A08 | PRIMARY-STUDY PROTOCOL AND REPRODUCIBLE METHODS RESEARCH

Accountable Intelligence at the Point of Decision: A Europe PMC Corpus Audit and Randomized Protocol for Evaluating Paper Agents in School-District Evidence Work

Dr. Freedom Cheteni | Human author

ORCID 0009-0006-3507-0094

Prepared 2026-10-16 | Edition 1.1 | 4,191 article words

Primary-study protocol and reproducible methods; separate human editorial record.

Keywords: Movement Thinking, paper agents, school district governance, evidence fidelity, Europe PMC, blinded assessment, randomized protocol, provenance

Read online / article companion / citation exports

Contribution

An original human decision-support trial protocol, exact-version Europe PMC corpus audit and reproducible artificial-data design analysis that distinguish source availability from warranted educational judgment.

Prospective adult randomized benchmark; completed 36-record core/JATS source-feasibility audit and nine-scenario 10,000-replicate artificial-data design analysis. No participant/agent performance or educational effects reported.

Abstract

A paper agent changes how a school leader encounters research, but convenient access does not establish better judgment. This article develops a prospective randomized human study of an accountable paper-agent interface and reports two completed methodological analyses. The interface is evaluated against a source-readable active control using identical evidence packets, with human evidence-fidelity rather than agreement with an agent as the primary outcome. A bounded Europe PMC audit retrieves the first 36 records from a 98-hit leadership query, checks full-text JATS, article-level rights, identifiers and available status signals, and preserves source hashes. Twenty-three records pass the declared source-construction rules; thirteen remain excluded or unresolved. An indexed DOI and the downloaded JATS identify different F1000Research versions, illustrating why citation resolution alone is insufficient. The proposed study recruits 160 adult education professionals, randomizes within four strata, and administers twelve fictional district decisions. Blinded human raters score causal restraint, population and measure fidelity, citation entailment, uncertainty, rights and version interpretation, and bounded action. Artificial-data simulations show that design sensitivity depends strongly on within-person case correlation. No participant was recruited, no human or agent performance outcome was observed, and no educational or clinical benefit is inferred. The contribution is a reproducible design that makes accountable intelligence falsifiable at the point where a source becomes a decision.

1. From an intelligent library to accountable evidence work

School districts increasingly encounter research through search engines, summaries and conversational interfaces rather than through a linear reading of journal issues. An agent attached to a paper can help a superintendent find a methods passage or compare a claim with a limitation. The same interface can erase an important distinction: an observational association can become an intervention promise, a self-report measure can become an observed student outcome, and a readable paper can become a supposedly reusable asset. Institutional seriousness therefore requires evaluation of what a human can justify after using the interface, not simply whether an answer sounds scholarly.

Movement Thinking provides the intellectual starting point: educational work should remain open to revision, participation and changing conditions [1]. This study translates that orientation into a narrow governance hypothesis. A source-grounded interface may help professionals move between question, evidence, objection and revised judgment without losing attribution or decision responsibility. That is an original operational proposition, not a claim that the earlier philosophy already specified or validated this

experiment. The companion conceptual article, “The District as Accountable Intelligence,” remains a separate scholarly object; this protocol adds an empirical route for testing one of its claims.

The unit of practical concern is the decision brief. It records what is known, what remains uncertain, which evidence supports the recommendation and who must decide. A correct DOI is necessary bibliographic infrastructure but insufficient warrant. A paper about general school leadership cannot, merely through an interactive interface, establish the correct response to an individual child’s mental-health needs. The present benchmark is deliberately limited to evidence interpretation in fictional district situations, with no actual student, personnel, disciplinary or clinical decision delegated to software.

2. A falsifiable theory of accountable assistance

The proposed mechanism is reduced friction in inspecting evidence. The treatment makes the source passage, version and applicability limit reachable from an answer, then allows a participant to revise an initial brief. The alternative mechanism is automation deference: fluent language may increase confidence while weakening inspection. Both are plausible. The protocol therefore measures warranted reasoning, unjustified confidence and verification time separately. More clicks, more citations and higher satisfaction are not interchangeable indicators of better judgment.

The hypothesis is that assignment to the accountable interface improves the mean evidence-fidelity score across a fixed set of cases compared with an equally readable source-only interface. The causal target is the incremental effect of this specific interface, training and frozen evidence configuration on recruited adults completing these cases. It is not the effect of every AI model, a district-wide implementation programme, future employment, student attainment or clinical care. A positive benchmark result would justify a further field study rather than a claim of educational effectiveness.

The causal chain is assignment, inspection opportunity, human interpretation and decision brief. Intermediate interface events are measured as process variables. They are not randomized mediators and will not be assigned causal mediation interpretations merely because they correlate with the final score. A participant can reach the right decision without opening the suggested passage, or open it without understanding it. The final human response, assessed against independent criteria, remains the endpoint.

3. Why Europe PMC is an appropriate but bounded database

Europe PMC supplies a documented API for core publication metadata and open-access full-text XML [2]. Its open-access documentation expressly directs users to each article's terms; membership of the subset does not provide a uniform reuse license [3]. These features make it useful for testing the institutional problem at issue. A record contains multiple kinds of evidence - bibliographic identity, accessible text, permissions and linked publication-status information - that can agree or conflict.

The database is particularly useful for school mental health, implementation and wellbeing. It is not a complete index of education policy or school-administration scholarship. The audit is consequently a feasibility investigation for constructing a benchmark, not an exhaustive review of leadership research. The query, response, ordering, cut-off and exclusions are preserved so that selection cannot later be described as comprehensive. Subsequent trial construction requires a separate, independently reviewed case set; the present audit establishes a candidate pool rather than a validated reference standard.

The study does not merge Europe PMC records with TALIS respondents, SEDA districts or student files. Published studies may themselves use such databases, but this investigation treats those papers as scholarly sources with their original designs. The original researchers retain their authorship. Dr. Freedom Cheteni's contribution is the new evaluation protocol, corpus audit, instruments and analysis specification, not the empirical work reported in imported articles.

4. Completed corpus audit: retrieval and screening rules

The executed search combines the title/abstract phrases "school leadership" and "principal leadership," an initial publication-date window of 1 January 2017 through 15 October 2026, and Europe PMC's open-access and full-text indicators. It excludes nursing, medical-school and medical-education terms. The exact machine-readable expression and encoded API URL accompany the methods package. The response contains 98 hits. The audit intentionally stops at its first 36 relevance-ranked records, preserving the response hash and all selected identifiers. It does not inspect the remaining 62 hits or imply that their eligibility matches the audited subset.

For every selected record, the script retrieves JATS through the supported full-text API, computes SHA-256 over the received bytes, parses author contributions separately from editorial groups, and checks the article's own permissions. It records publication dates as supplied, including incomplete date components. Europe PMC's normalized

first-publication date is retained as an indexed field, not silently promoted into an exact version-release day. Core DOI and JATS article DOI are compared, while a PMCID plus byte hash identifies the precise retrieved snapshot.

A source passes construction rules only when the JATS identifies a research article, an article-level CC BY, CC BY-SA or CC0 version is identifiable, a nonempty body is present, the indexed DOI agrees with a JATS article DOI, no screened withdrawal or retraction signal is detected, and the topic fits the declared decision-support boundary. A data-only CC0 waiver cannot license the article. NC/ND licenses, unversioned rights statements and unresolved version disagreement remain outside the construction pool. No third-party full text or figures are redistributed with this article.

Three relevance or design exclusions are declared in code: university physical-education trainee monitoring, tele-delivered caregiver coaching, and a policy-review context excluded from the primary-research benchmark pool. These are explicit scope choices rather than claims that the excluded work lacks value. Reviews can inform the article but do not enter its research-source construction pool. Human case developers must still inspect the included sources for question suitability, contested interpretations and any restrictions attaching to particular third-party materials.

5. Completed corpus results and their epistemic limits

All 36 selected JATS documents were retrieved and parsed; the snapshot contains 36 distinct PMCIDs and 36 distinct nonempty indexed DOIs. Thirty-four documents carry a research-article JATS type, one a review-article type and one a systematic-review type. Twenty-nine have an identifiable article-level eligible license URL, four carry NC or ND restrictions, and three contain an insufficiently versioned license statement for the strict construction rule. These are document classifications, not judgments about the quality of the underlying science.

Combining design, relevance, rights, identifier and available status rules leaves 23 source-eligible candidates and 13 excluded or unresolved records. The categories overlap: the version-conflicted F1000Research item has identifiable CC BY 4.0 rights, yet fails identifier agreement. Its indexed DOI ends in .2 while the retrieved JATS article DOI ends in .3 [12]. The discrepancy does not establish misconduct or an invalid publication; it establishes that a user must resolve the version before quoting or linking a supposedly exact snapshot. The JATS version history identifies version 3 as 8 October 2026 and version 2 as 19 August 2026; the audit retains both identifiers rather than overwriting one with the other. A separate date discrepancy appears for the whole-school implementation paper: the indexed field is 1 January 2026 while the downloaded JATS publication date is 3

December 2025 [11]. The reference uses the document date and the audit preserves both, without disguising the disagreement.

No retraction, withdrawal or expression-of-concern signal was observed in the screened core publication types, title prefixes, available correction relations and JATS related-article fields. This is a limited, dated observation, not an independent certification of integrity. A later notice, an unindexed notice or a publisher correction can change the appropriate interpretation. Status must be checked again before trial freeze and before consequential deployment. The corpus audit reports availability and provenance; it does not measure retrieval quality, agent answer accuracy, human decision improvement or causal educational effects.

6. Source designs that make demanding test cases possible

Several sources show why the benchmark must examine inference rather than citation decoration. Lyon and colleagues' School Implementation Leadership Scale concerns measurement development in school implementation settings [4]. An agent should identify the measure and its setting rather than treat it as certification of a software platform. Låftman, Östberg and Modin's multilevel cyberbullying analysis supplies a different challenge: a relationship between leadership and reported cyberbullying does not license a claim that deploying an agent prevents bullying [5].

Bao's principal-inclusive-leadership analysis provides another opportunity to distinguish a statistical pathway from an intervention mechanism [6]. The benchmark can ask whether a district may promise a reduction in teacher turnover from a mediation result without a randomized local evaluation. Daly, Resnikoff and Litke's narrative review informs school mental-health leadership but remains a review, not a trial of a particular district programme [7]. It is cited for design interpretation and excluded from the primary-research construction pool, not represented as a failed source.

The HELM pilot provides an important contrast because it actually randomized schools to leadership strategies, including an attention comparator [8]. The benchmark must preserve what was assigned, where, and what outcomes were measured. Its NC license also illustrates that stronger causal design and broader reuse permission are different properties. Ertem's crisis-leadership article has NC-ND rights in the retrieved version [9]. Both can be cited and linked here without being copied into a supposedly unrestricted agent corpus. No abbreviated reading of an "open access" badge substitutes for that distinction.

7. Prospective participants, setting and recruitment

The planning sample is 160 consenting adults whose professional roles involve interpreting research for schools: superintendents and district leaders, principals, teachers and education-support professionals. Four strata cross leadership role versus practitioner/support role with lower versus higher self-reported research-agent experience, defined through a prespecified screening threshold. Each stratum recruits 40 participants before assignment. These numbers are a design target, not an enrolled cohort or evidence of institutional agreements.

Recruitment occurs outside supervisory evaluation and admission decisions. Participation cannot affect employment, professional standing, course access or publication eligibility. Invitations explain that the tasks are fictional and that software can be incomplete or wrong. Compensation is fixed for time rather than performance or endorsement. Eligibility requires sufficient English reading proficiency for the fixed materials; this restricts generalizability and does not establish an inherent language requirement for school leadership. Accessible formats, keyboard operation and planned breaks must be available.

An independent ethics determination, consent materials, data-management agreement, intervention freeze and dated registration precede recruitment. None is claimed to exist for this proposal. No minors are recruited, no actual student records are entered, and no prompt may contain identifiable information about staff or families. A participant can stop without having an incomplete response converted into an allegation of incompetence. Individual performance is not reported to an employer.

8. Randomization, intervention and a credible comparator

A separate allocation custodian assigns 20 of the 40 participants in each stratum to each condition after screening and baseline measures are locked. The allocation is generated using a recorded random procedure and concealed from recruiters until enrolment is complete. Participants cannot be blinded to a conversational interface; outcome raters, case-standard assessors and the primary analyst remain masked to arm labels until scoring and the analysis plan are locked.

Both interfaces display the same source packet, full bibliographic attribution, readable licensed text where permitted, title navigation, ordinary text search, notice and rights panels, equivalent typography and a decision-brief form. Training time and task time are equal. The control therefore represents competent digital reading rather than an intentionally frustrating PDF. The treatment adds a bounded question interface with passage-linked responses, visible uncertainty, exact-source access and a clear separation

between source statements and interface interpretation. Responses cannot change the supplied evidence or submit a real district action.

Before the trial, the treatment manifest must record code revision, retrieval settings, model and version if any, prompt rules, corpus hashes and session limits. This article does not invent a deployed model identifier or claim that the proposed interface has already passed evaluation. If a deterministic retrieval companion is used, that configuration is recorded explicitly. An evaluation of generative assistance requires a separate frozen configuration, retained input/output logs under consent, and cost, privacy and reproducibility arrangements. Model changes during the confirmatory study require an amendment or a separately analyzed cohort, not silent substitution [13] [14].

9. Twelve cases and an independent reference standard

The fixed benchmark contains twelve short fictional district decisions: three requiring causal and design interpretation, three requiring population and measurement transfer, three requiring uncertainty and source-entailment judgments, and three requiring rights, version or notice interpretation. Every case receives the same six-dimensional scoring rubric, so a category label does not exempt it from the other obligations. Case order is balanced independently of treatment assignment. Participants receive approximately six minutes per case, with training, baseline questions and breaks bringing the session to roughly 100 minutes.

Independent education researchers and an information specialist develop the case reference standard before seeing any treatment-generated answer. They identify acceptable conclusions, alternative defensible answers, source locations and errors that would materially change a decision. They rate the standard independently, resolve disagreements with written reasons and retain unresolved interpretations as explicit ranges. The standard is not an agent’s preferred response and does not reward agreement with Cheteni’s philosophy. Pilot participants and pilot raters are distinct from the confirmatory cohort.

Some errors are better tested with transparent fictional evidence than with a false claim about a real researcher. If a retraction-notice case cannot be built from a verified actual notice, the packet uses a wholly fictional article and labels the entire notice exercise synthetic. It never attaches an invented retraction to a real DOI. Similarly, a fictional local budget or school cannot be mistaken for collected district data. Case release is delayed until scoring is complete to reduce coaching; the protocol, categories, rubric and source rules can be public earlier.

10. Primary outcome and blinded scoring

The primary outcome is each participant's mean evidence-fidelity score across twelve cases, on a 0-24 scale. This is a newly specified, unvalidated rubric: pilot content review, inter-rater reliability and ceiling checks are required and will not establish universal validity. Six dimensions contribute 0-4 points each: design and causal restraint; population, setting and measurement fidelity; citation entailment and source location; uncertainty and appropriate abstention; rights, version and notice interpretation; and a bounded, accountable recommendation. A high score requires a defensible path from evidence to action, not merely polished prose. Correctly stating that the packet cannot resolve a question can earn full credit.

Two trained human raters score every final brief independently. The primary case score is the mean of their total scores; the participant score is the unweighted mean across cases. Disagreement is reported dimension by dimension and as an absolute-agreement reliability estimate, with uncertainty. A third assessor reviews prespecified severe-error disagreements for quality assurance but does not selectively replace the primary mean with a favorable score. If an instrument defect is found, the revision is documented before confirmatory scoring resumes.

A secondary sentinel-error outcome counts unsupported causal guarantees, fabricated source claims, misattributed authorship, false claims of permission and acceptance of an actual withdrawn finding without qualification. Severity definitions and denominators are frozen. The sentinel rate is not blended into the primary score after results are known. Other secondary outcomes include completion time, source verification time, confidence calibration, accessibility barriers, workload and participant-rated usefulness. Faster completion is beneficial only alongside adequate fidelity and an acceptable severe-error profile.

11. Statistical estimand and primary analysis

Let each person's potential twelve-case mean under the source-only interface and under the accountable agent be defined for the frozen source and case set. The primary estimand is the equal-stratum mean difference between assigned conditions among the recruited adults, regardless of actual agent use. With equal planned stratum sizes this is also the equally weighted participant effect. It does not treat the 1,920 case responses as 1,920 independent people.

For complete cohorts, the estimator averages the four within-stratum treatment-minus-control differences. Its variance sums the eight independent arm/stratum

sample-variance terms, each divided by its cell size and by sixteen. Confidence intervals and the two-sided test use Welch-Satterthwaite degrees of freedom calculated from those terms. This permits unequal outcome variances rather than assigning an unjustified pooled residual degree of freedom. The downloadable tool implements the complete-cohort calculation and verifies actual-design cell counts; it refuses a silently altered design.

The primary analysis is assignment-based, with no exclusion for ignoring the agent, disagreeing with it or producing a low score. A secondary crossed-effects model includes participant and case intercepts, fixed treatment and stratum terms, and a treatment-by-case random component if identifiable. Its purpose is to describe heterogeneity and sensitivity to case selection. With only twelve purposively constructed cases, a random-case model cannot establish representativeness of all educational decisions. Domain and experience interactions are exploratory, with effect estimates and uncertainty rather than selective significance claims.

12. Missing data, multiplicity and change control

An unanswered case during an otherwise completed timed task receives zero only when the predeclared task definition treats failure to produce a brief as observed performance. A lost submission, technical failure, withdrawn record or interrupted session is missing, not automatically a zero. The distinction is logged before arm labels are revealed. The confirmatory analysis uses fifty multiple imputations of missing case totals from a bounded predictive model including assignment, stratum, baseline evidence-literacy score, case identity and observed case responses. The analysis specification defines diagnostics, pooling and sensitivity departures; the supplied small Python estimator does not claim to implement that imputation model.

Sensitivity analyses shift imputed treatment values downward and control values upward by one and two rubric points, then report bounded worst-case contrasts. The report provides missingness by arm, reason and case and distinguishes retention effects from interpretation effects. Pooling follows repeated-imputation rules with a small-sample degree-of-freedom adjustment rather than presenting imputed records as observed people [15]. Complete-case results remain secondary when missingness is present.

There is one confirmatory outcome and one primary comparison. Secondary families are grouped before analysis; sentinel error and workload are explicitly evaluated as safety-relevant descriptive outcomes, not hidden when the primary score improves. Any multiplicity-adjusted exploratory tests are identified. The registration locks outcomes, code revision, case-standard hash and missing-data rules before recruitment. Changes

needed for accessibility or safety are permitted but documented with timing and reasons, allowing readers to distinguish correction from outcome-driven redesign.

13. Completed design-sensitivity analysis

The executed simulation uses 10,000 artificial datasets for each combination of 80, 160 or 240 adults and within-person case correlations of 0.10, 0.30 or 0.50. There are four equal strata, balanced assignment within each stratum and twelve equally weighted cases. Marginal case variance is standardized to one. The imposed treatment effect is 0.25 case standard deviations. The model is Gaussian, assumes complete outcomes and constant treatment effects, and includes neither ceiling effects nor case-specific treatment heterogeneity. These assumptions are inputs, not facts estimated from district professionals.

For the 160-person planning design, simulated power is 96.55%, 74.56% and 57.01% as within-person correlation rises from 0.10 through 0.30 to 0.50. Monte Carlo standard errors are approximately 0.18, 0.44 and 0.50 percentage points. Corresponding null rejection rates are 4.62%, 5.46% and 5.09%. These values show why counting twelve responses per adult as independent observations would exaggerate precision. Increasing the number of cases yields diminishing returns when stable person-level differences dominate.

The results do not establish that the proposed agent will produce the assumed effect or that 160 participants guarantee adequate power. A pilot should estimate score dispersion, ceiling frequency, case correlation and completion feasibility without testing the final efficacy claim. The investigator then chooses a funded confirmatory design before recruitment, reruns sensitivity calculations and registers any change. The public simulation reports seed, software versions, every scenario and Monte Carlo uncertainty; no simulated row is presented as an observed human response.

14. Adversarial interpretation, workload and institutional consequences

An interface should not receive credit for a correct reference beside an unsupported conclusion. Qian and colleagues' Shanghai analysis explicitly reports an exploratory moderation finding that does not reproduce under one scoring alternative [10]. A defensible benchmark response preserves that uncertainty rather than selecting whichever result supports an enthusiastic recommendation. Burns and colleagues' whole-school implementation study requires attention to its embedded case design and actual contributing schools [11]. Neither source becomes a universal district intervention result through citation alone.

The benchmark tests authority boundaries as well as evidence recall. When asked to approve an intervention, diagnose an individual, change a publication status or claim that AI completed peer review, a successful response explains the evidential and human process required. The participant's brief should likewise retain the named accountable professional. Software may draft or retrieve; it cannot bear scholarly responsibility or become the independent human peer whose review underwrites a journal decision.

A useful interface may still fail the institutional test if verification requires excessive effort or produces inaccessible interaction. Source-checking time, assistance requests and screen-reader or keyboard barriers are recorded separately from total time. A future deployment threshold must be set before a field trial and include severe-error bounds, maintenance cost and a correction route. This protocol does not convert a laboratory score into permission to automate personnel evaluation, admission, safeguarding or clinical action. Those decisions require distinct governance, evidence and authorization.

15. Limitations and conditions that would contradict the hypothesis

The corpus is a small relevance-ranked slice of a health-oriented scholarly index, not a probability sample of school research. Rights are interpreted conservatively from exact downloaded versions; unresolved wording may exclude material that a later permissions check would clear. Conversely, a clear article license does not remove restrictions on embedded third-party content. The audit is not a comprehensive retraction service, and its status observations expire as the scholarly record changes.

The human experiment is prospective. A fixed-case task cannot establish changes in school practice, equity, student achievement, wellbeing or clinical outcomes. English-language recruitment, volunteer participation, professional experience and a structured time limit

restrict transfer. Participants may recognize the software condition, and apparently neutral briefs may reveal its wording to raters. Masking success is assessed by asking raters to guess arm with confidence after scoring, without using those guesses to edit unfavorable results.

The hypothesis would be contradicted if the treatment failed to improve fidelity within practically relevant uncertainty, increased severe unsupported claims, raised confidence without accuracy, or saved time by discouraging necessary checking. A trade-off cannot be settled through a composite score chosen afterward. The strongest result would be improved warranted reasoning with usable verification, followed by independent replication. Even that result would support the evaluated configuration and case domain, not a claim that the institution possesses an infallible intelligence layer.

Method and data statement

This original article reports a completed database and full-text feasibility audit and a completed artificial-data design analysis, alongside a prospective primary human study. There were zero trial participants and no observed treatment outcomes. Europe PMC core metadata and 36 full-text JATS snapshots were inspected through documented API endpoints; raw third-party documents remain outside the public package. Public audit records preserve original authors, identifiers, dates, licenses, hashes, exclusions and access limitations. The methods package includes instruments, an analysis specification, executable audit and simulation code, results and tests. Exact byte identity supports reproducibility, not truth, permission or endorsement. Registration, ethics approval and final case validation remain prerequisites for collecting primary participant data [13].

AI, authorship and editorial disclosure

ChatGPT assisted with research organization, code, source inspection and drafting on 15 October 2026 [16]. The tools executed real downloads, metadata parsing and explicitly artificial simulations; they did not recruit participants or supply a human peer-review report. Dr. Freedom Cheteni is the named human author, with responsibility for final approval and scholarly accountability. Source authors retain their contributions, and neither Europe PMC nor the source institutions are coauthors or endorsers of this protocol. This is a new companion manuscript to A03. The earlier conceptual article’s publisher-confirmed review does not constitute review, acceptance or authorization of this new protocol. No acceptance date, publication date, registration identifier, ethics approval or DOI is invented. The Editor in Chief’s authorship requires independent human handling of any subsequent editorial decision.

Research materials and reproducibility

Inspect the prospective study specification, original draft instruments, executed public-file audit, design simulations and documented tools. No local participant trial outcomes are reported.

Complete research package · ZIP

Methods guide

Study specification

Assessment instruments and scoring manual

Statistical analysis specification

Executable audit and design tools · Python

Completed corpus audit

Executed design simulations

Granular source-verification register

Reproducibility tests

Read the corresponding conceptual foundation

Evidence and limits

Claim	Evidence	Limit
A paper-agent interface can be evaluated without delegating human authority.	Four-stratum adult randomization, source-readable active control and independently blinded human evidence-fidelity scoring. [1, 13, 14]	Prospective design; no human performance effect observed.
Open access and exact-version reuse eligibility are distinct.	Executed audit:36 JATS bodies, 29 eligible article-license URLs, 4 restricted and 3 unresolved rights records; 23 pass all construction rules. [2, 3, 8, 9]	A bounded 36 of 98 query slice, not a systematic review or integrity certification.
A valid identifier can still point to a different source version.	Observed F1000Research indexed DOI .2 versus downloaded JATS article DOI .3; both retained. [12]	Version conflict requires human resolution; no allegation of misconduct or source invalidity.
Twelve responses per adult do not supply twelve independent participants.	Executed simulations under three within-person case correlations, with 10,000 replicates/scenario. [13, 15]	Artificial Gaussian assumptions; no observed agent effect, dropout or case-treatment heterogeneity.

References and rights

[1] Freedom Cheteni. Embracing the AI Revolution in Education: A Movement Thinking Approach. K12 Digest. 2023-11-09.

https://www.k12digest.com/embracing-the-ai-revolution-in-education-a-movement-thinking-approach/

Citation only; no republication rights asserted. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[2] Europe PMC. Europe PMC RESTful Web Service. EMBL-EBI Europe PMC. Accessed 2026-10-15.

https://europepmc.org/RestfulWebService

Citation only; no republication rights asserted. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[3] Europe PMC. Open access subset: copyright and web services. EMBL-EBI Europe PMC. Accessed 2026-10-15.

https://europepmc.org/downloads/openaccess

Citation only; no republication rights asserted. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[4] Aaron R Lyon; Catherine M Corbin; Eric C Brown; Mark G Ehrhart; Jill Locke; Chayna Davis; Elissa Picozzi; Gregory A Aarons; Clayton R Cook. Leading the charge in the education sector: development and validation of the School Implementation Leadership Scale (SILS). Implementation Science : IS. 2022-07-19.

https://doi.org/10.1186/s13012-022-01222-7

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[5] Sara B Låftman; Viveca Östberg; Bitte Modin. School Leadership and Cyberbullying - A Multilevel Analysis. International Journal of Environmental Research and Public Health. 2017-10-15.

https://doi.org/10.3390/ijerph14101226

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[6] Yuangen Bao. A dual mediation model of the association between principal inclusive leadership and teacher innovative behavior. Scientific Reports. 2024-05-29.

https://doi.org/10.1038/s41598-024-63332-0

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[7] Brian P Daly; Annie Resnikoff; Shannon Litke. Effective School Leadership for Supporting Students’ Mental Health: Findings from a Narrative Literature Review. Behavioral Sciences. 2025-01-01.

https://doi.org/10.3390/bs15010036

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[8] Jill Locke; Catherine M Corbin; Roger Goosey; Vaughan K Collins; Mark G Ehrhart; Kurt Hatch; Christine Espeland; Aaron R Lyon. Not getting better but not getting worse: A cluster randomized controlled pilot trial of a leadership implementation strategy. Implementation Research and Practice. 2025-01-29.

https://doi.org/10.1177/26334895241312405

Citation only; restricted or conflicting source rights statements. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[9] Hasan Yücel Ertem. School leadership fostering mental health in the times of crisis: synthesis of school principals' views and PISA 2022. BMC Psychology. 2024-11-26.

https://doi.org/10.1186/s40359-024-02195-6

Citation only; restricted or conflicting source rights statements. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[10] Mengting Qian; Chunshun Yan; Xinyi Wang. Perceived school-wide teacher-student relations and teacher job satisfaction: An affective events theory analysis of Shanghai teachers. PLOS One. 2026-09-03.

https://doi.org/10.1371/journal.pone.0357279

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[11] Sharyn Burns; Roisin Glasgow-Collins; Hanna Saltis; Jacqueline Hendriks. Implementation of a Whole-School Relationship and Sexuality Education Project in Western Australian Schools: Evaluation Data From a Multiple, Embedded Case Study. Health Promotion Journal of Australia. 2025-12-03.

https://doi.org/10.1002/hpja.70138

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[12] Yihenew Belay; Mateb Tafere. Leading beyond the Job Description: Investigating the Association between Principals' Leadership Styles and Teachers' Organizational Citizenship Behaviors in North Wollo Secondary Schools. F1000Research. 2026-10-08 (downloaded JATS version 3).

https://doi.org/10.12688/f1000research.183874.3

CC BY 4.0. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[13] An-Wen Chan; Isabelle Boutron; Sally Hopewell; David Moher; Kenneth F Schulz; Gary S Collins; Ruth Tunn; Rakesh Aggarwal; Michael Berkwits; Jesse A Berlin; Nita Bhandari; Nancy J Butcher; Marion K Campbell; Runcie C W Chidebe; Diana R Elbourne; Andrew J Farmer; Dean A Fergusson; Robert M Golub; Steven N Goodman; Tammy C Hoffmann; John P A Ioannidis; Brennan C Kahan; Rachel L Knowles; Sarah E Lamb; Steff Lewis; Elizabeth Loder; Martin Offringa; Philippe Ravaud; Dawn P Richards; Frank W Rockhold; David L Schriger; Nandi L Siegfried; Sophie Staniszewska; Rod S Taylor; Lehana Thabane; David J Torgerson; Sunita Vohra; Ian R White; Asbjørn Hróbjartsson. SPIRIT 2025 statement: updated guideline for protocols of randomised trials. The BMJ. 2025-04-28.

https://doi.org/10.1136/bmj-2024-081477

Citation only; restricted or conflicting source rights statements. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[14] Tammy C Hoffmann; Paul P Glasziou; Isabelle Boutron; Ruairidh Milne; Rafael Perera; David Moher; Douglas G Altman; Virginia Barbour; Helen Macdonald; Marie Johnston; Sarah E Lamb; Mary Dixon-Woods; Peter McCulloch; Jeremy C Wyatt; An-Wen Chan; Susan Michie. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014-03-07.

https://doi.org/10.1136/bmj.g1687

Citation only; no republication rights asserted. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[15] John Barnard; Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. Biometrika. 1999-12.

https://doi.org/10.1093/biomet/86.4.948

Citation only; no republication rights asserted. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[16] OpenAI. ChatGPT [AI-assisted drafting and coding software]. Software/tool citation, not an author or peer reviewer. Accessed 2026-10-15.

https://chatgpt.com/

Tool use disclosed; no exact model version asserted. Original authors and rights holders retain their contributions; no source text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

Original primary-study protocol and reproducible corpus/design methods by Dr. Freedom Cheteni. Original source authors, titles, venues, dates, versions and rights remain distinct. No third-party full text, figure, participant record or reviewer report is redistributed. New-manuscript licensing and author approval are separate from source permissions.