Stanley JC (1963). Experimental and Quasi-Experimental Designs for Research. Houghton Mi fflin.*
6. ^*Shadish WR, Cook TD, Campbell DT (2002). Experimental and Quasi-Experimental Designs for Generalize d Causal Inference. Houghton Mifflin.*
7. ^a, ^b*Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, Lancaster GA (2016). "CONSORT 2010 Statement: Extension to Randomised Pilot and Feasibility Trials." BMJ. **355**:i5239.*
8. ^*Cronbach LJ, Meehl PE (1955). "Construct Validity in Psychological Tests." Psychol Bull. **52**(4):281–302.*
9. ^*Shoemaker PJ, Vos TP (2009). Gatekeeping Theory. Routledge.*
10. ^*Hovland CI, Weiss W (1951). "The Influence of Source Credibility on Communication Effectiveness." Public Opinion Q. **15**(4):635–650.*
11. ^*Metzger MJ, Flanagin AJ, Medders RB (2010). "Social and Heuristic Approaches to Credibility Evaluation O nline." J Commun. **60**(3):413–439.*
12. ^*Sundar SS (2008). "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credi bility." In Metzger MJ, Flanagin AJ (Eds.), Digital Media, Youth, and Credibility. pp. 73–100. MIT Press.*
13. ^*Tandoc EC Jr, Lim ZW, Ling R (2018). "Defining "Fake News": A Typology of Scholarly Definitions." Digit J. **6** (2):137–153.*
14. ^*Klapper JT (1960). The Effects of Mass Communication. Free Press.*
15. ^*Knobloch-Westerwick S (2015). Choice and Preference in Media Use: Advances in Selective Exposure Theor y and Research. Routledge.*
16. ^*Messing S, Westwood SJ (2014). "Selective Exposure in the Age of Social Media: Endorsements Trump Parti san Source Affiliation When Selecting News Online." Commun Res. **41**(8):1042–1063.*
17. ^a, ^b*Aufderheide P (1993). "Media Literacy: A Report of the National Leadership Conference on Media Literac y." Aspen Institute.*
18. ^a, ^b*Livingstone S (2004). "Media Literacy and the Challenge of New Information and Communication Tech nologies." Commun Rev. **7**(1):3–14.*
19. ^a, ^b*Maksl A, Ashley S, Craft S (2015). "Measuring News Media Literacy." J Media Lit Educ. **6**(3):29–45.*

20. ^Garrison DR, Anderson T, Archer W (2000). "Critical Inquiry in a Text-Based Environment." Internet High Educ. 2(2–3):87–105.

21. ^Mayer RE (2021). Multimedia Learning. 3rd ed. Cambridge University Press.

22. ^Leppink J, Paas F, Van der Vleuten CPM, Van Gog T, Van Merriënboer JJG (2013). "Development of an Instrument for Measuring Different Types of Cognitive Load." Adv Health Sci Educ. 18:1058–1072.

23. ^Schulz KF, Grimes DA (2002). "Generation of Allocation Sequences in Randomised Trials: Chance, Not Choice." Lancet. 359(9305):515–519.

24. ^Suresh KP (2011). "An Overview of Randomization Techniques: An Unbiased Assessment of Outcome in Clinical Research." J Hum Reprod Sci. 4(1):8–11.

25. ^Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, Robson R, Thabane M, Giangregorio L, Goldsmith CH (2010). "A Tutorial on Pilot Studies: The What, Why and How." BMC Med Res Methodol. 10:1.

26. ^Freedland KE, Mohr DC, Davidson KW, Schwartz JE (2011). "Usual and Unusual Care: Existing Practice Control Groups in Randomized Controlled Trials of Behavioral Interventions." Psychosom Med. 73(4):323–335.

27. a, bShiffman S, Stone AA, Hufford MR (2008). "Ecological Momentary Assessment." Annu Rev Clin Psychol. 4:1–32.

28. a, bBolger N, Laurenceau J-P (2013). Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research. Guilford Press.

29. ^Csikszentmihalyi M, Larson R (1987). "Validity and Reliability of the Experience-Sampling Method." J Nerv Ment Dis. 175(9):526–536.

30. ^Stone AA, Shiffman S (1994). "Ecological Momentary Assessment in Behavioral Medicine." Ann Behav Med. 16(3):199–202.

31. ^Trull TJ, Ebner-Priemer UW (2013). "Ambulatory Assessment." Annu Rev Clin Psychol. 9:151–176.

32. ^Messick S (1995). "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." Am Psychol. 50(9):741–749.

33. ^Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL (2018). "Best Practices for Developing and Validating Scales." Front Public Health. 6:149.

34. ^McDonald RP (1999). Test Theory: A Unified Treatment. Lawrence Erlbaum Associates.

35. ^Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Mons B (2016). "The FAIR Guiding Principles." Sci Data. 3:160018.

36. ^*Dane AV, Schneider BH (1998). "Program Integrity in Primary and Early Secondary Prevention." Clin Psychol Rev. **18**(1):23–45.*

37. ^*Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S (2007). "A Conceptual Framework for Implementation Fidelity." Implement Sci. **2**:40.*

38. ^*Laird NM, Ware JH (1982). "Random-Effects Models for Longitudinal Data." Biometrics. **38**(4):963–974.*

39. ^*Raudenbush SW, Bryk AS (2002). Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd ed. SAGE.*

40. ^*Raykov T (1997). "Estimation of Composite Reliability for Congeneric Measures." Appl Psychol Meas. **21**(2):173–184.*

41. ^*Embretson SE, Reise SP (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates.*

42. ^*Little RJA, Rubin DB (2019). Statistical Analysis with Missing Data. 3rd ed. Wiley.*

43. ^*Enders CK (2010). Applied Missing Data Analysis. Guilford Press.*

44. ^*Bonferroni CE (1936). "Teoria Statistica Delle Classi E Calcolo Delle Probabilità" [Statistical Theory of Classes and Probability Calculus]. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. **8**:3–62.*

45. ^*Benjamini Y, Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." J R Stat Soc B. **57**(1):289–300.*

46. ^*Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018). "The Preregistration Revolution." Proc Natl Acad Sci USA. **115**(11):2600–2606.*

47. ^*Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, Glennerster R, Green DP, Humphreys M, Imbens G, Laitin D, Madon T, Nelson L, Nosek BA, Petersen M, Sedlmayr R, Simmons JP, Simonsohn U, Van der Laan M (2014). "Promoting Transparency in Social Science Research." Science. **343**(6166):30–31.*

48. ^*Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D (2008). "Improving the Reporting of Pragmatic Trials: An Extension of the CONSORT Statement." BMJ. **337**:a2390.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.