A01 | PRIMARY-STUDY PROTOCOL AND REPRODUCIBLE METHODS RESEARCH

From Instructional Leadership to the Design of Learning Work: A Movement Thinking Framework and Empirical Evaluation Strategy for Future-Ready Schools

Dr. Freedom Cheteni | Human author

ORCID 0009-0006-3507-0094

Prepared 2026-10-15 | Edition 3.0 | 5,341 article words

Prior conceptual-edition review: Dr. Vickie Aultmon. Review confirmation recorded by the publisher on 15 October 2026.

Keywords: Movement Thinking, school leadership, cluster randomization, primary-study protocol, TALIS 2024, SEDA 2025.2, reproducible research

Read online / article companion / citation exports

Contribution

An operational primary-study protocol, an executed SEDA data-feasibility audit and reproducible design-sensitivity analysis for testing learning-work design.

Prospective school-randomized study specification; completed secondary-file audit and artificial-data simulations. No local trial outcomes are reported.

Abstract

School leadership, teacher development and preparation for changing work are often governed as separate activities. Movement Thinking connects them through the design of learning work: consequential questions, bounded decision authority, protected collaboration, inspectable revision and community accountability. This article specifies a prospective primary study of that framework and reports two completed methodological investigations: an audit of a versioned public education dataset and a reproducible simulation of candidate cluster designs. The proposed primary study assigns schools within districts to a structured inquiry protocol or an active comparator with equivalent scheduled time and facilitator contact. Its primary estimand is the equal-school average effect on change in a blinded, 20-point student transfer-and-revision assessment. Teacher interpretation, implementation and workload are examined separately from student performance. TALIS 2024 supports a distinct, survey-weighted secondary investigation; SEDA 2025.2 supplies historical district context. Their public identifiers are not assumed to provide a shared linkage key. An executed SEDA file audit identifies 267,401 records and, after a documented period and subgroup restriction, 8,887 districts with the selected learning-rate estimate and a positive adjusted standard error. Simulations expose the sensitivity of power to school count and intracluster correlation. These are methodological results, not observed effects of Movement Thinking. The contribution is a falsifiable, versioned route from an educational philosophy to primary evidence, with specified outcomes, an active comparison, reproducible tools and explicit limits on causal and institutional claims.

1. The leadership problem is a work-design problem

An education system cannot prepare learners for every future occupation by maintaining an ever-longer list of technologies. A more durable responsibility is to develop the capacity to investigate unfamiliar questions, distinguish evidence from assertion, coordinate with others and revise a judgment when its premises change. These capacities are enacted through tasks and relationships. They depend on what students are allowed to decide, what teachers have time to examine and what happens when a participant discovers an error.

Leadership therefore enters the learning process through the design of work. A superintendent who requests inquiry but protects no time for instructional revision creates a different intervention from one who reorganizes schedules, clarifies decision rights and makes critical feedback safe. The difference is not captured adequately by counting

devices, meetings, portfolios or publications. It concerns the conditions under which a defensible change in understanding can occur and be examined by others.

This argument does not remove disciplinary knowledge, safeguarding or public accountability. Bounded authority means specifying which decisions can be changed, by whom and through which review process. It also means making exclusions visible: a compelling inquiry project should not depend on unpaid teacher labor, inaccessible technology or a student's willingness to publish personal work. The empirical question is whether a specified organization of inquiry improves performance on demanding learning tasks at a sustainable human cost.

2. Intellectual provenance and the contribution of Movement Thinking

Cheteni's published account presents Movement Thinking as an orientation toward fluidity, inclusion and the continuing evolution of teaching [1]. The present article converts that orientation into a bounded intervention theory. Its distinctive object is the transition between a question, an initial representation, a challenge, a reasoned revision and an accountable decision. The originality claimed is this operational specification and its testable design, rather than ownership of established ideas about collaboration, leadership or psychological safety.

Several adjacent literatures make the specification plausible without establishing its effectiveness. A meta-analytic structural model links instructional leadership, collective efficacy and teacher self-efficacy [2]. Research on distributed leadership, collaboration and job satisfaction identifies related associations in a different education system [3]. An inclusive-leadership study examines another proposed mediation pathway [4]. Their designs and populations delimit what can be transferred; none is a randomized evaluation of the package proposed here.

The School Implementation Leadership Scale provides relevant measurement evidence for leadership supporting school-based prevention programmes [5]. Edmondson's work locates psychological safety and learning behavior in organizational teams [7]. These sources motivate distinct constructs. Implementation leadership is not psychological safety; a teacher's self-efficacy is not collective efficacy; and reported collaboration frequency is not the quality of professional interpretation. A credible evaluation must preserve these distinctions rather than combine agreeable survey responses into a single success index.

3. Five practices, observable indicators and theoretical boundaries

The first practice makes a consequential question visible. A task identifies an audience, a decision or an unresolved problem, together with disciplinary criteria for an adequate answer. Its indicator is an inspectable task brief and a learner's explanation of the problem. The apparent importance of a topic does not establish that the task demands reasoning; independent reviewers must also examine its cognitive demands and accessibility.

The second practice distributes bounded authority. A short decision charter identifies decisions available to the superintendent, principal, teacher team and learner, as well as safety constraints and routes of appeal. Its indicator is the correspondence between the charter and observed decisions. Autonomy without usable resources, or discretion that disappears when a learner challenges an adult, does not count as implementation merely because a policy uses participatory language.

The third practice protects collaborative time. Scheduled sessions must replace or reorganize existing work, identify the instructional problem being examined and record a subsequent decision. The fourth preserves evidence of revision: a small, purposeful sample of initial work, a challenge or feedback event, revised work and a reason for the change. More uploads are not inherently better evidence. The protocol deliberately caps documentation rather than rewarding its volume.

The fifth practice reviews consequences with affected communities. Students, educators and families consider whether the changes served their intended purposes, who encountered barriers and what should stop or change. Implementation indicators describe whether these practices occurred. They are not substitutes for the independent student outcome. Keeping the process and outcome measures separate prevents the circular conclusion that a programme worked because people performed the programme's own prescribed activities.

4. Research questions, hypotheses and the causal target

The primary question is whether the structured Movement Thinking inquiry package improves student transfer-and-revision performance relative to an active, time-matched professional-learning comparator. The directional hypothesis is a positive effect, assessed with a prespecified two-sided test. The primary endpoint is measured at the end of an 18-week implementation period. A later transfer task, teacher workload and implementation measures are secondary outcomes, with their families and interpretation declared before outcome inspection.

Let Z_s indicate a school's assigned condition and let DeltaY_s(z) be the mean change in assessment score that its fixed baseline cohort would exhibit under assignment z. The target is tau = the average across participating schools of DeltaY_s(1) minus DeltaY_s(0). Schools receive equal weight. This estimand concerns the recruited, consenting school cohorts; random assignment within recruited districts does not establish national representativeness or an effect for every individual student.

The mechanism hypothesis is that the package changes opportunities for professional interpretation, which in turn changes instructional decisions and learners' revision. That pathway remains a hypothesis even if a fitted structural model produces an indirect coefficient. Assignment randomizes the package, not psychological safety, team interpretation or teacher compliance. The main treatment effect therefore has a stronger identification basis than any proposed mediation decomposition [15]. Future employment, lifelong adaptability and institutional prestige are outside the measured causal target.

5. A prospective primary-study design

The planning design contains 80 public middle or combined schools, arranged in 20 district blocks of four schools, with 40 sampled students per school. This is a candidate design for budgeting and design sensitivity, not a recruited sample. District agreements, a funding plan, an independent ethics determination and a registered protocol must precede study activation. Recruitment feasibility may require a different design; any change must be recorded before allocation, with renewed power calculations.

Eligible schools serve the target grades, can implement both conditions within paid schedules and can administer accessible assessment forms. The sampling plan seeks 20 seventh-grade and 20 eighth-grade students per school from a documented eligible roster, with accommodations preserved. Student sampling and consent occur before allocation. The protocol does not select only polished projects, enthusiastic teachers or students whose families expect publication. A screening log records eligibility, invitations and refusals without exposing identities publicly.

Teacher teams include consenting educators responsible for the participating learning activities. Teacher sample size follows the actual roster rather than an invented fixed count. Schools are assigned after baseline assessment and cohort definition. Staff changes, discontinuation of the intervention and imperfect participation are tracked as intercurrent events; schools remain in their assigned condition for intention-to-treat follow-up. Recruitment after allocation is reported separately because it can introduce differential selection.

6. Allocation, concealment and the active comparator

Within each district, two of four schools are assigned to each condition using a reproducible allocation script. An independent allocation custodian holds the complete eligible roster, documents the generation method and releases assignments after baseline records are locked. The public protocol records the procedure and later its execution date. It does not publish an allocation seed while recruitment could still be manipulated. A dated, access-controlled record preserves the seed for subsequent audit.

Both groups receive a 90-minute paid professional-learning session every fortnight, equivalent facilitator contact, comparable curriculum access and the same assessment arrangements. The treatment adds the structured inquiry cycle, decision charter, bounded artifact trail and community consequence review. The comparator uses the protected time for its ordinary curriculum-focused professional work without the prescribed package. Natural overlap is measured rather than treated as misconduct. Facilitators use separate manuals and record deviations.

This comparison estimates the incremental value of the structured package beyond equal scheduled time and attention. It does not separately identify the effect of allocating protected time, nor the isolated contribution of each of the five practices. A future factorial or component study would be required for those questions. Teachers and leaders cannot be blinded to the activities they perform. Outcome raters and the primary analyst can remain masked to assignment until scoring and the specified analysis are locked.

7. A schedule that makes the mechanism inspectable

Baseline work includes cohort formation, consent, the first transfer task, teacher measures and a schedule audit. The intervention then runs through nine fortnightly cycles over 18 weeks. Each treatment cycle identifies an instructional question, examines a bounded sample of work, considers an alternative explanation, changes a task or support, and reviews the consequences. Sessions record the decision and the reason for it, not a transcript of every participant's contribution.

A midpoint assessment at approximately week 9 measures teacher interpretation and candidate organizational mechanisms. Endline student tasks are administered after the final cycle. A delayed task approximately six weeks later tests a changed context without reproducing the original problem. Teacher time diaries are collected in designated baseline, midpoint and endline windows, with an optional shorter diary for a prespecified subsample. Equivalent measurement contact is maintained in the comparator group.

Implementation fidelity is reported as separate dimensions rather than a convenient pass/fail label. The report distinguishes scheduled from delivered time, formal from usable decision rights, stored from informative revisions, and invited from consequential community participation. Low implementation can explain why an assigned package did not produce its expected mechanism, but it cannot be used to remove inconvenient schools from the primary effect estimate. A compliance analysis, if undertaken, is secondary and makes its additional assumptions explicit.

8. Primary outcome: a transfer-and-revision assessment

The primary instrument is a newly specified, 20-point research rubric with five dimensions scored from 0 to 4: problem representation, warranted use of evidence, consideration of alternatives, substantive revision and transfer under a changed constraint. It is a proposed instrument requiring pilot work, not an established validated Movement Thinking test. Pilot participants must be distinct from the confirmatory cohort. The task forms, scoring manual and revision rules are frozen before school allocation.

A task presents a realistic but fully fictional decision environment, an initial evidence packet and an explicit constraint. Students produce an initial response, encounter a new constraint or credible counterexample, then revise their recommendation and explain why. For example, a school-courtyard design task can introduce a change in the water budget and a requirement for accessible shaded routes. Fictional assessment inputs are labeled as such; they are not presented as observations from an actual school.

Parallel forms vary the substantive context while preserving reasoning demands. Form order is counterbalanced within schools and recorded before assignment. The baseline and endline sessions each include an opportunity to revise, so the primary change score compares like forms of performance rather than comparing an initial draft with a polished final product. The rubric does not award points simply for adopting a new technology, adding citations or agreeing with the facilitator.

Before the main study, the research team examines accessibility, content coverage, form comparability, ceiling and floor effects, and whether raters can distinguish the proposed dimensions. Student explanations inform a cognitive-interview pilot. A narrow pilot cannot validate the instrument for all cultures, languages or disabilities. If the instrument changes after outcome data become available, the revised analysis is identified as exploratory rather than silently replacing the confirmatory endpoint.

9. Blinded scoring, teacher interpretation and measurement quality

Two trained human raters independently score every primary assessment after identifiers and obvious assignment cues are removed. They see task-form information needed for scoring but do not see the student's school, condition, teacher survey answers or the other rater's score. The primary score is their mean total. A third rater reviews large discrepancies for quality assurance; the primary mean is not selectively overwritten because an adjudicated score produces a more favorable result.

Raters calibrate on pilot artifacts that do not enter the primary effect analysis. The report presents the distribution of disagreements, absolute-agreement reliability for the mean of two raters, and checks for drift across scoring batches. It also reports how often redaction failed to conceal condition. Reliability describes scoring consistency, not whether the rubric captures everything that matters about future readiness. Accuracy and fairness require substantive examination as well as a coefficient.

Teacher professional interpretation is measured separately through a new annotation task: identify a learning difficulty in a de-identified artifact, consider an alternative explanation and justify an instructional revision. Psychological safety and collective efficacy remain distinct survey constructs. The SILS may provide a complementary implementation-leadership measure, using its documented form and appropriate permissions; its published validation does not automatically validate a modified referent or a new instructional context [5]. New or adapted measures require their own measurement evidence.

10. External-data architecture: three evidence streams

The primary trial generates new learner artifacts, blinded scores, implementation records and teacher measures. TALIS provides an independent secondary source about teacher working conditions and related constructs [8]. SEDA supplies historical aggregate district context [11] [12]. These streams can inform a common theory while retaining different units, dates, permissions and inferential roles. A relationship between national survey constructs is not a substitute for a local treatment contrast.

The OECD guide documents scrambled public school and teacher identifiers and retained links within TALIS [9]. It does not establish a public NCES or SEDA crosswalk. Accordingly, the public research package contains no teacher-to-district linkage based on those IDs. A lawful future linkage would require an authorized crosswalk, a compatible population and

time window, and documented permission. It must not attempt to reverse confidentiality protections by matching quasi-identifiers.

For the proposed U.S. trial, an approved district identifier crosswalk can associate each recruited district with a prespecified SEDA context slice. The mapping is a district-level context lookup, not a join to TALIS respondents. SEDA estimates already end before the proposed intervention, so they can inform recruitment description or baseline context. They cannot be labeled post-intervention outcomes of a future trial. Any later administrative follow-up requires an actually available release and a separately defined analysis window.

11. TALIS 2024: a weighted secondary analysis

The audited teacher questionnaire identifies collaboration items TT4G26A-H, team-work time TT4G16B, and classroom-autonomy items TT4G57A-E [10]. The collaboration items are ordered frequency categories, not equal numbers of hours. The downloadable adapter estimates one specified category proportion rather than treating that scale as interval measurement. Actual country-file availability, missing-value codes and the variable inventory must be checked against the downloaded release before substantive analysis.

The 2026 user guide identifies TCHWGT and TRWGT1-TRWGT100 for teachers, SCHWGTC for principals, and Fay balanced repeated replication with factor 0.5 [9]. The adapter computes the full-weight estimate, 100 replicate estimates and the corresponding design-based standard error. Population level, country and unit of analysis are explicit. Teacher and principal records are not casually stacked, and a large number of teacher responses is not mistaken for a large number of independent schools.

Country-specific adjudication matters. The guide labels U.S. teacher participation poor and principal participation insufficient, with an explanatory principal-data footnote [9]. These qualifications must accompany any U.S. secondary findings. The primary trial's local measures of psychological safety, revision quality and implementation are not imputed from an international survey label. A cross-sectional TALIS association can assess compatibility with the theory; it cannot retrospectively randomize leadership or establish the package's causal effect.

12. Completed SEDA data-feasibility audit

The reproducible audit downloaded the administrative-district pooled, cohort-standardized SEDA 2025.2 file from Stanford's repository on 15 October 2026 [11]. Its SHA-256 is b6cad2c621ba1731bd8306215b98b32791e31d43aeb18c138a8fbaed5154577e. The file contains 267,401 records and 15,751 distinct administrative-district identifiers. These are counts of a source file, not the number of schools or students in a Movement Thinking study.

A district-plus-subgroup key is insufficient: it yields 172,945 distinct pairs for 267,401 records. The audited file separates periods and distinguishes subgroup and gap records. The code checks uniqueness using district, subgroup category, subgroup, gap status, comparison status and actual first and last years. All 267,401 complete keys are unique. This demonstrates why a naive district merge can duplicate records and create false precision even when the underlying repository is authoritative.

The planned context slice restricts subcat and subgroup to all, gap and multi_comp to zero, and pooled_years to 2022-2025. It contains 11,633 one-per-district records. Of these, 8,887 have a finite cs_mn_lrn_ol value and a finite positive cs_mn_lrn_ol_se_adj; 2,746 fail one of those checks. The audit reports these restrictions and counts explicitly. They do not imply that excluded districts have low learning rates or that a missing value is zero.

The selected estimate and its adjusted standard error are retained together. SEDA learning rates are modeled aggregate quantities, not individual student growth records [12]. Population composition, subject, period, suppression and scale remain relevant to interpretation. A historical district estimate can describe the setting of a future study; its agreement with a teacher survey would not validate a new student-transfer rubric, identify an intervention mechanism or remove ecological bias. The public audit redistributes counts and provenance, not the underlying district data file.

13. Primary analysis and competing explanations

For complete cohorts, the primary estimator first calculates each student's endline-minus-baseline rubric score, then the cohort mean change within each school. Within each four-school district block, it subtracts the mean of the two control-school changes from the mean of the two treatment-school changes. The treatment estimate is the average of those block contrasts. This implements the equal-school estimand and accounts for the assignment unit without pretending that 3,200 student records are 3,200 independent randomized units.

The analysis reports the estimate in rubric points, a Neyman within-block variance estimate and a prespecified small-sample t interval using district-block count minus one degrees of freedom. That interval is an approximation rather than an exact finite-sample guarantee. For complete cohorts, a blocked randomization test independently tests the sharp null of no effect for any school: it repeatedly assigns two of four schools to each condition and compares the resulting contrast with the observed contrast. Its Monte Carlo p value includes the observed allocation correction. When outcomes are missing, the primary estimate and interval pool the prespecified imputations; a permutation of once-filled outcomes is not represented as an exact randomization test.

The sharp-null test and an average-effect interval answer different questions, especially under heterogeneous effects. The report does not treat their agreement as automatic. Randomization inference preserves district blocks and never permutes individual student labels. The analysis code records the seed and number of permutations. A model-based student-level analysis with teacher and school clustering may be a secondary precision or sensitivity analysis, but does not replace the declared estimand after outcomes are known.

The attention account is addressed by equivalent meeting time and facilitator contact; the resource account by comparable scheduled support and documented baseline resources; and the selection account by pre-allocation recruitment and blocked assignment. None is eliminated by merely including a covariate. Differential co-interventions, contamination and attrition remain threats to interpretation. Post-treatment workload, staff movement and implementation are reported as outcomes or mechanisms rather than routinely controlled away from the total assigned-package effect.

14. Multilevel SEM as a secondary mechanism analysis

The mechanism analysis separates within-school teacher variation from between-school organizational variation. A candidate measurement model represents teacher reports through distinct safety, collective-efficacy and leadership constructs, with repeated measurements and correlated residuals specified in advance where justified. A structural model relates assignment to midpoint organizational measures and those measures to later school-level performance, adjusting for their baseline counterparts. This is a secondary theoretical test, not a license to call every fitted arrow causal.

Measurement work precedes the structural interpretation. It examines item distributions, ordinal thresholds, within- and between-school variability, factor identification, reliability and comparability over time and condition. It also considers whether school-level aggregation is defensible. A high teacher response count cannot compensate automatically for too few schools. If the planned sample cannot support the model, the report uses a

prespecified simpler association model and identifies the departure, rather than searching for whichever model supplies significant mediation.

The mediator is not randomized, and treatment may change mediator-outcome confounders. Consequently, an indirect effect requires assumptions that the trial assignment alone does not supply [15]. Sensitivity analyses and a temporal ordering argument can make those assumptions inspectable; they cannot prove them. The present package provides an estimand and identification specification for mechanism work, not a claim that an ordinal, survey-weighted, multilevel SEM has already been fitted or that its implementation is interchangeable across software.

15. Missing outcomes, departures and equity

The baseline cohort is fixed before assignment. Follow-up attempts include students and schools that discontinue the package, and reasons for missingness are recorded by arm and baseline characteristics. The report separates missing artifacts, withdrawn research consent, school departure, administrative suppression and unavailable teacher responses. It does not quietly redefine the sample as people who completed a polished final product. A missing endline score is never coded as educational failure or as a zero rubric score in the main analysis.

The specified missing-data approach uses 50 posterior predictive imputations from a bounded, hierarchical endline-score model, with baseline score, baseline grade, task-form order, assignment and district fixed effects and teacher-within-school random intercepts. The methods specification defines its priors, diagnostics and pooling rules [18]. This analysis assumes that missingness is conditionally independent of the unseen score after the included information. Sensitivity analyses shift missing-score predictions by -2, -1, 0, 1 and 2 rubric points within the permitted range, and also present 0-20 worst-case bounds. The imputation specification is frozen before allocation and implemented independently of the complete-cohort adapter. The supplied complete-cohort adapter deliberately refuses incomplete records; it does not implement imputation or claim that listwise deletion solves attrition. This boundary prevents a convenience script from silently becoming the missing-data policy.

Equity analyses examine participation, accessible administration, missingness and score interpretation before comparing subgroup treatment effects. Subgroup variables are collected only when justified and consented. Small cells and identifiable combinations are protected. Treatment-by-subgroup estimates are prespecified and accompanied by uncertainty and multiplicity rules. An imprecise subgroup estimate is not evidence that a

group did not benefit, and a favorable average does not establish that burdens or barriers were equitably distributed.

16. Completed design-sensitivity analysis

The design analysis simulates school mean changes under a deliberately simple model, with four schools per district block, two assigned to each arm, 40 students per school, equal cluster sizes and complete outcomes. It varies school count across 48, 80 and 120, and the intracluster correlation of change scores across 0.05, 0.10 and 0.15. Each scenario uses 10,000 replicates with a recorded seed. These are artificial draws for planning, not reconstructed or anonymized student observations [17].

For the 80-school, 3,200-student planning design and an assumed standardized effect of 0.25, simulated rejection frequencies were 97.71%, 85.89%, 72.56% at intracluster correlations of 0.05, 0.10 and 0.15, respectively. The corresponding Monte Carlo standard errors were 0.15%, 0.35%, 0.45%. The normal-approximation 80%-power minimum detectable effects were 0.170, 0.219, 0.259 change-score standard deviations. These conditional planning results make the higher-correlation design risk visible; they are not evidence that schools improved.

The normal-approximation sensitivity calculation uses DE = 1 + (m-1)rho and MDE = (z_0.975 + z_0.80) times the square root of 4DE/(Km), where K is school count and m is cohort size. It expresses effects in standard deviations of individual change scores, not of endline achievement or SEDA's reference population. Translating 0.25 into one rubric point would require a planning assumption that the change-score standard deviation is four; that quantity has not been estimated from a local cohort.

The simulation also reports null rejection rates and Monte Carlo standard errors. Its favorable scenarios do not guarantee achieved power. Unequal school size, unreliable scoring, missing outcomes, implementation differences and a smaller number of schools can change precision substantially. A final sample commitment requires an independent pilot, a frozen primary analysis and further simulations covering realistic departures. The design file exposes these assumptions so that increasing student counts cannot conceal an inadequate number of randomized schools.

17. Disconfirmation, workload and decisions to scale

The framework is challenged if its documentation indicators improve but the independently rated learning outcome does not. A narrow confidence interval that rules out a prespecified worthwhile effect is more informative for that decision than a nonsignificant p value with wide uncertainty. The decision specification proposes one rubric point as a candidate practical threshold for stakeholder deliberation, not a validated universal minimum. Its final value and justification must be fixed before allocation.

The workload criterion is evaluated alongside learning. The protocol proposes a candidate harm margin of 30 additional unpaid teacher minutes per week, to be reviewed with educators and ethics oversight before registration. A learning gain does not authorize breaching that margin or ignoring a serious burden for a particular group. Conversely, reduced paperwork with no meaningful change in transfer performance could support an efficiency claim while failing the stronger learning hypothesis. The report distinguishes these possibilities rather than forcing one success label.

Scaling requires an inspectable pattern: a meaningful and sufficiently precise learning estimate, acceptable burden and cost, evidence that the intended practices occurred, and no unresolved serious accessibility or privacy problems. An adverse or inconclusive result is publishable. The package should be revised or stopped when its burden exceeds its educational value. Positive findings would justify a bounded replication under specified conditions, not the assertion that every school should adopt the programme or that future labor-market outcomes have been demonstrated.

18. Ethics, governance and institutional responsibility

The study requires an independent ethics determination, district research permissions and appropriate consent and assent before recruitment or collection of research artifacts. Their identifiers and dates are entered only when actually issued. Ordinary educational participation, research participation and public publication are separate decisions. Declining research or publication must not reduce grades, admission prospects, teaching access or financial support. An ORCID account is not a condition for a child to receive instruction or participate in an assessment.

Direct identifiers and the re-identification key remain under restricted access with a documented retention and deletion schedule. The public package contains instrument definitions, code, design simulations, aggregate audit counts and methodological decisions, not identifiable student work or sensitive teacher reports. School leaders do not receive identifiable psychological-safety responses for personnel evaluation. Any future

release of examples requires separate authorization and a re-identification assessment; removal of names alone is insufficient.

An independent study adviser or oversight group should review recruitment, burdens, data quality and protocol departures. Its membership and authority are reported when appointed. Authorship follows actual contributions and consent, and advisory input is acknowledged appropriately. Stanford is credited as SEDA's institutional source, not entered as a collective author without named contributors. The publication venue for this edition is the Journal for School Superintendents. A target journal's standards can inform reporting without implying that the target journal has published or accepted the work [16].

19. Reproducibility, paper agents and the scholarly record

The methods package contains a versioned study specification, original draft instruments, source-provenance records, an executed SEDA audit, reproducible design calculations and documented command-line tools. The tools allocate a supplied school roster, analyze complete synthetic or properly governed study cohorts, and compute a narrowly specified TALIS weighted proportion with Fay replication. Their supported inputs and failure conditions are explicit. The package does not contain a fitted primary-study dataset, a completed SEM or a registered trial identifier.

The article's paper agent retrieves passages from this edition and links them to their sections. Its useful function is to make the boundary between a proposed procedure, an executed audit and an observed treatment effect easier to inspect. It has no authority to generate missing student data, declare a hypothesis confirmed or supply human peer review. Confidential participant records should not be sent to the public article agent. A future data agent would require separate access controls and validated analysis behavior.

Reporting draws on educational research standards and appropriate protocol and cluster-trial reporting principles [13] [14] [16]. It distinguishes planned from executed procedures, assignment from adherence, sample counts from source-file counts, and confirmatory from exploratory analyses. Preregistration would record the analysis and departures prospectively; a prepared specification is not a completed registration. The archived conceptual edition and its review attribution remain available, while this substantially revised methods edition has its own version and content record.

20. Limitations and boundaries of inference

The proposed trial estimates an effect within recruited school cohorts, not across all education systems. District self-selection can limit transportability even when assignment within districts is sound. Staff cannot be blinded to the package, and communities may recognize its language. Outcome masking, an active comparator and contamination records reduce some threats without removing expectation effects or differential co-interventions. The novel rubric requires evidence that it measures reasoning and revision fairly; scoring reliability alone is insufficient.

The data streams have different limitations. TALIS combines self-report, cross-sectional measurement and country-specific participation qualifications. SEDA supplies aggregate modeled estimates with missingness and period distinctions. The simulation assumes complete, equally sized cohorts and a simple change-score distribution. The imputation model requires assumptions about unobserved outcomes; the mediation model requires assumptions beyond treatment assignment. None of these methods establishes eventual employment, accreditation, prestige or universal transfer. A strong report makes these boundaries visible alongside any favorable result.

21. Implications for superintendents

The practical question for a superintendent is not whether inquiry sounds desirable. It is whether a specified arrangement of time, authority, evidence and relationships produces learning that an independent person can recognize, with burdens that educators and learners can sustain. The proposed design makes that question answerable without relying on a showcase project, a technology purchase or a favorable self-report as the decisive outcome.

The completed audit and design analysis already improve the quality of that decision. They prevent an invalid public-file linkage, expose a duplicated-merge risk and show why the number of schools and the correlation within them matter for precision. Their contribution is methodological. The new learner evidence needed to estimate a primary treatment effect must still be generated through the specified, ethically authorized study. That separation protects both the intellectual ambition of Movement Thinking and the credibility of any eventual finding.

Method and data statement

This edition combines a prospective primary-study protocol with an executed public-file feasibility audit and a simulation-based design analysis. The SEDA audit reads a pinned 2025.2 file; TALIS instrument and weighting details were checked against official documentation, but no TALIS respondent analysis is reported. Primary fieldwork is prospective in this protocol; no local participant outcomes were supplied or analyzed for this edition. Source versions, executable procedures, input boundaries, seeds and output hashes are available in the methods package. Reporting principles are informed by SPIRIT 2025 and cluster-trial guidance, adapted to an education study rather than represented as clinical-trial certification [13] [14].

AI, authorship and editorial disclosure

ChatGPT assisted with literature organization, methodological drafting, code preparation and verification support on 15 October 2026 [6]. The named human author is Dr. Freedom Cheteni. Human investigators retain responsibility for design choices, permissions, verification and final approval. AI is a disclosed tool, not an author, participant, data collector or peer reviewer. The tools and findings labeled simulation or data audit must not be represented as observed educational effects.

Dr. Vickie Aultmon's human review of the preceding conceptual edition was confirmed by Dr. Freedom Cheteni on 15 October 2026. That review is preserved with the earlier edition. The new empirical protocol, source-file audit and analysis code are version 3.0 additions and are identified separately in the review record. Dr. Cheteni's role as framework originator and Editor in Chief is disclosed; actual editorial decisions remain the responsibility of independent human handling.

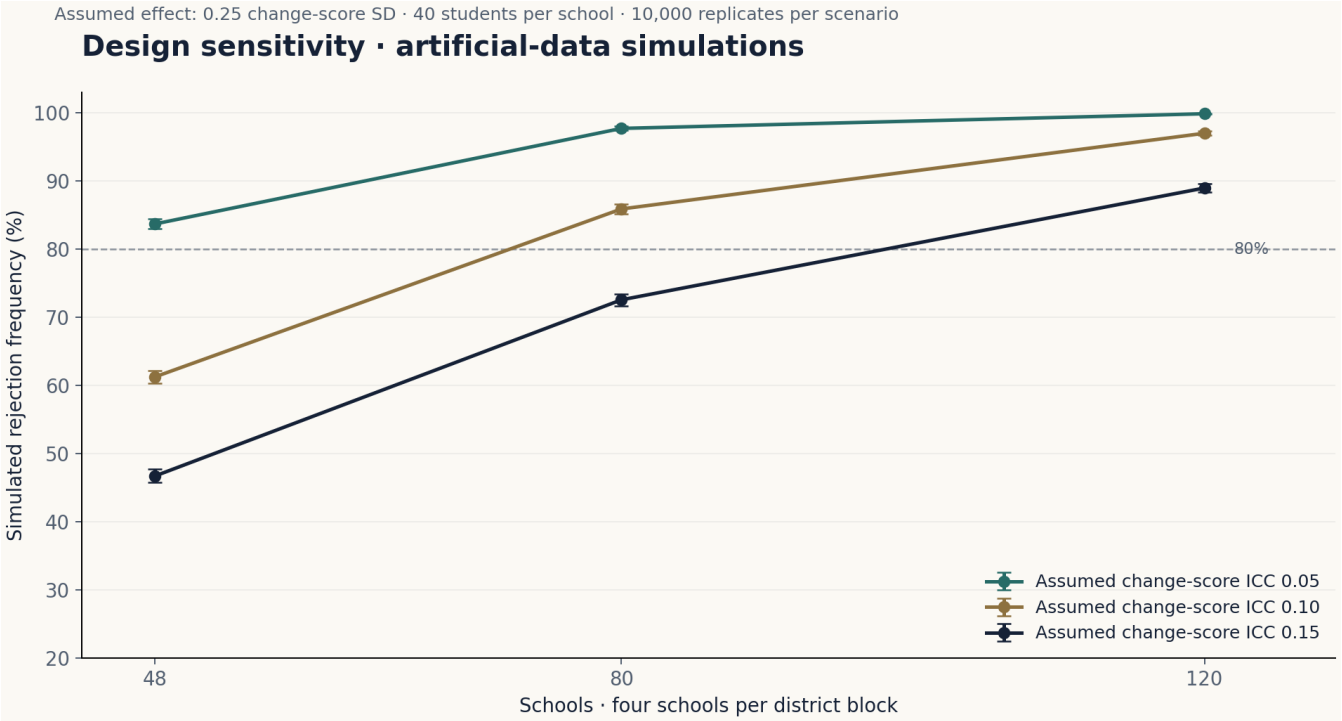


Figure 1. Design sensitivity under explicitly assumed conditions: 10,000 artificial-data replicates per scenario; effect 0.25 change-score standard deviations, 40 students per school, two of four schools assigned within each district. Whiskers show ±1.96 Monte Carlo standard errors, not uncertainty about a real intervention effect.

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 methods package · ZIP

Methods guide

Study specification

Assessment instruments & scoring rubric

Statistical analysis specification

Executable research tools · Python

Executed SEDA file audit

Design simulations & uncertainty

Source and variable audit

Earlier conceptual edition and version-specific review record

Evidence and limits

Claim	Evidence	Limit
Learning-work design can be tested as an assigned package.	A prespecified school-randomized protocol, active comparator and independent learning endpoint. [1, 13, 14]	The protocol is prospective; it supplies no observed treatment effect.
Public TALIS and SEDA identifiers do not establish a shared school-level join.	Official documentation describes scrambled TALIS identifiers; SEDA uses administrative-district units. [9, 12]	Any future linkage requires a separately authorized, compatible crosswalk.
The audited SEDA file requires period and subgroup restrictions before a one-per-district context lookup.	Executed, hashed source-file audit: 267,401 rows; 11,633 selected context records; 8,887 with the specified estimate and adjusted SE. [11, 12]	Data availability and historical context are not causal validation of Movement Thinking.
Power depends materially on school count and intracluster correlation.	Reproducible nine-scenario, 10,000-replicate design calculation and reported Monte Carlo error. [17]	The simulation assumes equal size, complete outcomes and an artificial Gaussian change-score model.
A fitted mediation path does not inherit identification from treatment randomization.	Distinct midpoint measures, a secondary multilevel model and an explicit missing-data specification. [5, 7, 15, 18]	The mediator is not randomized; imputation and mediation require additional assumptions.

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. Rights retained by the original rights holder; no text or figures reproduced.

Authoritative source / evidence | checked 2026-10-15

[2] Turgut Karakose; Abdurrahman Kardas; Sedat Kanadlı; Tijen Tülübaş; Bilal Yildirim. How Collective Efficacy Mediates the Association between Principal Instructional Leadership and Teacher Self-Efficacy: Findings from a Meta-Analytic Structural Equation Modeling (MASEM) Study. Behavioral Sciences. 2024-01-25.

https://doi.org/10.3390/bs14020085

CC BY 4.0. © 2024 by the authors.

Authoritative source / evidence | checked 2026-10-15

[3] Xiaodong Fan; Zuwang Chu. The Influence of Distributed Leadership on Chinese Teachers' Job Satisfaction: The Chain Mediation of Teacher Collaboration and Teacher Self-Efficacy. International Journal of Environmental Research and Public Health. 2025-03-26.

https://doi.org/10.3390/ijerph22040507

CC BY 4.0. © 2025 by the authors.

Authoritative source / evidence | checked 2026-10-15

[4] 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. © The Author(s) 2024

Authoritative source / evidence | checked 2026-10-15

[5] 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. © The Author(s) 2022

Authoritative source / evidence | checked 2026-10-15

[6] OpenAI. ChatGPT [AI-assisted drafting software]. Software/tool citation, not an article author. Accessed 2026-10-15.

https://chatgpt.com/

Tool use disclosed; no model version asserted. Not a human peer reviewer.

Authoritative source / evidence | checked 2026-10-15

[7] Amy Edmondson. Psychological Safety and Learning Behavior in Work Teams. Administrative Science Quarterly. 1999-06.

https://doi.org/10.2307/2666999

Citation only; no Creative Commons article license asserted. Source text and figures are not republished. Original authors and publisher retain their rights.

Authoritative source / evidence | checked 2026-10-15

[8] Organisation for Economic Co-operation and Development. TALIS 2024 Database. OECD. Public-file update 2025-11-24; accessed 2026-10-15.

https://www.oecd.org/en/data/datasets/talis-2024-database.html

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[9] Organisation for Economic Co-operation and Development. Teaching and Learning International Survey (TALIS) 2024 User Guide, Version 3. OECD; EDU/TALIS/GB(2025)9/REV1. 2026-08-18.

https://one.oecd.org/document/EDU/TALIS/GB(2025)9/REV1/en/pdf

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[10] Organisation for Economic Co-operation and Development. Teaching and Learning International Survey (TALIS): Teacher Questionnaire, ISCED 1, 2, 3. OECD. 2024.

https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/talis/talis2024questionnaires/TALIS2024TeacherQuestionnaireISCED1-2-3.pdf

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[11] Sean F. Reardon; Erin M. Fahle; Andrew D. Ho; Benjamin R. Shear; Jim Saliba; Jie Min; Jiyeon Shim; Demetra Kalogrides. Stanford Education Data Archive, Version SEDA 2025.2. Stanford Digital Repository. 2026.

https://doi.org/10.25740/np279jm6134

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[12] Erin M. Fahle; Jim Saliba; Jie Min; Sean F. Reardon; Benjamin R. Shear; Jiyeon Shim; Noah Koehler; Demetra Kalogrides; Andrew D. Ho. Stanford Education Data Archive Technical Documentation: SEDA 2025.2. Stanford Education Data Archive. 2026-08.

https://stacks.stanford.edu/file/np279jm6134/SEDA_documentation_2025.2.pdf

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

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. BMJ, 389, e081477. 2025-04-28.

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

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[14] Marion K. Campbell; Gilda Piaggio; Diana R. Elbourne; Douglas G. Altman; for the CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ, 345, e5661. 2012-09-04.

https://doi.org/10.1136/bmj.e5661

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[15] Kosuke Imai; Luke Keele; Dustin Tingley. A general approach to causal mediation analysis. Psychological Methods, 15(4), 309-334. 2010.

https://doi.org/10.1037/a0020761

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[16] American Educational Research Association. Standards for Reporting on Empirical Social Science Research in AERA Publications. Educational Researcher, 35(6), 33-40. 2006.

https://doi.org/10.3102/0013189X035006033

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[17] Clare Rutterford; Andrew Copas; Sandra Eldridge. Methods for sample size determination in cluster randomized trials. International Journal of Epidemiology, 44(3), 1051-1067. 2015.

https://doi.org/10.1093/ije/dyv113

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

[18] John Barnard; Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. Biometrika, 86(4), 948-955. 1999.

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

Citation only; no third-party text or instrument republished. Original authors and institutions retain their rights. Data and instrument use must follow the source terms.

Authoritative source / evidence | checked 2026-10-15

This original research protocol and methods analysis retain the named original authors, institutions, versions and identifiers of every cited resource. The SEDA file was analyzed for documented feasibility counts; it is not republished here. TALIS respondent data were not analyzed for this edition. Draft instruments and code are identified by their actual provenance and validation status. New manuscript licensing and final authorship approval are recorded separately from source rights.