Maurya M. The Confusion Fingerprint Index: Cognitive error taxonomy as an adaptive learning signal in learners. EdArXiv preprint, 2026. doi: https://doi.org/10.35542/osf.io/cr53e_v1 · Licensed under CC BY 4.0.

Does Knowing *Why* a Student Fails Matter More Than Knowing That They Did?

The Confusion Fingerprint Index: Cognitive Error Taxonomy as an Adaptive Learning Signal in Learners

Manik Maurya **[ORCID iD: 0009-0005-3554-693X]**

Department of Cognitive Science, Indian Institute of Technology (IIT) Kanpur, Kanpur 208016, Uttar Pradesh, India

Correspondence: manikmaurya.in@gmail.com · Research Intern, SAC Lab, Dept. of Cognitive Science, IIT Kanpur

Pre-registration: OSF.io Registries [DOI: https://doi.org/10.17605/OSF.IO/8PJQH] · AEA RCT Registry [ID: to be assigned] · Ethics: IIT Kanpur Human Ethics Committee [clearance required prior to data collection]

Protocol reported in accordance with SPIRIT 2013 (Chan et al., 2013). The trial is to be reported per CONSORT 2010 and the CONSORT pilot and feasibility extension (Eldridge et al., 2016). Intervention described per TIDieR (Hoffmann et al., 2014).

Note on authorship and research trajectory. *This protocol paper is solo-authored at the preprint and pre-registration stage to establish priority on the theoretical construct (the Confusion Fingerprint Index) and the trial design. The empirical phase will be conducted under formal institutional supervision; the manuscript reporting trial results, intended for peer-reviewed submission, may include additional co-authorship reflecting empirical and supervisory contributions. The Cognivia platform, the CFI construct, and the protocol reported herein are the intellectual contribution of the sole present author.*

Background. Standard adaptive learning systems treat student errors as binary signals—correct or incorrect—and adjust instructional difficulty accordingly. This correctness-binary paradigm, dominant since Bayesian Knowledge Tracing was formalised by Corbett and Anderson (1995), conflates mechanistically distinct cognitive processes under a single label: *wrong.* The cognitive and learning sciences distinguish, at minimum, Recall Failure, Partial Knowledge, Confabulation, and Interference—error types that arise from different memory mechanisms, carry different diagnostic information, and respond to different instructional interventions.

Objective. This pre-registered randomised pilot trial pursues two interrelated aims: (i) to test whether error-taxonomy-aware adaptive learning—in which the cognitive type of each incorrect response governs subsequent question selection and targeted Hindi-language remediation—produces superior 2-week delayed retention compared to correctness-only adaptive learning and standard classroom instruction; and (ii) to introduce and validate the Confusion Fingerprint Index (CFI), a novel individual-difference construct capturing each learner's characteristic distribution of cognitive error types.

Method. Ninety students (Classes 6–8; ages 11–14 years) enrolled in three government junior high schools in Kanpur Dehat district, Uttar Pradesh, India, will be randomly assigned—stratified by school and class level—to: (A) error-taxonomy-aware Cognivia adaptive learning with differentiated Hindi remediation (n = 30); (B) correctness-only Cognivia adaptive learning (n = 30); or (C) standard classroom instruction (n = 30). Paper-based NCERT Science assessments—pre-test, immediate post-test, and 2-week delayed retention test—constitute the primary and secondary outcome measures.

Expected Contribution. This study constitutes the first pre-registered empirical test of cognitive error taxonomy as a real-time adaptive sequencing signal, the first validation of the Confusion Fingerprint as a retention-predictive individual-difference construct, and the first pre-registered randomised adaptive learning trial conducted with rural, Hindi-medium, government-school learners in India—a population of approximately 250 million children that is structurally absent from the extant evidence base.

Keywords: *adaptive learning · cognitive error taxonomy · Confusion Fingerprint Index · metacognitive calibration · knowledge tracing · spaced repetition · rural India · educational technology · randomised pilot trial · Cognivia · desirable difficulties · hypercorrection*

INTRODUCTION

Error is not a homogeneous event. When a student incorrectly answers a question, at least four categorically distinct cognitive situations may be responsible: the student may lack any encoded representation of the target concept; she may possess a partial, directionally correct mental model that is not yet fully developed; she may hold a coherent and confidently maintained misconception that actively interferes with the correct response; or she may retrieve a well-encoded concept from an adjacent topic, misapplying it through inter-concept confusion. These situations share a surface form—the wrong answer—but diverge in their underlying memory structures, in the type of corrective feedback most likely to produce durable learning, and in the long-term retention trajectories they predict. Adaptive educational technology, used by hundreds of millions of learners worldwide, treats them identically.

The dominance of the correctness-binary approach to knowledge modelling is traceable to Corbett and Anderson's (1995) formalisation of Bayesian Knowledge Tracing (BKT), which estimates a latent binary knowledge state—learned or not learned—from sequences of correct and incorrect responses. BKT's scientific influence has been substantial: Wang et al.'s (2024) meta-analysis of 156 randomised studies estimates a mean

effect of d = 0.52 for adaptive versus non-adaptive instruction. Two decades of extension work—from Performance Factor Analysis (Pavlik et al., 2009) through Deep Knowledge Tracing (Piech et al., 2015), memory-augmented networks (Zhang et al., 2017), self-attentive knowledge tracing (Pandey and Karypis, 2019), and context-aware attention models (Ghosh et al., 2020)—has refined the predictive machinery without altering its foundational assumption: the incorrect response is a single, unanalysed evidential category. The question motivating this study is whether enriching that signal—by identifying the cognitive type of the error, not merely its occurrence—produces meaningfully better learning outcomes.

The errorful learning literature provides a principled basis for expecting that it should. Three decades of experimental work established that not all errors carry the same informational value for subsequent memory (Metcalfe, 2017). High-confidence errors produce a hypercorrection effect—they are corrected more dramatically than low-confidence errors immediately following feedback (Metcalfe and Finn, 2011)—yet the incorrect memory trace underlying a confident wrong answer is not extinguished; it competes with the corrected trace at delayed retention tests (Butterfield and Metcalfe, 2001). Unsuccessful retrieval attempts, by contrast, paradoxically strengthen subsequent memory for the target through the generation of desirable difficulties (Bjork, 1994; Bjork et al., 2013). And inter-concept interference errors—arising when two related, well-encoded concepts compete for retrieval—are responsive to discrimination practice rather than re-encoding (Kornell and Bjork, 2008). Each error type, in short, calls for a theoretically distinct instructional response. No adaptive system has implemented these prescriptions in real time, and none has subjected them to prospective empirical test.

A concurrent theoretical development makes empirical validation particularly timely. Verma (2025) recently formalised the EDGE framework—a Bayesian state-space model for misconception-aware adaptive scheduling—demonstrating mathematically that error-type information should improve adaptive sequencing decisions under a range of plausible generative models, and explicitly identifying the absence of empirical validation as the primary limitation. The present study is designed to provide that validation in the most ecologically relevant setting available.

The population in which this validation is conducted is itself a scientific contribution. Systematic reviews of adaptive learning efficacy (Wang et al., 2024; Gligorea et al., 2023) document a near-total absence of evidence from low- and middle-income countries. India, where approximately 250 million children attend government primary and secondary schools—many in rural areas characterised by large class sizes, shared device access, intermittent connectivity, and Hindi-medium instruction—has produced no pre-registered randomised trial of any adaptive learning platform. These contextual features are not merely implementation obstacles; they are moderators of adaptive learning efficacy whose estimation requires a study designed for and conducted within this population.

The present study makes three separable scientific contributions. First, it provides the first pre-registered, field-experimental test of cognitive error taxonomy as a real-time adaptive sequencing signal. Second, it introduces and validates the Confusion Fingerprint Index (CFI) as a novel individual-difference construct—a learner-level profile of characteristic error types, computed from early session logs, that is hypothesised to predict 2-week delayed retention independently of overall accuracy. Third, it generates the first pre-registered estimates of adaptive learning efficacy in a rural, Hindi-medium, government-school context, with full open data enabling replication and meta-analysis.

THEORETICAL BACKGROUND

A Cognitive Taxonomy of Learning Errors

The proposal that qualitatively distinct error types carry differential diagnostic information is grounded in both cognitive science and science education research, though the two traditions have not previously been integrated into an adaptive sequencing framework. From the cognitive science side, the errorful learning literature (Metcalfe, 2017; Potts and Shanks, 2014) establishes that errors followed by corrective feedback enhance long-term memory for the correct response, and that this benefit is modulated by the nature of the error. From the science education side, research on conceptual change (Vosniadou, 1994; Chi, 1992) demonstrates that learners enter formal instruction with prior knowledge frameworks that actively resist replacement—a phenomenon documented across physics, biology, chemistry, and mathematics, and directly relevant to the NCERT Science curriculum used in the present study (Treagust, 1988).

We propose a four-category taxonomy grounded in the intersection of these two literatures. **Recall Failure (RF)** denotes the production of a response that is random, implausible, or explicitly uncertain, indicating the absence of any consolidated memory representation of the target concept. The remedial implication, consistent with the generation effect literature (Roediger and Karpicke, 2006), is initial encoding: the learner requires re-exposure at reduced difficulty, not correction of an existing trace. **Partial Knowledge (PK)** denotes a response that is directionally correct but insufficiently specific, indicating an incomplete mental model. The remedial implication is scaffolded progression: the learner requires a question that targets the missing conceptual component. **Confabulation (CF)** denotes the selection of a response corresponding to a documented misconception—a coherent, plausible, and confidently held incorrect belief. The hypercorrection literature (Metcalfe and Finn, 2011; Butterfield and Metcalfe, 2001) establishes that CF-type errors, though dramatically corrected in the immediate post-feedback period, produce incorrect memory traces that are not extinguished but rather weakened and displaced, and that can resurface as competing responses at delayed retention tests. This dynamic, formalised in the memory reconsolidation literature (Nader et al., 2000), requires an explicit contrast-and-correct intervention: a targeted explanation of why the chosen response is incorrect paired with a follow-up probe of the corrected concept. **Interference (INT)** denotes the retrieval of a correctly encoded

concept from an adjacent topic studied in the same session, applied to a question where it does not belong. INT errors indicate that both competing concepts are encoded, but cannot yet be discriminated under retrieval pressure—a situation for which Kornell and Bjork's (2008) interleaving research prescribes discrimination practice.

Cognitive Mechanisms and Dual-Process Underpinnings

The four error types map onto distinct positions in Sweller's (1988) cognitive load framework. RF errors arise when intrinsic load exceeds working memory capacity, leaving no resources available for schema formation; CF errors, by contrast, reflect a fully loaded and confidently executed schema that is nonetheless incorrect. The implication for instructional design is that RF and CF errors require opposite responses along the difficulty dimension: difficulty reduction for RF, difficulty maintenance with conceptual contrast for CF. Chi and Wylie's (2014) ICAP framework provides a complementary lens: RF errors suggest passive or inactive engagement with the target concept, while CF errors suggest active but misdirected engagement with a competing schema. The adaptive implication is that error-type-specific remediation addresses the underlying engagement mode, not merely the surface-level correctness signal.

Feedback theory provides further theoretical grounding. Kluger and DeNisi's (1996) feedback intervention theory distinguishes feedback directed at task-level performance, process-level strategies, and self-regulatory processes, demonstrating that task- and process-level feedback produces stronger learning gains than outcome-only feedback. Hattie and Timperley's (2007) elaboration identifies the explanatory richness of corrective feedback—not merely whether feedback is given, but what it communicates about the nature and source of the error—as the primary moderator of feedback effectiveness. Black and Wiliam's (1998) foundational analysis of formative assessment arrives at the same conclusion from the classroom research tradition: feedback that identifies the gap between current and target understanding is categorically more effective than feedback that reports only correctness. The present intervention implements this principle at the error-type level, matching the explanatory content of remediation to the cognitive mechanism of the error.

The Confusion Fingerprint as an Individual-Difference Construct

Individual learners do not distribute their errors randomly across the four error types. A student whose errors are predominantly confabulations occupies a systematically different epistemic position from one whose errors are predominantly recall failures, even when their overall accuracy—and therefore their position on a standard adaptive difficulty curve—is identical. We propose the Confusion Fingerprint as a construct that captures this difference: the characteristic distribution of cognitive error types exhibited by a learner across a domain and an early time window. The Confusion Fingerprint Index (CFI) operationalises this construct as a three-component vector computed from session logs across sessions 1–4 (Table 1).

Table 1. Components of the Confusion Fingerprint Index (CFI), computed from session logs for sessions 1–4 of the Cognivia intervention.

Component	Formula	Theoretical Interpretation
CFI-CF	CF ÷ total errors (sessions 1–4)	Confabulation index. Proportion of errors arising from active misconceptions. Predicted negative predictor of 2-week retention (Metcalfe and Finn, 2011).
CFI-INT	INT ÷ (INT + RF) [0 if denom. = 0]	Interference index. Cross-concept confusion relative to pure encoding absence. Predicted to require discrimination practice (Kornell and Bjork, 2008).
CFI-RF	RF ÷ total errors (sessions 1–4)	Encoding-absence index. Predicted weaker negative predictor of retention than CFI-CF, because RF errors generate no competing memory trace.

The theoretical prediction follows directly from the errorful learning literature. High CFI-CF scores should negatively predict delayed retention because confabulated responses generate incorrect memory traces that resist correction and compete with the target response at retrieval (Butterfield and Metcalfe, 2001). High CFI-RF scores should be a weaker predictor because recall failures create no competing trace; standard spaced repetition is expected to be sufficient. High CFI-INT scores should predict intermediate outcomes: interference errors signal adequate encoding of two competing concepts, a problem more amenable to targeted discrimination practice than to pure re-encoding.

The diagnostic value of the CFI is separable from the adaptive mechanism it was designed to support. Even in the absence of an adaptive platform, a classroom teacher with access to a learner's CFI profile during formative assessment would possess actionable information about which students require misconception correction rather than re-instruction—a distinction that the correctness signal alone cannot provide. The CFI thus constitutes a diagnostic instrument of potential value across instructional contexts.

Knowledge Tracing: From BKT to Attention-Based Models

Corbett and Anderson's (1995) BKT model estimates a latent binary knowledge state using four parameters: prior probability of knowledge, probability of learning, slip probability given knowledge, and guess probability given non-knowledge. BKT has demonstrated predictive validity in intelligent tutoring system studies (VanLehn, 2011) and has motivated a productive family of extensions. Performance Factor Analysis (Pavlik et al., 2009) distinguishes correct from incorrect practice trials while retaining a correctness-binary input signal. Deep Knowledge Tracing (Piech et al., 2015) replaced the hidden Markov model with long short-term memory networks, improving next-response prediction accuracy but reducing interpretability and introducing data-leakage risks documented by Liu et al. (2022).

Memory-augmented networks (Zhang et al., 2017), self-attentive architectures (Pandey and Karypis, 2019), and context-aware attention models (Ghosh et al., 2020) have extended the representational capacity of deep KT further, but have not altered the foundational assumption: the incorrect response remains a single, unanalysed evidential category across the entire lineage from BKT to AKT.

The Cognivia system employed in this study uses a modified SM-2 spaced repetition algorithm (Wozniak, 1990) as its core scheduling engine. SM-2 was chosen over deep KT architectures for three reasons specific to this deployment context: (i) interpretability and auditability in a low-resource setting where model debugging requires transparency; (ii) low computational overhead compatible with shared-device, low-bandwidth operation; and (iii) established psychometric validity for spaced-practice scheduling. The error-type classification layer operates as an additional signal above the SM-2 scheduler, enriching the adaptive input without requiring a neural architecture—a design choice that maximises interpretability and reproducibility.

Metacognitive Calibration

Metacognitive monitoring—the capacity to assess the accuracy of one's own knowledge and to regulate learning accordingly—develops substantially during middle childhood and early adolescence (Flavell, 1979; Veenman et al., 2006). Students who are well-calibrated—whose subjective confidence tracks their actual accuracy—allocate study effort more effectively and show stronger learning outcomes (Dunlosky and Rawson, 2012; Hacker et al., 2008). Confabulation, by definition, constitutes a calibration failure: the student is maximally confident in a response that is incorrect. Error-taxonomy-aware adaptive instruction that explicitly targets the overconfidence implicit in confabulation—by informing students not only that their response is wrong but precisely *why* it is wrong and why the correct response is right—should, on this account, improve metacognitive calibration as measured by the Brier score (Brier, 1950), a proper scoring rule for probabilistic confidence judgements. Carpenter et al.'s (2019) demonstration that adaptive training improves Brier scores in adult samples provides direct empirical precedent; the present study extends this prediction to early adolescent learners in a low-resource setting.

The Population Context: Rural India as Scientific Contribution

The external validity of the adaptive learning evidence base is severely constrained by its sampling frame. Wang et al.'s (2024) analysis of 156 randomised studies finds that high-income, English-medium, post-secondary, individually-accessed populations account for the substantial majority of the evidence. India, home to the world's second-largest student population and to approximately 250 million children enrolled in government primary and secondary schools, has produced no pre-registered randomised controlled trial of any adaptive learning platform. The contextual features of this population—shared device access, intermittent network connectivity, teacher-administered rather than student-administered sessions, Hindi-medium instruction, and large within-class heterogeneity in prior knowledge—are not merely implementation constraints but potential moderators of adaptive learning efficacy that cannot be estimated from the existing evidence base. This study generates the first prospectively registered estimates of these effects, contributing to the generalisability literature regardless of the direction of its findings.

STUDY AIMS AND PRE-REGISTERED HYPOTHESES

This study pursues two primary aims: (1) to test whether error-taxonomy-aware adaptive learning produces superior delayed retention relative to correctness-only adaptive learning and standard instruction; and (2) to test whether the Confusion Fingerprint Index, computed from early session logs, predicts 2-week delayed retention above and beyond overall session accuracy. The study is designed around four pre-registered hypotheses, registered at OSF.io Registries prior to data collection. The first three are confirmatory; the fourth is explicitly exploratory and registered as such. All four will be reported in full regardless of outcome direction, consistent with the registered reports model (Chambers, 2013; Nosek et al., 2018). Effect sizes will be reported for all analyses irrespective of statistical significance.

Table 2. Pre-registered hypotheses.

Hypothesis	Pre-Registered Prediction
H1 — Primary, Confirmatory	Students in Group A (error-taxonomy-aware) will show significantly higher 2-week delayed retention than Group B (correctness-only adaptive) and Group C (standard instruction), after controlling for pre-test performance (ANCOVA; Bonferroni-corrected $\alpha = .025$ per planned contrast). *Sub-hypotheses:* H1a: Group A > Group B (mechanism contrast); H1b: Group B > Group C (replicating the basic adaptive effect in this population).
H2 — Secondary, Confirmatory	CFI-CF, CFI-INT, and CFI-RF (computed from sessions 1–4) will significantly predict 2-week retention above and beyond pre-test score and overall session accuracy (hierarchical multiple regression; F-change for Block 2, $p < .05$). CFI-CF is predicted to be a significant negative predictor (H2a). A moderation analysis will test whether the CFI-CF $\rightarrow$ retention relationship is attenuated in Group A relative to Group B, providing mechanistic evidence that error-taxonomy-aware adaptation targets confabulation persistence (H2b).
H3 — Secondary, Confirmatory	Group A will exhibit lower Brier scores (better metacognitive calibration) at the post-test than Groups B and C (one-way ANOVA, $p < .05$; Tukey HSD post-hoc). A secondary analysis will test whether Brier score improvement (post − pre) correlates positively with 2-week retention across the full sample (Pearson r).
H4 — Exploratory, Registered	CFI-CF from sessions 1–4 will correlate positively with CFI-CF from sessions 5–6 in Group B (construct stability absent targeted adaptation), and this correlation will be lower in Group A (construct responsiveness to error-taxonomy-aware intervention). This hypothesis is registered as exploratory and will not be used to confirm or disconfirm primary conclusions.

METHOD

This protocol is reported in accordance with SPIRIT 2013 (Chan et al., 2013). The completed SPIRIT checklist will be deposited at the OSF pre-registration record. The trial will be reported per CONSORT 2010 and the CONSORT extension to pilot and feasibility trials (Eldridge et al., 2016) upon completion of data collection. The intervention is described per TIDieR (Hoffmann et al., 2014) in Table 3.

Study Design

A three-arm, parallel-group, pre-registered randomised pilot trial with three assessment occasions (pre-test, immediate post-test, and 2-week delayed retention test). Randomisation is stratified by school and class level. The study is powered as a pilot trial for effect-size estimation rather than confirmatory efficacy testing, consistent with the framework of Lakens et al. (2018). Pre-specified feasibility outcomes include: recruitment rate (at least 80% of eligible students consented), session attendance rate (at least 80% of scheduled sessions attended per student), retention test completion rate (at least 85% of enrolled students completing the primary outcome), and equipment failure rate (no more than 5% of sessions affected).

Participants

Ninety students enrolled in Classes 6, 7, and 8 (ages 11–14 years) at three government junior high schools in Kanpur Dehat district, Uttar Pradesh, India, will be recruited for participation. Schools were identified through the investigator's established working relationships arising from a five-year rural education initiative across forty-plus schools in the district. Thirty participants will be enrolled per school, with approximately equal representation of class levels and an approximately equal gender distribution within each school.

Inclusion criteria: Class 6, 7, or 8 enrolment; written informed consent from a parent or guardian; verbal student assent at session 1.

Exclusion criteria: Fewer than 70% session attendance (minimum 13 of 18 sessions) for Groups A and B; absence from the 2-week retention test without a successful make-up within seven days; school transfer during the intervention.

Sample size and power: The primary contrast (Group A vs. Group C) is powered at a conservative $d = 0.40$, reflecting uncertainty about the incremental benefit of error-type classification over baseline adaptive instruction, below the meta-analytic mean of $d = 0.52$ (Wang et al., 2024). With $n = 30$ per group, an ANCOVA including pre-test as covariate, and a Bonferroni-corrected $\alpha = .025$, estimated power $(1 - \beta) = 0.76$, computed using G*Power 3.1. This power level is appropriate for a pilot trial whose primary purpose is effect-size estimation to inform a subsequent, adequately powered confirmatory trial (Lakens et al., 2018).

Ethics and Pre-Registration

Ethical clearance for this study will be sought from the IIT Kanpur Human Ethics Committee before any data collection begins. No data will be collected prior to receiving written approval; receipt of clearance will be reported in the final manuscript. Written informed consent will be obtained from the parent or guardian of each participant prior to enrolment; student verbal assent will be sought at the first session in Hindi:

‘*Kya aap is study mein participate karna chahte hain?*’ Non-participation will have no effect on academic standing. All data will be pseudonymised at the point of collection using alphanumeric codes (S001–S090); no names will appear in any analysis file or supplementary material.

The study will be pre-registered at OSF.io Registries before any data collection commences, using the OSF Preregistration template for randomised controlled trials. The pre-registration will specify all hypotheses, the complete analysis plan, feasibility outcomes, exclusion criteria, and stopping rules as detailed in this manuscript. Any deviations from the pre-registered plan will be explicitly flagged as amendments in the final manuscript.

Intervention Description (TIDieR)

Table 3. Intervention description per TIDieR (Hoffmann et al., 2014).

TIDieR Item	Group A: Error-Taxonomy-Aware	Group B: Correctness-Only
Brief name	Cognivia adaptive learning — error-taxonomy-aware	Cognivia adaptive learning — correctness-only
Why (rationale)	Error type determines memory mechanism; taxonomy-matched remediation targets the specific cognitive failure mode (Metcalfe, 2017; Kluger and DeNisi, 1996).	Correctness signal governs difficulty adjustment; no error-type information used (Corbett and Anderson, 1995).
What (materials)	240-item NCERT Science bank; 80 Hindi error-type-matched remediation tips (20/domain, 5/error type); Cognivia web platform.	Same 240-item bank; no remediation tips; Cognivia platform in correctness-only mode.
Who provided	Investigator (Manik Maurya); questions delivered verbally in Hindi; responses entered by investigator.	Same investigator; identical verbal delivery protocol.
How (mode)	Researcher-administered, group setting (5–6 students), seated separately; verbal question-and-answer format.	Identical delivery mode.
Where	Government junior high school classroom, Kanpur Dehat, Uttar Pradesh.	Same schools; separate sessions from Group A.
When/dose	18 sessions over 6 weeks (3/week); 45 min/session; 12–15 questions/session.	Identical schedule and session structure.
Tailoring	Next question type (RF/PK/CF/INT logic) and tip content tailored per error type per student per item.	Next question difficulty tailored by SM-2 correctness history only.
Fidelity	Standardised script (Hindi/English); 20% audio-recorded; 10% log-cross-checked.	Same fidelity procedures; blind to Group A protocol.

The Cognivia Platform and Error Classification

Cognivia is an open-source AI-assisted adaptive learning platform developed by the investigator. It is implemented as a progressive web application with a React/Vite/TypeScript frontend (Vercel), a Flask 3.x REST backend (Render), and a Supabase PostgreSQL data layer with row-level security on all research tables. The deployment is optimised for low-bandwidth operation, shared devices, and intermittent connectivity, and is installable as a native-feeling application across Android, iOS, Windows, and macOS. The platform is publicly accessible at *cognivia.vercel.app*; the research instrument is gated behind a researcher login. A WhatsApp-compatible interface (via Twilio) enables asynchronous access where individual device access is available. The core scheduling engine implements a modified SM-2 algorithm (Wozniak, 1990). For this study, four additional research-instrument components were implemented in a dedicated, production-isolated module (Table 3, above): the Adaptive Question Intelligence (AQI-1.0) rule-based error classifier; the Adaptive Tip Engine (ATE-1.0) driven by a modified UCB1 exploration-exploitation algorithm (Auer et al., 2002); a per-question Research Logger; and the CFI Engine, which computes CFI components automatically upon completion of sessions 4 and 6.

Each of the 240 questions in the NCERT Science question bank has four response options. At question design time, each incorrect option was assigned an error-type label (RF, PK, CF, or INT) through a two-source systematic analysis: (i) documented misconceptions in Indian science education research (Treagust, 1988), and (ii) a logical analysis of the cognitive state presupposed by each distractor. Pre-study inter-rater reliability was evaluated by an independent second rater, blind to the original classifications, who classified a stratified random sample of 40 distractor pairs (10 per domain). A pre-specified threshold of Cohen's kappa of 0.70 or higher was required for proceeding to main study enrolment.

Question Bank

The study employs a purpose-built 240-item NCERT Science question bank spanning four topic domains: Light (Class 8), Nutrition in Plants (Class 7), Nutrition in Animals (Class 7), and Motion and Time (Class 7). Domains were selected for high misconception density (enabling CF-type distractor construction), high examination salience, and sufficient conceptual independence to generate INT-type errors when adjacent domains appear in the same session. The bank comprises 120 intervention items (30 per domain) and 120 assessment items distributed across three parallel forms (pre-test, post-test, retention test; 40 items per form, 10 per domain) matched for topic coverage and difficulty (pre-test M = 2.42, SD = 0.89; post-test M = 2.50, SD = 0.87; retention M = 2.48, SD = 0.91; five-point scale). No item appears in more than one form.

Randomisation and Allocation Concealment

Randomisation will be conducted by the investigator after pre-test completion and before the first intervention session, using a computer-generated sequence (Python 3.11 random.shuffle, seed recorded and deposited at OSF) stratified by school and class level. Within each stratum, students will be assigned cyclically to Groups A, B, and C in a 1:1:1 ratio. The assignment sequence is generated and stored in a sealed file opened only after all pre-test data are entered. Classroom teachers will not be informed of group assignment. Students will be informed only that they are participating in a ‘Cognivia learning study’; they will not be told

their assigned condition.

Procedure

Pre-test (Week 0). A 30-item paper-based NCERT Science assessment (30 min) administered simultaneously to all 90 participants at each school. Verbal confidence ratings (1 = not sure, 2 = somewhat sure, 3 = very sure; administered in Hindi) collected after 10 randomly selected items establish baseline metacognitive calibration.

Intervention (Weeks 1–6; Sessions 1–18). Groups A and B complete 18 Cognivia sessions (3 sessions/week, 45 min/session) in groups of 5–6, seated separately. Each session: (i) 5-min warm-up review; (ii) 12–15 questions per student administered per group-specific adaptive logic; (iii) 3-min reflection; (iv) 2-min preview. After each incorrect response, Group A receives a Hindi remediation tip matched to the classified error type (ATE-1.0); Group B receives no tip. Group C receives standard NCERT instruction from assigned teachers; no platform access.

Post-test (Week 6). All 90 participants complete a 30-item parallel-form paper assessment on the final intervention day, with full confidence ratings on all items. A manipulation check is administered to Groups A and B immediately afterwards.

Retention test (Week 8; Primary Outcome). All 90 participants complete a third parallel-form paper assessment administered without advance notice, 14 days after the post-test. Full confidence ratings collected. Absent students are visited again within seven days; those still absent are excluded per the pre-registered exclusion protocol.

Outcome Measures

Primary outcome: 2-week delayed retention score (percentage correct on the 30-item retention test, Week 8). This is the sole primary outcome measure, pre-specified as such at OSF prior to data collection.

Secondary outcomes: (i) Confusion Fingerprint Index (CFI-CF, CFI-INT, CFI-RF), computed from session logs of sessions 1–4 for Groups A and B; (ii) immediate learning gain (post-test − pre-test, percentage points); (iii) Brier score, computed per student per assessment occasion as $\text{mean}[(\text{confidence}_{\text{prob}} - \text{accuracy})^2]$, where $\text{confidence}_{\text{prob}} \in \{0.0, 0.5, 1.0\}$ for ratings {1, 2, 3} and accuracy $\in \{0, 1\}$; lower Brier scores indicate better calibration.

Process measure: Structured field journal entries after each school visit, to be analysed thematically (Braun and Clarke, 2006) following the Fixsen et al. (2005) implementation science framework.

Statistical Analysis Plan

All analyses are pre-registered at OSF prior to data collection. Software: JASP 0.17.2 (open-source). All tests two-tailed; $\alpha = .05$ unless Bonferroni correction is specified. Effect sizes reported for all analyses irrespective of significance.

H1: One-way ANCOVA; pre-test as covariate; three planned contrasts (Bonferroni $\alpha = .025$): A vs. C, A vs. B, B vs. C. Effect sizes: partial η^2 (overall model), Cohen's *d* (pairwise). Assumption checks: Levene's test; homogeneity of regression slopes. Violation of the latter triggers Welch's heteroscedastic ANCOVA as sensitivity analysis.

H2: Two-block hierarchical multiple regression. Block 1: pre-test score and overall session accuracy. Block 2: CFI-CF, CFI-INT, CFI-RF (entered simultaneously). Primary test: *F*-change for Block 2. Standardised β with 95% CI for each predictor; VIF inspected for collinearity. H2b moderation: CFI-CF × Group interaction term.

H3: One-way ANOVA; Tukey HSD post-hoc. Pearson *r*: Brier change (post − pre) vs. retention score (full sample).

H4 (exploratory): Pearson correlations, early vs. late CFI-CF, within Groups A and B separately. Mixed ANOVA: block (early/late) × Group (A/B) interaction with CFI-CF as dependent variable.

Missing data: Complete-case analysis as primary approach. Last-observation-carried-forward (post-test score as proxy retention) as pre-specified sensitivity analysis. If differential attrition exceeds 10 percentage points between conditions, a baseline characteristic comparison between completers and dropouts will be reported.

ANTICIPATED RESULTS

Data collection will commence upon receipt of IIT Kanpur Human Ethics Committee approval. Results will be reported in full in a companion publication, in strict accordance with the pre-registered analysis plan. The following subsections describe the anticipated reporting structure; all tables and figures shown here are structural placeholders.

Participant Flow and Baseline Characteristics

Participant flow per CONSORT 2010 (Schulz et al., 2010) will be reported in Table 4, capturing enrolment, eligibility screening, randomisation, exclusions, and analysed sample sizes. Baseline equivalence across groups (pre-test mean, class distribution, sex, school) will be verified using one-way ANOVA on pre-test scores and reported alongside Table 4.

Stage of trial	Group A	Group B	Group C	Total
Assessed for eligibility	—	—	—	[n]
Excluded (criteria/refusal)	—	—	—	[n]
Randomised	30	30	30	90
Discontinued intervention	[n]	[n]	n/a	[n]
Lost to retention test	[n]	[n]	[n]	[n]
Analysed (primary outcome)	[n]	[n]	[n]	[n]

Table 4. Anticipated participant flow per CONSORT 2010 (Schulz et al., 2010). Counts to be inserted upon trial closure. A diagrammatic CONSORT flow chart will accompany the results manuscript.

Primary Outcome — H1

The primary contrast (Group A vs. Group C) is anticipated in the range d = 0.40–0.65, calibrated against the meta-analytic mean for adaptive vs. non-adaptive instruction (d = 0.52; Wang et al., 2024) and conservatively adjusted for the low-resource, researcher-administered context. The mechanism contrast (Group A vs. Group B) is expected to be smaller and constitutes the theoretically decisive test. Results will be reported in Table 5.

Group	n	Adj. M (%)	SE	95% CI	d (vs. C)
A — Error-taxonomy-aware	[n]	[M]	[SE]	[L, U]	[d]
B — Correctness-only	[n]	[M]	[SE]	[L, U]	[d]
C — Standard instruction	[n]	[M]	[SE]	[L, U]	—

Table 5. *Anticipated reporting template for the primary outcome (H1).* Adjusted means and 95% CIs from one-way ANCOVA controlling for pre-test score. Cohen's *d* computed against Group C. Planned contrasts: A vs. C, A vs. B, B vs. C, with Bonferroni correction (α = .025).

CFI Predictive Validity — H2

If CFI-CF significantly predicts 2-week retention above and beyond overall session accuracy (Block 2 F-change < .05), the Confusion Fingerprint achieves status as a genuine individual-difference construct: two learners with identical global accuracy but divergent CFI-CF profiles would carry meaningfully different expected retention trajectories, with implications for low-cost diagnostic practice in any classroom with access to formative assessment data. Results will be reported in Table 6.

Predictor	β	95% CI	*t*	*p*	VIF
Block 1 (R^2 = […])					
Pre-test score	[β]	[L, U]	[t]	[p]	[VIF]
Overall session accuracy	[β]	[L, U]	[t]	[p]	[VIF]
Block 2 (ΔR^2 = […], *F* change = […])					
CFI-CF	[β]	[L, U]	[t]	[p]	[VIF]
CFI-INT	[β]	[L, U]	[t]	[p]	[VIF]
CFI-RF	[β]	[L, U]	[t]	[p]	[VIF]

Table 6. *Anticipated reporting template for H2.* Two-block hierarchical regression predicting 2-week delayed retention. Standardised β with 95% CI. Block 2 F-change is the primary inferential test. CFI-CF is predicted to be a significant negative predictor.

DISCUSSION

Overview and Interpretive Framework

The Discussion section of the results manuscript will be organised around the pre-specified interpretive framework deposited at OSF.io prior to data collection. This pre-specified framework, reproduced here, ensures that result-contingent interpretations are not post-hoc constructions fitted to obtained outcomes. We describe, for each of the three primary outcome scenarios, the interpretation we are committed to reporting.

Interpreting the Primary Outcome (H1)

A significant Group A > Group C contrast in 2-week retention, combined with a significant Group A > Group B mechanism contrast, would constitute unusually clean evidence that error-type classification—independent of screen time, content exposure, session structure, and investigator contact—drives a retention benefit. The experimental isolation of the mechanism is a distinctive design feature: Groups A and B are equated on all elements of the intervention except what the system does with a wrong answer. This design choice was deliberate, following the logic that mechanism isolation, rather than treatment-package efficacy, is the scientifically informative comparison at this stage of the research programme.

If Group A outperforms Group C but not Group B (Scenario 2), the most parsimonious interpretation is that the basic adaptive learning effect generalises to rural India, but that error-type classification does not add measurable value at n = 30 per group or within a six-week window. This finding would be substantively important: it would provide the field's first pre-registered estimate of whether the basic adaptive learning effect—consistently positive in high-resource, English-medium, individually-accessed settings—generalises to one of the world's largest and most structurally underserved learner populations.

A null result across both contrasts (Scenario 5) would itself constitute a genuine scientific contribution. A pre-registered null, with full open data, deposits the field's first reliable estimate of the boundary conditions of adaptive learning efficacy in rural India, and generates hypotheses about which contextual moderators—device-sharing, investigator-administration, within-session group format, Hindi-medium delivery, NCERT curriculum structure—may attenuate the standard effect. Chambers' (2013) argument for registered reports applies here with particular force: the informativeness of a result is determined by the quality of the design, not its direction.

The Confusion Fingerprint Construct (H2)

The theoretical prediction grounding H2 is specific and mechanistically derived. Metcalfe and Finn's (2011) hypercorrection studies establish that high-confidence errors are corrected more dramatically than low-confidence errors in the immediate post-feedback period. But the same research tradition demonstrates that the incorrect memory trace underlying a confident wrong answer is not erased; it is weakened and displaced in the short term and can resurface as a competing response at delayed tests (Butterfield and Metcalfe, 2001). The memory reconsolidation framework (Nader et al., 2000)

formalises the biological mechanism: reactivated memories return to a labile state during which they can be updated or strengthened; inadequate post-reactivation intervention leaves the incorrect trace intact. A student whose early-session error profile is dominated by confabulations is, on this account, accumulating a pool of weakened but intact incorrect traces that will compete with correct responses at the 2-week retention test. The CFI-CF component operationalises the magnitude of this pool.

If CFI-CF predicts retention significantly above and beyond overall accuracy, the Confusion Fingerprint achieves status as a genuine individual-difference construct. Its diagnostic value would be separable from the adaptive platform: any educator with access to qualitative error observation during formative assessment could identify high-CFI-CF students and prioritise explicit misconception correction before high-stakes assessments. The construct would thus be applicable in the vast majority of rural Indian classrooms that lack device access entirely. The moderation analysis (H2b)—testing whether the CFI-CF → retention relationship is attenuated in Group A—will provide mechanistic evidence about whether the contrast-and-correct tip delivery successfully reconsolidates confabulated memory traces.

Metacognitive Calibration (H3)

The Brier score analysis tests a theoretically independent mechanism: whether error-taxonomy-aware feedback improves metacognitive calibration by explicitly addressing the overconfidence implicit in confabulated responses. Black and Wiliam's (1998) formative assessment framework and Hattie and Timperley's (2007) power-of-feedback analysis both converge on the prediction that feedback which identifies the nature and source of an error—not merely its occurrence—produces superior calibration and regulation outcomes. The error-type-specific tips delivered in Group A operationalise this principle at the item level. A significant H3 result would provide evidence that cognitive-mechanism-specific feedback produces calibration benefits in early adolescent learners in a low-resource context, extending Carpenter et al.'s (2019) adult findings across populations and delivery modes.

Limitations

Five limitations require explicit acknowledgement. *Investigator as administrator:* the study investigator developed the Cognivia platform, constructed the question bank, and will administer all sessions—a potential source of experimenter expectancy effects, mitigated by pre-specified rubric-based error classification, blind scoring, and audio-recorded fidelity checks. *Pilot sample size:* n = 30 per group is adequate for effect-size estimation but underpowered for the H2b moderation analysis; no strong confirmatory claims will be made. *Rule-based classification:* error types are assigned at question design time rather than through free-response analysis; idiosyncratic student reasoning not conforming to pre-specified distractor logic will be misclassified. *Mechanism confound:* Group A receives both error-type-specific question sequencing and tip delivery; a four-arm design isolating each component is required for future confirmatory work. *Single domain and site:* NCERT Science in Kanpur Dehat; replication across Mathematics, other states, and other low-resource contexts is required before claims of generalisability.

Future Directions

This study is the first step in a programmatic research agenda. The natural sequel is a pre-specified confirmatory RCT, adequately powered on effect sizes obtained here, with a four-arm design independently manipulating error-type sequencing and error-type tip delivery, a blinded second administrator, and n equal to or greater than 120 per arm. A second priority is the longitudinal stability of the Confusion Fingerprint: a two-year tracking study from Class 6 to Class 8 would determine whether the CFI profile constitutes a stable cognitive style or a context-specific response pattern—a prerequisite for its use as a diagnostic instrument in educational practice. Third, extension of the CFI taxonomy to Mathematics, where error taxonomies have been independently developed (Newman, 1977; Booth et al., 2014), would test domain generality. Finally, a teacher-facing CFI dashboard—aggregating class-level error-type profiles without individual device access—could translate the construct's diagnostic value into a low-technology intervention deployable at scale across the 250 million-learner government school system.

CONCLUSION

Standard adaptive learning systems ask one question of a student's error: did it happen? This study asks a second: what kind of error was it? The Confusion Fingerprint hypothesis—that the cognitive type of a learning error carries predictive and instructionally actionable information beyond the correctness signal alone—has been theoretically motivated for decades across the errorful learning, metacognitive calibration, and conceptual change literatures, but has never been subjected to a prospective empirical test in a naturalistic classroom setting. This study provides that test.

It does so in a population—rural, Hindi-medium, government-school learners in India—whose near-total absence from the adaptive learning evidence base represents both a scientific gap and an equity concern. The approximately 250 million children enrolled in India's government primary and secondary schools face the highest-stakes educational assessments of their lives with little access to individualised instruction and with adaptive technology that has been designed, trained, and validated almost entirely in settings that bear little resemblance to theirs. This study generates the first pre-registered estimates of adaptive learning efficacy in this context, contributing to the field's external validity regardless of the magnitude or direction of the findings.

The Confusion Fingerprint Study is, at its core, a case for a richer theory of instructional error. Adaptive systems that attend to what kind of wrong a student's error represents—not merely that it occurred—have the theoretical potential to produce durable learning benefits that correctness-binary systems cannot. This paper is the evidence base that will determine, empirically, whether that potential is real.

OPEN SCIENCE STATEMENT

All data will be fully anonymised and deposited as open data at OSF.io upon completion of data collection. Deposited materials will include: the complete 240-item NCERT Science question bank with all error-type labels and the inter-rater reliability dataset; the error classification rubric; the Cognivia source code (Python/Flask; MIT licence); the JASP analysis file with all pre-registered models; and the researcher's structured field journal (redacted of identifying information). The OSF preregistration timestamps all hypotheses, the analysis plan, feasibility outcomes, exclusion criteria, and stopping rules prior to any data collection, consistent with the registered reports model (Chambers, 2013; Nosek et al., 2018).

ACKNOWLEDGEMENTS

The author thanks the principals, teachers, students, and families of the participating schools in Kanpur Dehat district for their cooperation and support. Appreciation is also extended to colleagues, educators, and mentors whose discussions and feedback helped inform the broader development of the ideas presented in this work.

This study received no external funding. The Cognivia platform, study protocol, and all study materials were developed independently by the author, who is solely responsible for the conception, design, methodology, analysis plan, and preparation of this manuscript.

DECLARATION OF INTERESTS

The author is the developer of the Cognivia platform used as the study instrument. This dual role is acknowledged as a potential source of expectancy bias and is addressed through the design features detailed in the Method section.

AUTHOR CONTRIBUTIONS

M.M.: Conceptualisation, Methodology, Software (Cognivia), Formal Analysis Plan, Investigation (data collection), Data Curation, Writing—Original Draft, Writing—Review and Editing, Project Administration.

FUNDING

No external funding was received for the development of the Cognivia platform, the design of this protocol, or the preparation of this manuscript.

HOW TO CITE THIS PREPRINT

Maurya, M. (2026). The Confusion Fingerprint Index: Cognitive error taxonomy as an adaptive learning signal in learners. EdArXiv. https://doi.org/10.35542/osf.io/cr53e_v1. Platform and study materials: cognivia.vercel.app. ORCID: 0009-0005-3554-693X.

REFERENCES

1. Auer P, Cesa-Bianchi N, and Fischer P. Finite-time analysis of the multiarmed bandit problem. *Mach Learn* 47: 235–256, 2002.
2. Bjork RA. Memory and metamemory considerations in the training of human beings. In: *Metacognition: Knowing About Knowing*, edited by Metcalfe J and Shimamura A. Cambridge, MA: MIT Press, 1994, p. 185–205.
3. Bjork RA, Dunlosky J, and Kornell N. Self-regulated learning: Beliefs, techniques, and illusions. *Annu Rev Psychol* 64: 417–444, 2013.
4. Black P and Wiliam D. Assessment and classroom learning. *Assess Educ* 5: 7–74, 1998.
5. Booth JL, Barbieri C, Eyer F, and Paré-Blagoev EJ. Persistent and pernicious errors in algebraic problem solving. *J Problem Solving* 7: 3, 2014.
6. Braun V and Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 3: 77–101, 2006.
7. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78: 1–3, 1950.
8. Butterfield B and Metcalfe J. Errors committed with high confidence are hypercorrected. *J Exp Psychol Learn Mem Cogn* 27: 1491–1494, 2001.
9. Carpenter J, Sherman MT, Kievit RA, Seth AK, Lau H, and Fleming SM. Domain-general enhancements of metacognitive ability through adaptive training. *J Exp Psychol Gen* 148: 51–64, 2019.
10. Chambers CD. Registered reports: A new publishing initiative at Cortex. *Cortex* 49: 609–610, 2013.
11. Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleza-Jeric K, Hrobjartsson A, Mann H, Dickersin K, Berlin JA, Dore CJ, Parulekar WR, Summerskill WSM, Groves T, Schulz KF, Sox HC, Rockhold FW, Rennie D, and Moher D. SPIRIT 2013 Statement: Defining standard protocol items for clinical trials. *Ann Intern Med* 158: 200–207, 2013.
12. Chi MTH. Conceptual change within and across ontological categories: Examples from learning and discovery in science. In: *Cognitive Models of Science*, edited by Giere RN. Minneapolis, MN: University of Minnesota Press, 1992, p. 129–186.
13. Chi MTH and Wylie R. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educ Psychol* 49: 219–243, 2014.
14. Choi Y, Lee Y, Cho J, Baek J, Kim B, Cha Y, Shin D, Bae C, and Heo J. Towards an appropriate query, key, and value computation for knowledge tracing (SAINT). In: *Proc 10th Int Conf Learn Anal Knowl* (LAK 2020). New York: ACM, 2020, p. 325–334.
15. Corbett AT and Anderson JR. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model User-Adapt Interact* 4: 253–278, 1995.
16. Dunlosky J and Rawson KA. Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learn Instr* 22: 271–280, 2012.
17. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, and Bond CM. Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLoS One* 11: e0150205, 2016.
18. Fixsen DL, Naoom SF, Blase KA, Friedman RM, and Wallace F. *Implementation Research: A Synthesis of the Literature.* Tampa, FL: University of South Florida, NIRN, 2005.
19. Flavell JH. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *Am Psychol* 34: 906–911, 1979.
20. Ghosh A, Heffernan N, and Lan AS. Context-aware attentive knowledge tracing. In: *Proc 26th ACM SIGKDD Int Conf Knowl Discov Data Min* (KDD 2020). New York: ACM, 2020, p. 2330– 2339.
21. Gligorea I, Cioca M, Oancea R, Gorski AT, Gorski H, and Tudorache P. Adaptive learning using artificial intelligence in e-learning: A literature review. *Educ Sci* 13: 1216, 2023.
22. Hacker DJ, Dunlosky J, and Graesser AC (Eds). *Metacognition in Educational Theory and Practice.* Mahwah, NJ: Erlbaum, 2008.
23. Hattie J and Timperley H. The power of feedback. *Rev Educ Res* 77: 81–112, 2007.
24. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, Altman DG, Barbour V, Macdonald H, Johnston M, Lamb SE, Dixon-Woods M, McCulloch P, Wyatt JC, Chan AW, and Michie S. Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 348: g1687, 2014.
25. Kluger AN and DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull* 119: 254–284, 1996.
26. Kornell N and Bjork RA. Learning concepts and categories: Is spacing the 'enemy of induction'? *Psychol Sci* 19: 585–592, 2008.
27. Kornell N, Hays MJ, and Bjork RA. Unsuccessful retrieval attempts enhance subsequent learning. *J Exp Psychol Learn Mem Cogn* 35: 989–998, 2009.

28. Lakens D, Scheel AM, and Isager PM. Equivalence testing for psychological research: A tutorial. *Adv Methods Pract Psychol Sci* 1: 259–269, 2018.

29. Liu Z, Liu Q, Chen J, Lv Q, Zhao H, and Chen E. pyKT: A Python library to benchmark deep learning based knowledge tracing models. In: *Advances in Neural Information Processing Systems* 35, 2022, p. 25280– 25293.

30. Metcalfe J. Learning from errors. *Annu Rev Psychol* 68: 465–489, 2017.

31. Metcalfe J and Finn B. People's hypercorrection of high-confidence errors: Did they know it all along? *J Exp Psychol Learn Mem Cogn* 37: 437–448, 2011.

32. Nader K, Schafe GE, and LeDoux JE. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* 406: 722–726, 2000.

33. Newman MA. An analysis of sixth-grade pupils' errors on written mathematical tasks. *Vic Inst Educ Res Bull* 39: 31–43, 1977.

34. Nosek BA, Ebersole CR, DeHaven AC, and Mellor DT. The preregistration revolution. *Proc Natl Acad Sci USA* 115: 2600–2606, 2018.

35. Pandey S and Karypis G. A self-attentive model for knowledge tracing. In: *Proc 12th Int Conf Educ Data Min* (EDM 2019). International Educational Data Mining Society, 2019.

36. Pavlik PI, Cen H, and Koedinger KR. Performance factors analysis—a new alternative to knowledge tracing. In: *Proc 14th Int Conf Artif Intell Educ* (AIED 2009). Berlin: IOS Press, 2009, p. 531–538.

37. Piech C, Spencer J, Huang J, Ganguli S, Sahami M, Guibas L, and Koller D. Deep knowledge tracing. *Adv Neural Inf Process Syst* 28: 505–513, 2015.

38. Potts R and Shanks DR. The benefit of generating errors during learning. *J Exp Psychol Gen* 143: 644–667, 2014.

39. Roediger HL and Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol Sci* 17: 249–255, 2006.

40. Schulz KF, Altman DG, and Moher D, for the CONSORT Group. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ* 340: c332, 2010.

41. Skinner BF. Teaching machines. *Science* 128: 969–977, 1958.

42. Sweller J. Cognitive load during problem solving: Effects on learning. *Cogn Sci* 12: 257–285, 1988.

43. Treagust DF. Development and use of diagnostic tests to evaluate students' misconceptions in science. *Int J Sci Educ* 10: 159–169, 1988.

44. VanLehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ Psychol* 46: 197–221, 2011.

45. Veenman MVJ, Van Hout-Wolters BHAM, and Afflerbach P. Metacognition and learning: Conceptual and methodological considerations. *Metacogn Learn* 1: 3–14, 2006.

46. Verma AP. EDGE: A theoretical framework for misconception-aware adaptive learning. *arXiv preprint* arXiv:2508.07224, 2025.

47. Vosniadou S. Capturing and modeling the process of conceptual change. *Learn Instr* 4: 45–69, 1994.

48. Wang X, Huang R, Sommer M, Pei B, Shidfar P, Rehman MS, Ritzhaupt AD, and Martin F. The efficacy of artificial intelligence-enabled adaptive learning systems from 2010 to 2022 on learner outcomes: A meta-analysis. *J Educ Comput Res* 62: [in press], 2024.

49. Wozniak P. *Optimization of Learning.* Master's thesis, University of Technology, Poznan, 1990.

50. Zhang J, Shi X, King I, and Yeung D-Y. Dynamic key-value memory networks for knowledge tracing. In: *Proc 26th Int Conf World Wide Web* (WWW 2017). Geneva: IW3C2, 2017, p. 765–774.