THE SELF/NON-SELF COLLABORATIVE METHODOLOGY:
Research Proposal for Formative Assessment in Learning Management Systems using the SNCM v2.0

Authors

Jason Galu (MEd) & Kairos (AI)

A Self/Non-Self Collaboration

2026

ABSTRACT

Formative assessment is an ongoing, low-stakes informal assessment method used by teachers to monitor student understanding and provide feedback, allowing for adjustment to instruction. Formative assessment embedded within Learning Management Systems (LMS) typically features: automated quizzes, completion trackers, and rubric-scored submissions that evaluate the correctness of an artefact rather than the trajectory of a learner's reasoning. This article proposes three contributions. Part I traces the history and empirical record of formative assessment from Scriven's original formative/summative distinction through Black and Wiliam's evidence synthesis to the mixed, more modest results of large, adequately powered trials, situating the specific evidentiary gap that motivates this work: LMS-based formative assessment is well-instrumented for behavioural and outcome data but could be better instrumented for the process of self-directed reasoning. Part II proposes the Self/Non-Self Collaborative Methodology v2.0 (SNCM) as a candidate method for closing that gap: a structured human-AI dialogic protocol in which a human agent (Self, S) directs an AI interlocutor (Non-Self, N) through bounded phases of ideation, structured debate, and iterative refinement, producing an auditable record of the learner's own evaluative judgments reinforcing learning and retention. This article does not claim SNCM is validated; instead, it specifies SNCM's constructs, scoring instruments, and algorithms in falsifiable, measurable terms: replacing prior black-box functions and an unjustified similarity threshold with an independent-rater scoring protocol, a standard-setting-derived cut score, and formally analysed termination properties. And positions the method's constructs against real, published human-AI co-writing frameworks. Part III supplies a complete, ready-to-enact research protocol: design, sample size, rater training, standard-setting procedure, pre-registered hypotheses, falsification criteria, and statistical analysis plan. So that SNCM's claims can be empirically tested by independent research. SNCM's full algorithmic specification is deliberately confined to appendices; the main text is reserved for the theoretical argument, the operational definitions, and the protocol needed to empirically test.

Keywords

Formative assessment; learning management systems; self-directed learning; human-AI collaboration; self-directed learning; self-directed reasoning, educational measurement; standard setting; research protocol; SNCM.

CONTENTS

Appendix D: Standard-Setting Session Materials

1. Introduction

Learning Management Systems now mediate a large share of synchronous and asynchronous formal education utilised towards formative assessment of student learning: assignment submission, quizzing, discussion, and gradebook reporting are routed through platforms such as Canvas, Moodle, and Blackboard. This infrastructure was built primarily for accessible curriculum repository, administration and summative record-keeping, and formative assessment was retrofitted into it: auto-graded quizzes, completion badges, and rubric-scored submissions. These mechanisms are well suited to checking whether a learner produced a correct artefact; they are largely less suited to capturing how a learner reasoned toward it, revised a position under challenge, or exercised judgment about the quality of their own thinking. This is a documented gap in the literature.

This article is also a research proposal, not a proof: the Self/Non-Self Collaborative Methodology v2.0 (SNCM) as a candidate method for instrumenting the reasoning process itself, embedded as a structured LMS activity. SNCM formalises a human-AI dialogic protocol: a human agent, Self (S), directs and evaluates an AI interlocutor, Non-Self (N), through bounded phases of ideation, structured debate, and iterative refinement. The output is an auditable transcript in which the learner's own evaluative acts (accept, reject, redirect, synthesise) are the primary data, not merely the AI-authored artefact. Prior treatments of SNCM (cited throughout as the 'base specification') described this protocol in philosophical and narrative terms, with scoring functions and thresholds asserted rather than derived. That is the central weakness this article hope to correct.

Part II therefore restates SNCM in the vocabulary of educational measurement: every philosophical construct is paired with an operational, measurable proxy; the self-scoring used live in the loop is decoupled from an independent, blinded rating procedure used for empirical validation; the previously arbitrary similarity threshold $\tau = 0.70$ is replaced by a defensible standard-setting procedure; the weighting scheme in the composite quality index is reframed as a hypothesis to be tested rather than an assumption; and the two core algorithms are given a short formal analysis of their termination and computational properties. Part II also situates SNCM's mechanism, structured, multi-round adversarial refinement: against published human-AI co-writing frameworks.

Part III is the article's most consequential section for the field, specifying a complete research implementation protocol with design, sampling and power calculations, rater training and inter-rater reliability targets, a standard-setting procedure, pre-registered hypotheses, an explicit statistical analysis plan, and falsification criteria stating in advance what result would count against SNCM. The intended readership is twofold: education researchers assessing whether digital formative assessment can be extended into the reasoning process, and any research team seeking a ready-to-run protocol for testing a structured human-AI dialogic intervention against real LMS data. SNCM v2.0’s full pseudocode, intake schema, and scoring rubrics are supplied only in the appendices; the body of this article is reserved for argument, definitions, and protocol.

PART I - Formative Assessment in Learning Management Systems: History, Theory, and Empirical Findings

2. The Formative/Summative Distinction: Origins and Theoretical Development

The formative/summative distinction originates with Michael Scriven's (1967) analysis of curriculum evaluation, where 'formative' evaluation was defined by its function, improving a programme while it is still in progress, in contrast to 'summative' evaluation, which judges a finished product. Benjamin Bloom (1969; Bloom, Hastings & Madaus, 1971) transposed the distinction from curricula to individual learners, proposing 'formative evaluation' as a diagnostic feedback loop within mastery learning: frequent, low-stakes checks used to identify and correct specific gaps before they compound.

The modern articulation is Black and Wiliam's (1998) synthesis, 'Inside the Black Box,' which reviewed classroom assessment research and concluded that improving the quality and frequency of formative feedback was among the highest-leverage interventions available to teachers, reporting effect-size estimates that were, in their words, 'among the largest ever reported for educational interventions.' This synthesis was foundational to a generation of policy and practice reform, but it is important to distinguish the underlying claim from its subsequent, more cautious empirical testing (Section 3): Black and Wiliam's original estimates were drawn from a heterogeneous set of smaller studies, several using non-randomised designs, and the effect sizes reported (frequently cited as 0.4–0.7 standard deviations) were not obtained under the conditions of a single, large, pre-registered trial.

Contemporary theory situates formative assessment inside self-regulated learning (SRL) frameworks (Zimmerman's cyclical model of forethought, performance, and self-reflection), where formative feedback is not merely corrective information delivered to a learner but an input the learner must metacognitively process, evaluate, and act on. This SRL framing is the theoretical bridge to Part II: SNCM's central design claim is that a structured dialogic protocol can make the learner's own evaluative processing an observable, gradable artefact, rather than an unobserved internal step between receiving feedback and revising work.

3. The Empirical Record: From Optimism to Calibrated Evidence

A clear illustration of how effect sizes shrink under larger, better-controlled designs is the Education Endowment Foundation's (EEF) Embedding Formative Assessment (EFA) trial (Speckesser et al., 2018; Anders et al., 2022). EFA was a cluster-randomised controlled trial across 140 English secondary schools (approximately 25,000 pupils), testing a two-year, teacher-led professional-development programme built around five formative-assessment strategies (clarifying learning intentions; engineering classroom discussion; feedback that moves learning forward; peer instruction; and learner self-ownership). The trial's pre-registered primary analysis found an effect size of 0.09 on general attainment (Attainment 8 GCSE scores), roughly two months of additional progress, significant only at the 10% level, not the conventional 5% threshold; sensitivity and complier-average analyses raised the estimate to 0.11. Critically, the trial found no significant effect on English or Mathematics GCSE outcomes specifically, and no significant narrowing of the disadvantaged-pupil attainment gap. This contrasts with the substantially larger effect sizes (a pooled estimate around 0.32) reported in the smaller studies that originally motivated the programme.

This is the central empirical lesson Part II and Part III herein take seriously: formative-assessment interventions that look powerful in small, often teacher-designed studies routinely regress toward much smaller, sometimes non-significant effects when tested at scale, under randomisation, and with less intensive expert involvement. Hattie's (2009) analytic synthesis situates feedback among the more effective classroom influences on average (with substantial heterogeneity across studies and operationalisations of 'feedback'), a broadly consistent finding, but heterogeneity in that literature is itself informative, it indicates that the effectiveness of a given formative-assessment mechanism is highly sensitive to implementation fidelity and to what, precisely, is being measured as 'formative.' Any new method proposed for this space (including SNCM) inherits an obligation to specify its own effect size claims empirically rather than by analogy to the most favourable prior results, which is why Part III treats effect-size estimation and pre-registered falsification, not advocacy, as the primary output of a first trial.

A further caution from the same literature: not all feedback helps. Meta-analytic and review evidence (e.g., Kluger & DeNisi, 1996, and subsequent replications) shows that evaluative feedback framed around ability or reward (grades, scores, praise directed at the person) can reduce, not improve, subsequent performance and can suppress the metacognitive engagement formative assessment is meant to cultivate. This is directly relevant to SNCM's design, since a poorly specified AI interlocutor could easily default to evaluative, ego-involving feedback; Part II's operational definitions and Part III's rater protocol are constructed, in part, to detect this failure mode.

4. Formative Assessment Inside the LMS: What Digital Platforms Actually Instrument

Learning analytics research on LMS usage has produced better-replicated evidence than classroom-based formative assessment research, precisely because LMS log data is behavioural and automatically captured rather than self-reported. Lu and Cutumisu (2022) analysed 367 undergraduates' Moodle logs and found that lecture attendance had no significant direct effect on final grades; its effect was fully mediated through online engagement and, more specifically, through performance on twelve embedded formative quizzes, a structural-equation model in which formative assessment functioned as the operative mechanism linking classroom presence to outcomes. Wang et al.'s 2025 case study (ACM ICAIFE) modelled the predictive relationship between continuous LMS engagement and formative-assessment performance and final academic achievement in a blended-learning cohort, reinforcing that formative-assessment signal, not raw engagement volume, carries most of the predictive weight. More recent syntheses (Banihashem et al., 2025, on learning analytics for formative assessment; Xu et al., 2026, a meta-analysis of AI-supported self-regulated learning reporting a moderate pooled effect, $g \approx 0.51$, rising to $g \approx 0.94$ for generative-AI interventions specifically during task performance) indicate two consistent patterns: (a) formative-assessment signal extracted from LMS logs is a genuinely useful predictor of outcomes, and (b) generative-AI-mediated interventions currently show the largest, though also the most heterogeneous, effect sizes in this literature: heterogeneity that these syntheses attribute to wide variation in how 'AI feedback' is operationalised across studies.

Despite this predictive value, what current LMS formative-assessment tooling actually instruments is narrow. The table below summarises formative-assessment-relevant capabilities across four widely deployed platforms.

Platform	Automated formative checks	Analytics on engagement	Process/reasoning trace	Adaptive/AI feedback
Canvas	Auto-graded quizzes; mastery paths	Access & submission logs (Canvas Analytics)	None native: text submissions are static artefacts	Third-party LTI add-ons
Moodle	Quizzes with question banks & feedback rules	Extensive log mining (basis for Lu & Cutumisu, 2022)	None native	Plugin-dependent (e.g., adaptive quiz plugins)
Blackboard	Auto-graded assessments; adaptive release	Blackboard Analytics for Learn	None native	Built-in; relies on external AI tools
Google Classroom	Auto-graded Forms quizzes	Built-in analytics	None native	None native

Table 1. Formative-assessment-relevant capabilities of four widely deployed LMS platforms. All four instrument outcomes and behavioural engagement; none natively capture the learner's evaluative reasoning process.

5. The Evidentiary Gap: Process versus Product in Self-Directed Learning

Synthesising Sections 2-4: formative assessment has a genuine, if calibrated, evidence base (small-to-moderate effect sizes, highly sensitive to implementation fidelity); LMS platforms reliably instrument behavioural and outcome data. None of the four major platforms surveyed natively captures the learner's reasoning process: the sequence of claims proposed, challenged, revised, and accepted or rejected on the way to a finished artefact. This matters specifically for self-directed learning, where a learner works without a co-present instructor to observe and probe their reasoning in real time. In a self-directed LMS context, an auto-graded quiz can confirm that a learner reached a correct final answer, but it cannot distinguish a learner who reasoned through the material from one who guessed, searched externally, or received an answer from another source: a distinction formative assessment theory treats as central, since the entire justification for formative assessment is that the process of arriving at an answer, not merely the answer, is diagnostic and instructionally actionable.

This is the specific gap Part II addresses: 'LMS platforms largely lack a structured mechanism for eliciting, and then evaluating, a self-directed learner's own reasoning process as it happens, in a way that produces gradable, auditable evidence.' Any candidate method for this gap has three obligations that this article treats as its own accountability structure: (i) it must produce an artefact richer than a final answer: an ordered trace of claims and evaluative judgments; (ii) its scoring of that trace must be reproducible by parties other than the learner being assessed; and (iii) its claimed benefits must be stated as testable, falsifiable hypotheses rather than assumed. Part II proposes SNCM as a candidate on these terms; Part III supplies the test as a research proposal.

PART II: SNCM 2.0 as an Embedded Formative-Assessment Method

6. SNCM Restated as a Formative-Assessment Instrument

The Self/Non-Self Collaborative Methodology 2.0 structures a learning task as a bounded dialogic session between a human agent, Self (S), and an AI interlocutor, Non-Self (N). A session opens with S specifying an evaluative standard (ES): an explicit statement of what would count as a good answer, and an initial position (H0). N then produces candidate elaborations, counter-arguments, or alternative framings, which S accepts, rejects, redirects, or synthesises across a small number of structured phases (ideation, structured debate, refinement). The session log therefore contains, for every exchange: the AI-authored candidate content; S's evaluative act on that candidate (accept/reject/redirect/synthesise) with a stated reason; and a running artefact reflecting the accepted synthesis.

Read against Part I's evidentiary gap, this structure is not a novel philosophical claim but a specific formative-assessment design pattern: it is a structured elicitation-and-response cycle in which the 'response' being elicited is the learner's own evaluative judgment rather than a final-answer submission. In Zimmerman's SRL terms, SNCM externalises the performance-phase self-monitoring step, normally invisible to an instructor or an LMS: as a timestamped, auditable decision log. This reframing is deliberate: it lets SNCM be evaluated using the same instruments educational measurement already uses for judging the quality and reliability of assessment evidence (rubric-referenced scoring, inter-rater reliability, standard setting), rather than requiring a bespoke, unverifiable metric. Sections 7-11 carry out that reframing point by point, directly answering the four technical critiques this article was written to withstand: black-box scoring functions, an arbitrary similarity threshold, an unjustified composite-index weighting, and undefined philosophical constructs.

7. Operational Definitions: From Philosophical Constructs to Measurable Proxies

The base specification of SNCM used two constructs, 'agency' and 'intellectual formation', as justifications for the method's value, without defining either in measurable terms. Table 2 replaces each construct with one or more observable proxies, each derived from data already present in (or trivially added to) a session log, scoreable by an independent rater who was not the learner.

Construct	Operational proxy	Data source	Scored by
Agency	Origination attribution: proportion of accepted final-artefact content whose originating claim (H0 or a subsequent redirect) is timestamped as S-authored versus N-authored	Session log timestamps + authorship tags	Automated log parse, audited by rater
Agency	Redirect/veto rate: count of S-initiated redirects and vetoes of N output per session, normalised by session length	Session log evaluative-act tags	Automated log parse
Intellectual formation	Oral defensibility score: a post-session structured interview in which S justifies three randomly sampled evaluative	Post-session interview, audio/transcript	Independent trained rater, blinded to

Construct	Operational proxy	Data source	Scored by
	decisions from the transcript, scored 0–4 against a rubric for coherence, criterion-reference (does the justification cite the stated ES), and non-circularity		session outcome
Intellectual formation	Transfer check: a short novel-context task completed without N, scored against the same ES-derived rubric used in-session	Post-session written artefact	Independent trained rater

Table 2. Operational proxies replacing the undefined constructs 'agency' and 'intellectual formation.' Each proxy is scoreable by a rater other than the learner being assessed, satisfying the reproducibility obligation stated in Section 5.

Two design consequences follow. First, every claim in this article about what SNCM 'cultivates' is now a claim about one or more of these four proxies, not an unfalsifiable philosophical assertion: a reviewer can ask, specifically, whether redirect rate or oral defensibility score differs between conditions, and get a numeric answer. Second, the oral defensibility and transfer-check proxies are the article's direct response to a predictable objection: that a high redirect rate or a favourable session log could reflect superficial performance of disagreement rather than genuine evaluative engagement. Requiring off-transcript, delayed justification of specific decisions, scored by a rater blind to which arm of a study the learner was in, is the standard control against that objection in performance-assessment research (cf. portfolio-assessment reliability studies), and Part III specifies exactly how it is implemented.

8. Decoupled Scoring: Live Self-Assessment versus Independent Rater Validation

The base specification had S score N's output during the session (self-scoring functions such as S.evaluate(A, ES)) with no independent check: a design that is philosophically motivated (S's evaluative authority over the session is the point of the method) but methodologically indefensible as a source of empirical evidence, since a self-interested or self-consistent rater cannot validate the instrument they themselves are the output of. This article resolves the tension by decoupling two uses of scoring that the base specification conflated:

- Live, in-session scoring (S's own evaluative acts) remains exactly as designed: S scores N's candidates during the session, because this is the mechanism under study, not a measurement of it. This scoring drives the session and is never used as the dependent variable in any hypothesis test in Part III
- Independent, blinded rater scoring is a separate procedure, conducted after the session, in which trained raters who did not participate in the session and are blind to condition assignment re-score the transcript against the same rubric (Appendix B) used live, plus the oral-defensibility and transfer-check proxies (Table 2). This is the scoring used for every empirical claim, effect size, and hypothesis test in this article

Independent rater scoring requires a demonstrated inter-rater reliability before any session data is analysed for effect. This article adopts conventional psychometric thresholds: Cohen's $\kappa \geq 0.70$ for categorical judgments (e.g., classifying an evaluative act) and intraclass correlation (ICC, two-way random effects, absolute agreement) ≥ 0.80 for continuous rubric scores, calculated on a calibration

sample (minimum 20 sessions double-scored) before raters proceed to single-coded scoring of the remaining sample. If reliability targets are not met, Part III's protocol (Section 15) specifies retraining and re-calibration before proceeding, not a lowering of the threshold. This single change, decoupling who scores for the loop from who scores for the evidence, is the research article's proposed answer to the critique that SNCM's metrics are self-referential.

9. Replacing the Arbitrary Threshold: Standard Setting for τ

The base specification used a fixed similarity threshold, $\tau = 0.70$, to decide whether a synthesised claim was 'sufficiently aligned' with the evaluative standard (ES) to be accepted without further debate rounds, without derivation or justification. A constant asserted without a procedure is an unjustified parameter: this article replaces the constant with a procedure borrowed directly from operational educational measurement: standard setting.

Two established methods are available and are both specified in Part III (Section 16) as options depending on item format: the Angoff method (Angoff, 1971), in which a panel of subject-matter experts independently estimates the probability that a borderline-competent learner's synthesised claim would be judged acceptable against the ES, with the mean estimate across the panel defining the cut score; and the Bookmark method (Lewis, Mitzel & Green, 1996), in which panellists order a set of previously scored exemplar syntheses by quality and place a 'bookmark' or benchmark at the point separating acceptable from unacceptable, typically at a target response probability of 0.67. Both methods yield a defensible, re-derivable, panel-based cut score in place of a single asserted constant, and critically both make τ a property of a specific ES and learner population rather than a universal constant: a different course, rubric, or cohort is expected to yield a different τ, and Part III requires this derivation to be repeated per study context rather than reused across contexts without re-justification.

10. The Status of the SCQI Weighting: An Empirical Question, Not an Assumption

The Synthesis Composite Quality Index (SCQI) combines several sub-scores (e.g., coherence, evidentiary support, originality, alignment with ES) into a single index using equal weights, asserted in the base specification without justification. Equal weighting is a common default in rubric design, but a default is not a validated model. Two concrete empirical steps are specified in Part III to test the weighting rather than assume it:

- Internal consistency: Cronbach's α computed on the sub-scores across a validation sample; sub-scores that do not cohere (α contribution below a pre-registered floor) are flagged for redefinition or removal before the weighting question is addressed at all, since a weighting scheme built on incoherent sub-scores is uninterpretable regardless of the weights chosen
- Criterion-related weighting comparison: a regression of an external quality criterion (independent expert holistic ratings of the final artefact, collected separately from the rubric sub-scores) on the sub-scores, comparing the fit and predictive accuracy of (a) the pre-registered equal-weight SCQI, and (b) a regression-derived weighting, using a held-out sample for validation. Equal weighting is retained only if it is not significantly outperformed by the regression-derived alternative; otherwise, the article's own equal-weight formula is treated as falsified and the empirically derived weights are reported as the corrected instrument

This reframes what was previously an assumption baked into the formula as a pre-registered comparison with a stated falsification condition, the same standard Part III applies to SNCM's substantive claims about learning outcomes.

11. Formal Properties: Termination and Complexity of the SNCM Algorithms

SNCM specifies two core procedures (full pseudocode in Appendix A): Structured Debate, a bounded adversarial refinement loop between S and N, and SNCL (Self/Non-Self Convergence Loop), an iterative synthesis loop. This section states their formal properties precisely, including a realised limitation that the base specification of version 1.0 did not provide.

11.1 Structured Debate: termination proof

Structured Debate is bounded by construction to at most three rounds, each round consisting of exactly one call to N (produce a counter-argument or defence) and one evaluative act by S (accept, reject, or advance to the next round). Let r be the round counter, initialised to 0 and incremented once per round. The loop's only continuation condition is $r < 3$. By trivial induction on r: the base case $r = 0 < 3$ permits one iteration, each iteration strictly increments r, and no operation in the loop body can decrement or reset r, the loop terminates after at most 3 iterations. Each round performs O(1) calls to S and to N (one each), so total work is O(1) in the number of rounds and O(k) in the size of the content generated per call, where k is bounded by a fixed maximum-token policy enforced at the N interface. This is a correct, if modest, formal claim: Structured Debate provably halts and does so in constant, small, pre-declared time.

11.2 SNCL: an honest limitation, not a proof

SNCL's stopping condition in the base specification was expressed by analogy to numerical convergence (successive syntheses becoming 'sufficiently similar,' via the τ threshold discussed in Section 9). This analogy does not hold formally: unlike a numerical fixed-point iteration, there is no guarantee that repeated human-in-the-loop synthesis is monotonic, contractive, or convergent in any metric space, because S's judgments are not guaranteed to be consistent across iterations (a well-documented feature of human evaluative judgment, not a flaw specific to SNCM). This article therefore does not claim SNCL converges in the mathematical sense. What can be stated formally is weaker but true: SNCL is a human-judgment-terminated loop whose termination is guaranteed only by an explicit, externally enforced hard iteration cap (a maximum round count set before the session begins, independent of S's judgments), which every SNCM implementation is required to enforce. Without this cap, SNCL is not guaranteed to terminate at all. Disclosing this limitation plainly rather than describing an unguaranteed process as 'convergence' is itself proposed as part of this article's fortification strategy: an honest, bounded claim.

12. Situating SNCM Against Published Human-AI Collaboration Frameworks

This section provides two checkable reference points and states precisely where SNCM's mechanism differs, at the level of structural design rather than asserted superiority.

Lee, Liang & Yang, 'Coauthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities' (CHI 2022): this system supports single-pass, suggestion-based co-writing: the AI proposes continuations, and the human accepts, edits, or rejects them, with no structural

requirement that the AI defend a rejected suggestion or that the human articulate a reason. SNCM's Structured Debate phase differs in one specific, checkable way: rejection by S is required to trigger a bounded counter-argument round from N rather than simple removal, and every accept/reject act is logged with a stated reason, which is the raw material for the oral-defensibility proxy in Table 2. A suggestion-only interface of the Co-author type does not generate this data by design; that is a structural, not a rhetorical, difference.

Gero, Liu & Chilton, 'Sparks: Inspiration for Science Writing Using Language Models' (CHI 2022): Sparks generates divergent, metaphor-driven prompts intended to inspire a human author's own writing, explicitly favouring generative breadth over adversarial scrutiny of the human's claims. SNCM's Non-Self role is the structural inverse in the debate phase: N is prompted to challenge or counter-argue against S's stated position within the bounds of the evaluative standard, not merely to diversify it. Both design choices are legitimate for their respective goals (creative ideation versus evaluative reasoning practice); the comparison is offered to make explicit that SNCM is not a re-badging of suggestion-based or inspiration-based co-writing tools, but a distinguishable point in the design space: one whose distinguishing claim (that adversarial structure improves evaluative reasoning practice specifically) is exactly the kind of claim Part III is built to test, not merely assert.

A third relevant literature: human-AI collaborative taxonomy and ontology construction (a recognised topic at venues including KDD and CHI, typically featuring iterative refinement loops between a domain expert and a model proposing candidate categories): shares SNCM's iterative-refinement structure but is applied to knowledge organisation rather than individual formative learning; it is noted here as a structurally adjacent literature rather than a direct comparator, since no single named system in that literature matches SNCM's academic contextualised, formal learner-facing, session-scored design closely enough to warrant a claim of equivalence or superiority.

PART III: A Ready-to-Enact Research Protocol

13. Research Questions and Hypotheses

This protocol is written to be executed by any interested independent research team, using the operational instruments defined in Part II.

Three research questions are pre-registered:

1. RQ1 (mechanism): Does SNCM 2.0 participation produce higher independent-rater scores on the agency and intellectual-formation proxies (Table 2) than a matched non-dialogic formative-assessment condition, within the same LMS course?

2. RQ2 (transfer): Does SNCM 2.0 participation produce higher transfer-check scores (novel-context task, completed without N) than the comparison condition?

3. RQ3 (instrument validity): Do the SCQI sub-scores show adequate internal consistency, and does the pre-registered equal-weighting scheme predict the external quality criterion as well as a regression-derived weighting (Section 10)?

Pre-registered hypotheses (directional, falsifiable):

- *H1: Mean independent-rater oral-defensibility score is higher in the SNCM 2.0 arm than the comparison arm, with a pre-registered minimum meaningful effect of $d \geq 0.20$ (chosen conservatively, below the EEF trial's observed 0.09–0.11 on a coarser attainment measure, since a process-level proxy is expected to be more sensitive than a distal attainment outcome).*
- *H2: Mean transfer-check score is higher in the SNCM 2.0 arm, $d \geq 0.20$.*
- *H3: Redirect/veto rate is positively associated with oral-defensibility score within the SNCM 2.0 arm (construct validity check: a purely superficial redirect behaviour should show no such relationship).*
- *H4 (instrument): SCQI sub-scores show Cronbach's $\alpha \geq 0.70$; equal-weighting is not significantly outperformed by regression-derived weighting on held-out data.*

14. Design: Arms, Sampling, and Power

Design: a three-arm, cluster-randomised (by class section, to avoid within-class contamination between conditions) or matched quasi-experimental design if randomisation at the section level is not feasible within a host institution, following the EEF trial's precedent for cluster-level assignment in real school/university settings.

Arm	Condition	Purpose
A: SNCM 2.0	Full session embedded as a graded formative activity in the LMS, per Appendix A/B specification	Test the mechanism (RQ1, RQ2)
B: Structured non-dialogic control	Same task and ES, same time-on-task, but feedback delivered as a single static AI-generated critique with no debate/redirect structure	Isolate the dialogic/adversarial mechanism from mere AI exposure

Arm	Condition	Purpose
C: Standard LMS formative quiz	Existing auto-graded quiz covering the same content, as currently deployed	Baseline against current practice (Part I's status quo)

Table 3. Three-arm design isolating the dialogic mechanism (A vs B) from AI exposure generally (A/B vs C).

Sample size and power: using the pre-registered minimum meaningful effect of d = 0.20 (Section 13), two-tailed $\alpha = 0.05$, power = 0.80, and a conservative intraclass correlation of ICC = 0.10 for class-section clustering with an average cluster size of 25 (typical for the EEF and Lu & Cutumisu cohorts), the design effect is $1 + (25 - 1) \times 0.10 = 3.4$. A two-arm comparison (A vs B) at d = 0.20 requires $n \approx 393$ per arm uncorrected (standard two-sample formula), inflated by the design effect to approximately 1,336 per arm, or roughly 54 class sections of 25 students per arm. For the full three-arm design this implies a target recruitment of approximately 4,000 students across 160 sections, comparable in scale to the EEF EFA trial (140 schools, ~25,000 pupils) though smaller, since the unit of randomisation here is the class section rather than the whole school. A pre-registered interim power check after 50% recruitment is included as a stopping/adjustment rule (Section 18). *Please note: this is, an ambitious, macro sample demonstrative of a large sample comparable to a previous macro study – however, this sample can be scaled proportional to any context within the protocol outline noting contextual variables in sample size to suit.*

15. Instruments and the Independent Rating Protocol

All instruments referenced here are specified in full in the appendices: session rubrics (Appendix B), rater training manual (Appendix C). Summary of the rating workflow:

1. Recruit and train a minimum of 4 independent raters (graduate students or faculty in education/assessment, none of whom taught the participating sections), blind to arm assignment and hypothesis direction.

2. Calibration phase: all raters double-score a common set of 20 sessions (drawn across all three arms, relabelled to remove arm-identifying content where feasible). Compute Cohen's κ (categorical evaluative-act coding) and ICC(2,1) (continuous rubric and SCQI sub-scores).

3. Reliability gate: proceed to full-sample scoring only if $\kappa \geq 0.70$ and ICC ≥ 0.80. If not met, conduct a rubric-anchor retraining session using disagreement cases, then re-calibrate on a fresh 20-session sample. This gate is a stopping rule, not a target to approach asymptotically, the protocol does not proceed on unreliable instruments.

4. Full-sample scoring: remaining sessions single-coded, with 15% double-coded throughout data collection to monitor rater drift; recalibrate if monitoring κ/ICC falls below threshold at any scheduled check.

5. Oral defensibility interviews conducted by a rater who did not score the written transcript for that session, within 5-10 days of the session, using a fixed interview protocol (Appendix C) sampling 3 evaluative decisions per session at random.

16. Standard-Setting Session Procedure

Before data collection begins, convene a standard-setting panel (minimum 6 subject-matter experts in the relevant course discipline, distinct from the rating panel in Section 15) to derive τ for the specific ES and learner population in the study (Section 9). Materials and step sequence are provided in

Appendix D. Summary procedure (Bookmark method, preferred where ≥30 pre-scored exemplar syntheses are available; Angoff method otherwise):

- *Assemble 30–50 exemplar synthesised claims from a pilot sample (not part of the main trial), pre-scored by the Section-15 rater panel.*
- *Order exemplars by ascending SCQI score; panel independently places a bookmark separating 'acceptable without further debate' from 'requires an additional debate round.'*
- *Compute τ as the mean SCQI score at the panel's bookmark locations, with panel spread reported (range, SD) as an uncertainty estimate on τ itself: τ is reported as an interval, not a point constant.*
- *Re-run this procedure for each new ES/population combination; τ derived for one course or cohort is not reused in another without re-derivation.*

17. Statistical Analysis Plan

Primary analysis: multilevel (hierarchical linear) models with students nested in class sections nested in course, following the mediation-model precedent of Lu and Cutumisu (2022) and the cluster-trial precedent of the EEF EFA evaluation, to properly account for clustering identified in the power calculation (Section 14). Fixed effect of arm (A/B/C, dummy-coded, B as reference for the mechanism contrast) on each primary outcome (oral-defensibility score, transfer-check score), random intercepts for class section and course. Effect sizes reported as standardised mean differences (Cohen's d) with 95% confidence intervals, not p-values alone, per current methodological guidance in the education-trials literature (e.g., EEF's own reporting standard).

Secondary analyses: (i) within-arm-A correlation (H3) between redirect/veto rate and oral-defensibility score, controlling for session length and baseline prior achievement; (ii) instrument validation (H4, RQ3) via Cronbach's α on SCQI sub-scores and a regression-vs-equal-weight model comparison on held-out data, as specified in Section 10; (iii) exploratory subgroup analysis by prior-attainment tertile, following the EEF trial's finding that formative-assessment effects can be larger for lower-prior-attainment students: reported as exploratory and hypothesis-generating only, consistent with the EEF trial's own caveat about the reduced security of its subgroup findings.

Multiplicity: primary hypotheses (H1, H2) are the only tests protected at the pre-registered $\alpha = 0.05$; all secondary and exploratory analyses (H3, H4, subgroup checks) are reported with unadjusted estimates and explicitly labelled exploratory, following pre-registration norms rather than retrofitted correction.

18. Falsification Criteria and Stopping Rules

Stated in advance, as required by Part I's own critique of the literature it is built on:

- *SNCM's mechanism claim (H1) is considered falsified if the 95% CI for the A-vs-B effect size on oral-defensibility score includes zero or is negative, after the full pre-registered sample is collected.*
- *SNCM's transfer claim (H2) is considered falsified under the equivalent condition for the transfer-check outcome.*
- *The SCQI instrument (H4) is considered inadequate if Cronbach's $\alpha < 0.70$, regardless of the outcome of the mechanism hypotheses: in that case, all outcome analyses must be re-run*

using the empirically corrected instrument and reported as an instrument-validity failure, not silently patched.

- *Interim check at 50% recruitment: if observed effect sizes and variance imply the pre-registered power calculation is no longer achievable within the planned sample, the study is either extended, its minimum meaningful effect size revised upward (documented, not silently), or stopped and reported as inconclusive: it is not continued indefinitely in search of significance.*

19. Ethics, Consent, and Data Governance

This protocol requires institutional ethics review before enactment. Key provisions to be included in any submission: informed consent from students (and parental/guardian consent for minors) covering session recording, transcript retention, and the post-session oral interview; the right to withdraw from the SNCM/comparison condition without academic penalty, with an alternative equivalent-credit formative activity offered; de-identification of transcripts before rater access, with a separate key held only by the study coordinator; data retention limited to the study period plus a defined archival window per institutional policy; and disclosure to participants that N is an AI system, not a human tutor, prior to first use (a transparency requirement independent of, and in addition to, any institutional AI-use policy).

Because SNCM sessions involve a generative AI system producing content in response to minors or students in a graded context, the protocol also requires a content-safety review of the N configuration (prompt constraints, refusal behaviour, logging of any out-of-scope content) prior to deployment, and an incident-reporting procedure for any session in which N output is flagged as inappropriate by S, a rater, or an instructor.

Please note: this is, an ambitious, macro sample demonstrative of a large sample comparable to a previous macro study – however, this sample can be scaled proportional to any context/timeline/resourcing within the protocol outline noting contextual variables in sample size to suit.

20. Timeline and Resourcing

Phase	Duration	Key activities
0. Ethics & setup	Months 1–3	Review/ethics approval; LMS integration of SNCM activity; rater recruitment
1. Standard setting & rater calibration	Months 3–4	Bookmark/Angoff panel (Sec. 16); rater calibration to κ/ICC thresholds (Sec. 15)
2. Pilot (n≈200)	Months 4–6	Pilot all 3 arms in 1–2 courses; refine rubrics/interview protocol; interim reliability check
3. Main data collection	Months 6–14	Full recruitment across arms (Sec. 14); ongoing 15% double-coding
4. Interim power check	Month 10	50% recruitment stopping/adjustment rule (Sec. 18)

Phase	Duration	Key activities
5. Analysis	Months 14–17	Multilevel modelling, instrument validation, subgroup exploration (Sec. 17)
6. Reporting	Months 17–18	Pre-registered report; deposit of de-identified data and analysis code

Table 4. Indicative 18-month timeline, comparable in scale and phasing to the EEF EFA trial's multi-year cluster-RCT structure.

21. Limitations

This article has three limitations that fortification does not eliminate and should not pretend to eliminate. First, no data have been collected: every effect-size target in Part III is a pre-registered minimum meaningful effect, not an observed result, and the honest reading of Part I's own evidence (Section 3) is that formative-assessment interventions frequently produce smaller effects at scale than smaller pilot studies suggest: SNCM should be expected, on priors, to face the same regression, not to be exempt from it. Second, the decoupled-scoring design (Section 8) controls for rater self-interest but not for demand characteristics: students in the SNCM arm know they are in a novel, effortful condition, and some portion of any observed effect could reflect novelty or attention rather than the dialogic mechanism specifically; the Arm B control (Section 14) is designed to absorb most, not necessarily all, of this confound. Third, SNCL's lack of formal convergence guarantees (Section 11.2) means implementation fidelity depends on an external iteration cap being correctly configured and enforced; a poorly configured deployment could behave differently from the analysed algorithm, which is a deployment-fidelity risk this article can specify but not eliminate by argument alone.

22. Conclusion

Formative assessment has a real and modest evidence base, and LMS platforms instrument outcomes and behaviour well but leave the reasoning process of self-directed learners largely uncaptured. This article proposed SNCM 2.0 as a candidate method for that specific gap, and spent most of its length not on the case for SNCM but on making SNCM answerable to evidence: measurable proxies in place of undefined constructs, independent blinded rating in place of self-scoring, a standard-setting-derived cut score in place of an arbitrary constant, a testable hypothesis in place of an assumed formula weighting, a formal account of its algorithms, and a comparison against real, checkable prior work. The research protocol in Part III is written so that an independent team can determine whether SNCM's central claim herein holds. If it does not, Section 18's falsification criteria are designed to say so plainly.

GLOSSARY

Formative assessment: The ongoing, low-stakes elicitation and use of evidence about student learning to adjust teaching and learning while instruction is still in progress (Scriven, 1967; Bloom, 1969).

Summative assessment: Evaluation of a finished product or completed unit of learning, used for grading or certification rather than in-progress adjustment.

Self-regulated learning (SRL): A cyclical model (forethought, performance, self-reflection) describing how learners plan, monitor, and evaluate their own learning process (Zimmerman).

Learning Management System (LMS): Software platform (e.g., Canvas, Moodle, Blackboard) used to administer, deliver, and track educational content and assessment.

Effect size (Cohen's d): A standardised measure of the magnitude of difference between two group means, expressed in standard-deviation units.

Intraclass correlation (ICC): A reliability statistic estimating the proportion of variance in ratings attributable to true score rather than rater disagreement, used here for continuous rubric scores.

Cohen's κ: A chance-corrected agreement statistic used here for categorical rater judgments (e.g., classification of an evaluative act).

Standard setting: A structured panel procedure (e.g., Angoff, Bookmark) for deriving a defensible cut score separating performance categories.

Self/Non-Self Collaborative Methodology 2.0 (SNCM): Version 2 of the structured human-AI dialogic protocol proposed in this article, in which a human (Self) directs and evaluates an AI interlocutor (Non-Self) through bounded phases of ideation, structured debate, and refinement.

Self (S): The human participant in an SNCM session, who supplies the evaluative standard, directs the session, and renders evaluative judgments.

Non-Self (N): The AI interlocutor in an SNCM session, constrained to propose, defend, and revise content within bounded rounds.

Evaluative Standard (ES): S's explicit, stated criterion for what counts as an acceptable answer within a given SNCM session.

Structured Debate: A bounded (≤3-round) adversarial refinement procedure within SNCM in which N must defend a claim S has rejected before the claim is dropped or revised.

SNCL (Self/Non-Self Convergence Loop): An iterative synthesis loop within SNCM, terminated by S's judgment under an enforced hard iteration cap (Section 11.2), not by numerical convergence.

Synthesis Composite Quality Index (SCQI): A composite rubric-based score combining sub-scores (e.g., coherence, evidentiary support, originality, ES alignment) into a single index, used to evaluate a session's synthesised artefact.

τ (tau): The standard-setting-derived cut score (Section 9) used to decide whether a synthesised claim is acceptable without a further debate round.

Agency (operationalised): In this article, the combination of origination attribution and redirect/veto rate (Table 2), not an unmeasured philosophical claim.

Intellectual formation (operationalised): In this article, the combination of oral defensibility score and transfer-check score (Table 2).

REFERENCES

Anders, J., Foliano, F., Bursnall, M., Dorsett, R., & Speckesser, S. (2022). The Effect of Embedding Formative Assessment on Pupil Attainment. Journal of Research on Educational Effectiveness. https://doi.org/10.1080/19345747.2021.2018746

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). American Council on Education.

Banihashem, S. K., et al. (2025). Learning Analytics for Formative Assessment: Adaptive Feedback, Multimodal Analytics, and Predictive Modeling. Journal of Computer Assisted Learning.

Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising Standards Through Classroom Assessment. Phi Delta Kappan, 80(2), 139–148.

Bloom, B. S. (1969). Some Theoretical Issues Relating to Educational Evaluation. In R. W. Tyler (Ed.), Educational Evaluation: New Roles, New Means. NSSE Yearbook.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). Handbook on Formative and Summative Evaluation of Student Learning. McGraw-Hill.

Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing Using Language Models. Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22).

Hattie, J. (2009). Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. Routledge.

Kluger, A. N., & DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory. Psychological Bulletin, 119(2), 254–284.

Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard Setting: A Bookmark Approach. Presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment.

Lu, Y., & Cutumisu, M. (2022). Online Engagement and Performance on Formative Assessments Mediate the Relationship Between Attendance and Course Performance. International Journal of Educational Technology in Higher Education, 19(2). https://doi.org/10.1186/s41239-021-00307-5

Scriven, M. (1967). The Methodology of Evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), Perspectives of Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1.

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). Embedding Formative Assessment: Evaluation Report and Executive Summary. Education Endowment Foundation.

Wang, Y., et al. (2025). Explore the Prediction of Formative Assessment to Academic Success in Blended Learning: A Case Study. Proceedings of the 2024 International Conference on Artificial Intelligence and Future Education. https://doi.org/10.1145/3708394.3708427

Xu, Z., et al. (2026). Artificial Intelligence and Self-Regulated Learning: A Meta-Analysis. British Journal of Educational Technology.

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. Theory Into Practice, 41(2), 64–70.

Appendix A: SNCM 2.0 Algorithmic Specification (Pseudocode)

This appendix contains the full algorithmic specification of SNCM's two core procedures, for previous pseudocode for comparison see also the paper preceding this *(The Self/Non-Self Collaborative Methodology: SNCM v1.0 - Formal Iteration & Comparative Analysis - https://aixiv.science/abs/aixiv.260702.000002).* This code is provided for implementers and is not required reading for the article's argument in Parts I–III.

A.1 Structured Debate

```
function StructuredDebate(H0, ES, S, N):
  claim <- H0
  r <- 0
  while r < 3:
    candidate <- N.propose(claim, ES)
    act <- S.evaluate(candidate, ES)      # accept | reject | redirect
    log.append(timestamp, candidate, act, S.reason)
    if act == accept:
      return candidate, log
    elif act == redirect:
      claim <- S.redirect_claim
    r <- r + 1
  return claim, log   # terminates: r bounded by 3 (Sec. 11.1)
```

A.2 SNCL (Self/Non-Self Convergence Loop)

```
function SNCL(claim0, ES, S, N, MAX_ITER):   # MAX_ITER: externally enforced hard cap (Sec. 11.2)
  claim <- claim0
  i <- 0
  while i < MAX_ITER:
    synthesis <- N.synthesize(claim, ES)
    score <- IndependentRater.score(synthesis, ES)   # NOT S.score — see Sec. 8
    if score >= tau:                                 # tau from standard setting, Sec. 9/16
      return synthesis, i
    claim <- S.revise(claim, synthesis)
    i <- i + 1
  return claim, i   # cap reached without meeting tau: flagged, not silently accepted
```

Appendix B: SNCM 2.0 Instrumentation: Session Intake Fields, Evaluative Standard, and Scoring Rubrics

B.1 Session intake fields

- Session ID, timestamp, course, section, arm assignment (recorded but hidden from live raters)
- Evaluative Standard (ES): free-text statement authored by S before H0
- H0: S's initial position/claim
- Per-exchange log: candidate text (N), evaluative act (S), stated reason (S), timestamp
- Final synthesised artefact and SCQI sub-scores

B.2 SCQI sub-score rubric (0–4 anchors, abridged)

Sub-score	0 (absent)	2 (partial)	4 (strong)
Coherence	Internally contradictory or disjointed	Mostly consistent, minor gaps	Fully internally consistent argument
Evidentiary support	No support offered	Some support, not tied to ES	Support explicitly tied to ES criteria
Originality	Restates source/N verbatim	Minor rephrasing of N's content	Genuine S-originated synthesis beyond N's proposal
ES alignment	Unrelated to stated ES	Partially addresses ES	Fully addresses all stated ES criteria

Table B1. Abridged SCQI rubric anchors; full anchor set with worked exemplars is maintained in the rater training manual (Appendix C).

B.3 Oral defensibility interview rubric (0–4)

- 0–1: No coherent justification, or justification contradicts the session log
- 2: Justification present but does not reference the stated ES
- 3: Justification references ES and is internally coherent
- 4: Justification references ES, is coherent, and is non-circular (does not simply restate the decision as its own justification)

Appendix C: Independent Rater Training Manual (Abridged)

- Session 1 (2h): Introduction to SNCM session logs; walkthrough of Table 2 and Appendix B rubrics using 5 worked exemplars per SCQI anchor level.
- Session 2 (2h): Practice scoring on 10 archived pilot sessions (Section 20 pilot phase), independent then group discussion of disagreements.
- Session 3 (2h): Calibration scoring on the 20-session double-coded set (Section 15, step 2); compute κ/ICC; retrain on disagreement cases if gate not met.
- Ongoing: 15% double-coding throughout main data collection; monthly drift-check meeting reviewing any session where two raters differ by $\geq$2 rubric points.
- Oral interview raters additionally trained on a fixed interview script (randomised selection of 3 evaluative decisions per session, standardised prompts, no leading follow-ups).

Appendix D: Standard-Setting Session Materials

- Panel composition: minimum 6 subject-matter experts per course/ES combination, distinct from Appendix C raters.
- Materials: 30–50 pre-scored exemplar syntheses spanning the full SCQI range, from the pilot phase (Section 20).
- Bookmark procedure: exemplars ordered by ascending SCQI score; each panellist independently marks the boundary between 'acceptable without further debate' and 'requires an additional debate round'; boundaries averaged to derive τ, with panel range/SD reported alongside τ as an uncertainty band.
- Angoff alternative (used when <30 pre-scored exemplars exist): panellists independently estimate, per exemplar, the probability a borderline-competent learner's synthesis would be judged acceptable; τ is the mean of panellist estimates at the target performance level.
- Re-derivation requirement: τ is re-derived for each new course, ES, or learner population; no cross-context reuse of a previously derived τ without re-running this procedure.