

Making Deliberation Visible

A Three-Site Feasibility and Instrument-Validation Study of a Constrained Deliberative System in Secondary Education

PRE-REGISTRATION DRAFT V1.0

OPEN FOR METHODOLOGICAL COMMENT

Making Deliberation Visible: A Three-Site Feasibility and Instrument-Validation Study

Pre-registration

Administrative Information

Title:

Making Deliberation Visible: A Three-Site Feasibility and Instrument-Validation Study of a Constrained Deliberative System in Secondary Education

Principal Investigator:

Kennedy Ifeh

CrossBorderCart Inc. (Cuestora) Vaughan, Ontario, Canada

Academic Collaborators: None confirmed at the time of publication

Registration Status: Draft protocol, not yet registered

Planned Start of Data Collection: October 2026 (provisional)

Planned Measurement Period: Six weeks (provisional)

Funding and Conflict of Interest:

This study is self-funded and receives no external research funding. The principal investigator is the founder of the company that develops the software under study. There is a conflict of interest based on the ownership and research of the software under review. The pre-registration, specification of the study's measures and analyses, and transparent reporting of deviations prior to the research are part of a deliberate effort to constrain the discretion of the researcher.

Pre-registration and Immutability Clause:

This document, therefore, is the prospective study specification and will be fixed at the point of registration, before data collection begins. The hypotheses, outcome measures, decision thresholds,

and confirmatory analysis methods specified here will not be retrospectively altered in response to observed study data.

If circumstances require a departure from the registered protocol, the original specification will remain unchanged. The departure will instead be documented explicitly in the study report, together with the reason for the deviation, when it occurred, and its likely implications.

Any analyses introduced after registration that were not specified in this document will be identified as exploratory or post hoc and will not be presented as pre-registered confirmatory analyses.

Draft Status: This document is published as Pre-registration Draft v1.0 and is circulated for methodological comment prior to registration. Items marked "Open methodological item" are known gaps in the specification. They will be resolved and fixed before registration and before any study data are collected or inspected.

1. Background and Rationale

Most educational assessments evaluate what a learner produces, whether it's an essay, a written response, a solved problem, or a completed work. These artifacts are used as evidence of what the learner knows or understands. Generative AI complicates this model because AI can now contribute to the final work without showing the underlying deliberation that was done by the learner.

There is also a growing concern about the effects of heavy reliance on AI during intellectual activities. Kosmyna et al. (2025), for example, reported differences in neural connectivity, recall, and reported ownership of writing across different essay-writing conditions involving LLM assistance. These findings reinforce concerns about cognitive offloading, although, they should be interpreted within the limits of the study's sample, design, task conditions, and with the recognition that further research is needed.

To highlight the enormity of the problem, this study uses the term **cognitive annexation** to enunciate the process whereby AI begins to perform parts of the thinking process that would otherwise have been carried out by the learner. In this context, the term is used as a theoretical idea.

The study examines Cuestora, a software environment designed to protect human agency, preserve learner authorship, and make the process of deliberation more visible. Learners record units of their own reasoning as "ThoughtBlocks." They can revise these blocks, challenge their reasoning through an AI interlocutor that is constrained from supplying answers, conclusions, or rewritten learner text, and produce a lasting record of the process called a "ThinkPrint."

Whether this type of environment can support deliberative practice among secondary learners, and whether that practice can be observed reliably through the records it produces, has not yet been tested. This study examines those questions across three different learner populations.

The study draws partly on trace-based research in self-regulated learning (e.g., Winne; Winne & Perry; Winne & Hadwin). This research uses records of learner activity over time to study learning processes rather than relying only on what learners later report about what they did.

The present study takes a more limited approach to these records. It does not treat a behavioural trace as direct evidence of what happened inside a learner's mind. Instead, it examines observable changes in learner-authored work, what the learner first committed, what was later changed, what challenge came before the change, and how these events are connected in the record.

The study therefore describes what can be observed in the deliberative record while avoiding claims about internal cognitive changes that cannot be directly measured.

1.1 Aim of the Study

The aim of this study is to determine whether observable traces of deliberative activity, including changes in learner-authored work, can be reliably identified and described during six weeks of Cuestora use. It will also examine whether selected indicators of this activity change within individual learners over the six-week period.

The study does not claim to measure understanding, mastery, intelligence, comprehension, or any other internal cognitive state. Where an observable change may have a possible cognitive meaning, the study will clearly distinguish between **what was observed** and **what that observation might mean**.

2. Theoretical Framework

2.1 Constrained Deliberative System (CDS)

A **Constrained Deliberative System (CDS)** is an AI-mediated environment wherein the role of the AI is deliberately constrained to preserve the human thinking process. Unlike conventional AI systems designed mainly to generate answers or complete tasks, a CDS places rules around what the AI can do, when it can intervene, and how its contribution is recorded.

In Cuestora, these rules include requiring the learner to commit their own work before certain AI interventions, preserving learner-authored text, showing where contributions come from, separating the learner's committed work from system-generated interpretations, and preventing AI-generated text from being silently substituted for the learner's own deliberative work.

2.2 The Three-Layer Architecture

The CDS is organised around three layers: the **Deliberative Record**, **Constitutional Governance**, and **Deliberative Inference**.

Layer 1: Deliberative Record

This is the observable record of what happened during the learner's work. It includes committed ThoughtBlocks, revisions, timestamps, interactions, provenance, and other recorded events.

Layer 2: Constitutional Governance

This contains the rules that govern what the system is allowed to do. These rules are intended to maintain a clear separation between the learner's own work and AI intervention.

Layer 3: Deliberative Inference

This contains interpretations or classifications made from the deliberative record. These interpretations must remain separate from the underlying observations. They must not be presented as direct measurements of what a learner thinks, understands, or experiences internally.

2.3 Boundaries of What the Study Can Establish

The study is deliberately limited in what it seeks to establish.

2.3.1 This study can establish:

- Whether the intended deliberative activities occur during real use across the study populations (**feasibility**).
- Whether the system's automated observations correspond with what independent human reviewers identify in the same records (**instrument validity**).
- Whether selected indicators of deliberative activity change within individual learners over the six-week study period (**signal detection**).
- Where learners experience friction, stop using parts of the system, or encounter other difficulties (**design evidence**).

2.3.2 This study cannot establish:

- That Cuestora causes an improvement in reasoning. The study has no control group, randomisation, or comparison condition.
- That any changes observed in these populations would also occur in other populations.
- That changes in learner-authored artifacts represent changes in the learner's underlying cognitive ability or capacity.

This is a **feasibility and instrument-validation study with an exploratory signal-detection component**. It is not an efficacy trial, and its findings will not be presented as evidence that Cuestora improves reasoning or learning.

3. Research Questions and Hypotheses

3.1 Research Questions

Primary Research Question:

What observable patterns of deliberative activity emerge during six weeks of repeated Cuestora use among secondary-age learners, and how do these patterns change for individual learners over that period?

Secondary Research Question:

Do independent human reviewers identify the same kinds of changes in learner-authored work as the system?

3.2 Confirmatory Hypotheses

3.2.1 H1 — Feasibility

At least 50% of onboarded participants will commit 10 or more ThoughtBlocks during the six-week measurement period.

Ten committed ThoughtBlocks is set in advance as the minimum level of activity needed to show repeated use of the core capture–commit–reflection cycle. The 50% threshold is a feasibility criterion. It is not a measure of efficacy.

3.2.2 H2 — Revision Engagement

Within each participant, the proportion of committed ThoughtBlocks that are later revised will be higher in Weeks 5–6 than in Week 1.

A revision will count if it occurs within seven days of the ThoughtBlock's initial commitment. ThoughtBlocks committed during the six-week measurement period may continue to be observed during the seven-day follow-up period described in Section 10.1.

3.2.3 H3 — Revision Transformation

Among revisions that occur within the defined observation period, the proportion classified as **non-cosmetic transformations** will be higher in Weeks 5–6 than in Week 1.

This hypothesis distinguishes changes to a learner-authored artifact from changes that affect only its surface form. Cosmetic and non-cosmetic transformations will be defined in advance in the Transformation Codebook. These classifications describe changes in the learner's work. They will not be treated as evidence of changes in the learner's underlying cognitive state.

3.2.4 H4 — Tension Engagement

Within each participant, the proportion of flagged tensions that are followed within seven days by a revision to an involved ThoughtBlock will be higher in Weeks 5–6 than in Week 1.

A **flagged tension** and a **involved ThoughtBlock** will be defined in advance in the study measures. The analysis concerns the sequence of observable events. A revision that follows a flagged tension will not, on that basis alone, be described as having been caused by the tension or as occurring in response to it.

3.2.5 H5 — Classification Reliability and Agreement

H5a — Human Coding Reliability

Independent human reviewers, blinded to the system's classifications, will demonstrate above-chance agreement when classifying sampled before-and-after ThoughtBlock revision pairs using Transformation Codebook v0.1.

H5b — System–Human Classification Agreement

The system's transformation classifications will demonstrate above-chance agreement with the predefined human reference classification for the same sampled before-and-after revision pairs.

The statistical analysis in Section 6 will specify the agreement measures and thresholds, how the human reference classification will be constructed, how revision pairs will be sampled, and how disagreements and **Unknown** classifications will be handled.

4. Study Design and Participants

This is a prospective, multi-site, single-arm feasibility and instrument-validation study. Participants will use Cuestora over a six-week measurement period, allowing their activity to be observed over time.

There will be no randomisation, control group, or comparison condition. Participants and the investigator will know that Cuestora is being used.

The study is designed to examine whether Cuestora can be used in real educational settings, whether observable patterns of deliberative activity change over time, and whether selected system classifications can be reproduced by independent human reviewers.

Because there is no control group, changes observed during the study cannot be said to have been caused by Cuestora.

4.1 Study Sites

The study will take place across three different educational settings.

Site	Population	Setting
A	Secondary-age learners recruited through tutoring and community-based educational networks	Remote / Community
B	Students from a publicly funded secondary school serving a newcomer and lower-income community in the Greater Toronto Area	Classroom
C	Students from an independent secondary school in the Greater Toronto Area	Classroom

The three sites are included to examine whether the study can be carried out and whether observable deliberative activity can be captured across different learner populations and educational settings.

The sites are not intended to represent the wider population of secondary school students, and the study is not designed to determine whether one site or population performs better than another.

Language of instruction, primary language, and use of dictation will be recorded where relevant. These factors may affect the form of learner-authored artifacts or their revisions and will therefore be considered when interpreting the study findings.

Participation by each site depends on institutional agreement before data collection begins. If a planned site is not secured or later withdraws, the study will continue with the remaining sites. The change and its possible effect on the study will be reported.

4.2 Blinding for Instrument Validation

For H5, independent human reviewers will examine sampled before-and-after ThoughtBlock revision pairs.

The reviewers will not be shown:

- the classification produced by Cuestora;
- the participant's study site; or
- whether the revision occurred during the early or late part of the six-week period.

Reviewers will classify the changes using the preregistered Transformation Codebook v0.1.

This blinding applies only to the instrument-validation component of the study. Participants and the investigator will not be blinded to the use of Cuestora.

4.3 Instrument Freeze

A calibration period will take place before the six-week measurement period begins. During this period, prompts, classifiers, models, and relevant parts of the interface may be adjusted.

Once the measurement period begins, the following components will be frozen:

- the reasoning-move classification prompt and model;
- the Reflection Panel generation logic;
- the ThinkPrint generation logic;
- the ThoughtBlock commitment and revision data model; and
- Transformation Codebook v0.1 (Appendix A).

These components will not be changed during the measurement period unless a change is necessary for participant safety, privacy, security, data integrity, or critical system functionality.

If such a change is necessary, it will be dated, versioned, and documented together with the reason for the change. If the change could affect a study measure, observations produced before and after the change will remain distinguishable.

Bug fixes that do not change the definition or generation of a study measure may be made during the measurement period. These changes will be recorded in the study change log.

Relevant model, prompt, classifier, codebook, and system-version information will be retained where applicable so that automated classifications and derived observations can be traced to the version of the system that produced them.

5. Measures and Observable Indicators

Most measures in this study come from records produced through ordinary use of Cuestora. Participants will not be asked to perform additional measurement tasks beyond the participant and educator-reported measures described in Section 5.5.

5.1 Levels of Measurement

The study distinguishes between three types of measures according to what can reasonably be claimed from them.

Level 1 — Observable Events

These are events that can be verified directly from the system record without interpreting their meaning. Examples include a ThoughtBlock being committed, a revision being created, a Tension Report being generated, or a session taking place.

Level 2 — Observable Transformations

These are changes that can be verified by comparing two states of an artifact or two recorded events. Examples include text being added or removed, a citation appearing, one block becoming two, or the amount of textual change between two versions.

Level 3 — Interpretive Classifications

These require judgement about the meaning of an artifact or change. Examples include identifying that an assumption was made explicit, a claim was qualified, or a counterargument was introduced.

Level 3 classifications will remain clearly separated from the underlying observable record. They will not be presented as direct measurements of a learner's internal cognitive state.

H1, H2, and H4 rely primarily on Level 1 and Level 2 measures. H3 is different because it depends on Level 3 classification of ThoughtBlock transformations. H3 will therefore be interpreted together with the human-coding reliability and system–human agreement results specified in H5.

5.2 Primary Measures

M1 — Committed ThoughtBlocks (Level 1)

Number of ThoughtBlocks with committed status for each participant during each week of the study.

M2 — Seven-Day Revision Rate (Level 1)

The proportion of eligible committed ThoughtBlocks that receive at least one revision within seven days of their initial commitment.

ThoughtBlocks committed during the six-week measurement period are eligible for this measure. Events originating in Week 6 may continue to be observed during the seven-day follow-up period described in Section 10.1.

No new ThoughtBlocks originating during the follow-up period will enter this measure.

M3 — Revision Magnitude (Level 2)

The amount of textual change between an original ThoughtBlock and its revised version will be measured using normalized token-level edit distance.

Revision magnitude will be reported as a continuous measure.

For an additional descriptive analysis, revisions with a normalized edit distance greater than 0.20 will be identified as **higher-magnitude textual revisions**.

The 0.20 threshold is a provisional boundary chosen before inspection of study outcomes. It is not treated as a validated cutoff for meaningful or substantive revision. The continuous edit-distance measure remains the primary description of revision magnitude.

Revision magnitude describes only how much the text changed. It does not indicate whether the revision was important, improved the work, or reflected better reasoning.

The meaning of a revision will be examined separately using Transformation Codebook v0.1 and the human-review procedure.

M4 — Post-Tension Revision Rate (Level 1/2)

The proportion of flagged tensions that are followed within seven days by a revision to at least one ThoughtBlock involved in the flagged tension.

The generation of a Tension Report, the timestamp of the flag, the ThoughtBlocks named in the flag, and any later revision are observable records.

Whether the flagged relationship represents a meaningful tension is a system-generated interpretation.

This measure records the sequence of observable events. It does not establish that the flagged tension caused the later revision.

Events originating in Week 6 may continue to be observed during the seven-day follow-up period described in Section 10.1.

M5 — Reasoning-Move Variety (Level 3, Exploratory)

The study will describe the range and distribution of reasoning-move labels assigned to a participant's ThoughtBlocks during each week.

Because these labels are system-generated interpretations, this measure will be treated as exploratory.

Reasoning-move variety will not be interpreted as a direct measure of reasoning ability, quality, or cognitive development.

5.3 Adoption and Feasibility Measures

The following observable measures will also be used to describe adoption and continued use of Cuestora:

- onboarding completion rate;
- sessions per participant per week;
- active days per week;
- proportion of participants still active in Week 6;
- time from account creation to first committed ThoughtBlock;
- number of Tension Reports generated;

- number of ThinkPrints generated; and
- number of ThinkPrints shared with a tutor or teacher.

These measures describe use of the system and will not be interpreted as evidence of learning or improved reasoning.

5.4 Patterns That May Suggest Performative Use

An increase in recorded activity does not necessarily mean that the nature of the learner's work has changed.

For this reason, the study will look for patterns that could provide alternative explanations for increases in the measures used in H2–H4.

These include:

- revision frequency increasing while median revision magnitude remains stable or decreases;
- flagged tensions increasing while the proportion followed by revision within seven days decreases; and
- ThoughtBlock creation becoming increasingly concentrated in the hour before a ThinkPrint is generated or shared.

These patterns will not be treated as proof that a participant is deliberately trying to manipulate the system.

If observed, they will be reported as possible alternative explanations for changes in the study measures and considered alongside other available evidence.

No participant will be labelled as "gaming" based on these records alone.

5.5 Participant- and Educator-Reported Measures

5.5.1 Participant Confidence

Participants may optionally record how confident they feel about a topic using one of three options:

- Confident
- Uncertain
- Needs work

These ratings belong to the participant and will not be changed by the system. They may be examined alongside observable measures but will not be used to validate those measures.

5.5.2 Educator Observations

Participating educators will provide brief weekly structured observations about what they noticed in learners' reasoning or engagement.

These observations will be collected independently of the system's classifications and will be used as contextual evidence rather than confirmatory evidence for H1–H5.

5.5.3 Exit Interviews

Semi-structured exit interviews will be conducted with a purposive subsample of participating learners and with participating educators.

Interview findings will be used to understand participant experience, implementation issues, and possible explanations for patterns observed in the study.

They will not be used as confirmatory evidence for H1–H5.

6. Analysis Plan

6.1 Unit of Analysis

The main longitudinal analyses will examine change within individual participants.

Each eligible participant's activity in Weeks 5–6 will be compared with that same participant's activity in Week 1. This allows each participant to serve as their own baseline and reduces the influence of stable differences between participants.

Results will also be summarized across participants and by study site.

Within-person change does not establish that Cuestora caused the change. Other factors, including continued exposure, maturation, differences in participation over time, and participant dropout, may contribute to any patterns observed.

The study will therefore describe these findings as **changes observed during Cuestora use**, not changes caused by Cuestora.

6.2 Primary Analysis

For H2, H3, and H4, a baseline value and a late-period value will be calculated for each eligible participant.

Baseline: eligible observations originating in Week 1.

Late period: eligible observations originating in Weeks 5 and 6, pooled to produce one late-period value.

For each participant:

within-person change = late-period value – baseline value

The median within-person change will then be reported across eligible participants, together with a bootstrap 95% confidence interval based on 10,000 resamples.

Results will be reported for the pooled sample and separately by study site where sufficient observations are available.

The primary analysis focuses on estimating the size and direction of observed change rather than testing statistical significance. P-values will not be used to determine whether H2–H4 are supported.

Confidence intervals will be reported in full, including when they include zero.

6.2.1 H1 - Feasibility

The proportion of onboarded participants who reach at least 10 committed ThoughtBlocks during the six-week measurement period will be calculated.

A binomial 95% confidence interval will be reported for the pooled sample and separately by site.

6.2.2 H2 - Revision Engagement

For each eligible participant, the seven-day revision rate for Week 1 will be compared with the seven-day revision rate for Weeks 5–6.

A revision will count if it occurs within seven days of the ThoughtBlock's initial commitment. ThoughtBlocks committed during the six-week measurement period may continue to be observed during the seven-day follow-up period described in Section 10.1.

6.2.3 H3 - Revision Transformation

For each eligible participant, the proportion of revisions classified as non-cosmetic transformations in Week 1 will be compared with the corresponding proportion in Weeks 5–6.

A participant must have at least one eligible revision in both periods for this comparison to be made.

Transformation classifications will be made according to Transformation Codebook v0.1.

6.2.4 H4 - Tension Engagement

For each eligible participant, the proportion of flagged tensions followed within seven days by a revision to an involved ThoughtBlock will be calculated for Week 1 and Weeks 5–6.

A participant must have at least one eligible flagged tension in both periods for this comparison to be made.

The analysis concerns the observed sequence of events and does not establish that a flagged tension caused a later revision.

6.2.5 H5 - Classification Reliability and Agreement

A sample of 100 eligible before-and-after ThoughtBlock revision pairs will be selected for human review.

The sample will be drawn across the three study sites using a sampling procedure fixed before the classifications are examined. If fewer than 100 eligible revision pairs exist across the study, all eligible pairs will be included.

Independent human reviewers, blinded as described in Section 4.2, will classify each pair using Transformation Codebook v0.1.

Human reviewer agreement will be reported using:

- percentage agreement; and
- Krippendorff's alpha for nominal classifications.

After independent coding is complete, disagreements will be resolved using the adjudication procedure specified in the codebook. This will produce a human reference classification for each reviewed revision pair.

The system's classification will then be compared with the human reference classification using percentage agreement and the preregistered chance-corrected agreement measure.

Open methodological item: The chance-corrected statistic and decision threshold for system–human agreement have not yet been fixed. Both will be specified before preregistration and before any study data are collected or inspected.

The treatment of **Unknown** classifications will be specified in Transformation Codebook v0.1 before the study begins.

6.3 Exploratory Analysis

Additional analyses may be used to explore patterns that are not part of the confirmatory hypotheses.

These may include mixed-effects models examining change over time, with participant-level random effects and study site included where appropriate.

All such analyses will be clearly labelled as exploratory.

Patterns that were not predicted in this preregistration may be reported as possible findings or hypotheses for future research, but they will not be presented as preregistered confirmatory findings.

6.4 Eligibility, Exclusion, and Missing Data

Eligibility will be determined separately for each hypothesis because the measures require different observable events.

Participants who do not qualify for H2, H3, or H4 will still remain part of the feasibility and adoption analyses where applicable.

H2 eligibility: at least one eligible committed ThoughtBlock in both Week 1 and Weeks 5–6.

H3 eligibility: at least one eligible revision in both Week 1 and Weeks 5–6.

H4 eligibility: at least one eligible flagged tension in both Week 1 and Weeks 5–6.

Accounts created for testing, demonstration, investigator use, or educator use will be excluded from participant analyses.

Participants who withdraw consent will be handled according to the withdrawal procedures described in Section 8.

Participants will not be excluded simply because their level of activity is low. Low engagement is itself an important feasibility outcome, and excluding low-activity participants could make engagement appear higher than it actually was.

No missing outcome data will be imputed.

Where absence of activity has a meaningful value of zero, it will be recorded as zero. Where a measure cannot be calculated because the required event did not occur, it will be recorded as **not estimable** rather than assigned a zero.

For example, if a participant makes no revision, revision magnitude cannot be calculated and will therefore be recorded as not estimable.

For every longitudinal analysis, the study will report how many participants were enrolled, how many remained active, how many were eligible for that particular analysis, and how many did not contribute to it.

H2–H4 therefore describe within-person change among participants for whom the relevant activity can be observed in both the baseline and late periods. They do not estimate change for every participant who was initially onboarded.

6.5 Stopping Rules

Data collection will end at the completion of the preregistered six-week measurement period regardless of the results observed.

No interim analysis of H1–H5 will be used to decide whether data collection should be extended or stopped early.

Participant safety, system security, privacy, and study operations may be monitored throughout the study. This monitoring will not be used to evaluate the study hypotheses before the measurement period is complete.

7. Kill Conditions

The following conditions are defined before data collection begins. They will be reported publicly whether or not they are met.

These conditions are intended to make clear in advance what results would count against the study's current assumptions, measures, or system design.

7.1 KC1 - Longitudinal Signal

If the median within-person change is at or below zero for **H2 (revision engagement)**, **H3 (non-cosmetic transformation rate)**, and **H4 (post-tension revision rate)**, the study will be interpreted as showing no positive longitudinal signal on the preregistered indicators of deliberative activity over the six-week period.

This result will count against the current expectation that these observable patterns increase during repeated Cuestora use.

It will not be interpreted as evidence that learners' reasoning declined. It will mean only that the preregistered indicators did not show the expected positive change.

The relevant measures and product mechanisms will be reconsidered before stronger claims are pursued.

7.2 KC2 - Adoption

If fewer than 50% of onboarded participants reach 10 committed ThoughtBlocks during the six-week measurement period, the preregistered feasibility criterion will not have been met.

This conclusion will stand regardless of the results of the longitudinal analyses among participants who did use the system regularly.

7.3 KC3a - Codebook Reliability

If agreement between blinded independent human reviewers produces a Krippendorff's alpha below 0.60, Transformation Codebook v0.1 will be considered insufficiently reliable for research use in its current form.

In that case, analyses that depend on the codebook's transformation classifications, including H3, will not be treated as supported confirmatory findings.

The taxonomy will need to be revised and tested again before those classifications are used as validated research measures.

7.4 KC3b - System Classification Validity. If human coding reaches the prespecified reliability threshold but system–human agreement does not meet the decision threshold that will be fixed before preregistration, automated transformation classifications will not be treated as validated research measures.

Open methodological item: The system–human agreement statistic and decision threshold remain to be specified before preregistration.

8. Ethics, Consent, and Data Handling

8.1 Participant Protection

This study involves secondary-age learners and will follow established standards for research involving human participants.

The main protections include:

- parental or guardian consent for participants under 18;
- learner assent where required;
- voluntary participation;
- the right to withdraw;
- protection of privacy and confidentiality;
- responsible handling of participant data; and
- clear disclosure of the investigator's conflict of interest.

Participation in the study, refusal to participate, or withdrawal from the study will not affect a learner's grades, academic standing, tutoring, or access to normal educational activities.

Study data will be used only for the purposes described in the consent materials.

8.2 Parental Consent and Learner Assent

Most participants are expected to be under 18.

Written consent from a parent or guardian and assent from the learner will be obtained before a minor begins study participation or contributes data to the research dataset.

The consent materials will explain:

- what information will be collected;
- what information will not be collected;
- how the information will be used;
- where and how the information will be stored;
- which third-party services process information;

- that participation is voluntary;
- that participants may withdraw without penalty or effect on their schooling or tutoring; and
- that anonymised findings are intended for public reporting or publication.

Required institutional approvals will be obtained separately for each participating site.

No research data collection will begin at a site until the approvals required by that site are in place.

8.3 Data Collected

Depending on how a participant uses Cuestora, the study may collect:

- learner-authored ThoughtBlocks;
- ThoughtBlock revision history;
- timestamps;
- system-generated reasoning-move labels;
- source materials uploaded by the learner;
- ThinkSpace conversation content;
- ThinkPrint records; and
- product-use events such as event type, count, duration, and area of the product used.

Only information needed for the study and normal operation of Cuestora will be used.

8.4 Data Not Collected

The study will not collect:

- biometric data;
- keystroke-level activity;
- screen recordings;
- camera recordings;
- precise location data; or
- activity from applications or websites outside Cuestora.

Learner-authored text will not be included in product analytics events.

8.5 Data Storage and Third-Party Processing

Cuestora uses third-party services to operate the platform.

Component	Provider	Purpose
Application database	Neon (PostgreSQL), US East (N. Virginia) – us-east-1	Storage of application and learner records
Application hosting	AWS Amplify, US East (N. Virginia) – us-east-1	Hosting and delivery of the application
Identity	Clerk	Authentication
AI processing	OpenAI API	AI-supported classification and dialogue
Product analytics	PostHog	Product-use events and counts; no learner-authored text

Data sent through the OpenAI API is not used to train OpenAI's models by default. The study will use the API under the data settings applicable to the Cuestora account at the start of data collection.

8.6 Publication, Quotation, and Anonymisation

Individual participants will not be identified in publications or public study reports.

Study sites will normally be described by type rather than by name. A participating institution will be named only with its written permission.

Learner-authored text will not be quoted publicly simply because the learner participated in the study.

A ThoughtBlock or other learner-authored excerpt may be quoted only where specific additional permission has been obtained from the participant and, where required, their parent or guardian. Any excerpt selected for publication will also be reviewed for information that could identify the learner or another person.

Participating sites will be provided with a summary of the study findings.

8.7 Withdrawal

Participants may withdraw from the study at any time without penalty.

After withdrawal, no new information from that participant will be added to the research dataset.

The consent materials will explain what happens to information already collected and the circumstances under which deletion can be requested or completed.

8.8 Conflict of Interest

The investigator designed and owns the software being studied and has a commercial interest in Cuestora. A favourable study result could therefore benefit the investigator and the company.

This conflict of interest will be disclosed to participating institutions, parents, learners, research collaborators, and in any publication arising from the study.

The following measures are intended to reduce the influence of this conflict on the research:

- preregistration of the hypotheses, measures, analysis plan, thresholds, and failure conditions before data collection;
- freezing measurement-related parts of the system during the study;
- blinded independent human coding for the instrument-validation component;
- public reporting of results whether or not the preregistered criteria are met;
- public reporting of deviations from the registered protocol; and
- publication of the analysis code alongside the study findings.

9. Named Threats to Validity

The following limitations are identified before data collection so that they can be considered when the findings are interpreted.

9.1 Effect of Being Observed

Making deliberative activity visible may itself change how learners behave. Learners may create or revise ThoughtBlocks partly because they know these activities are being recorded.

The patterns described in Section 5.4 will be examined as possible signs of activity performed mainly to satisfy the system's visible measures. Educator observations and exit interviews will provide additional information for interpreting these patterns.

These measures cannot fully remove this possibility.

9.2 Language and Input Mode

System classifications may be affected by a learner's language fluency, grammar, transcription quality, or method of entering text.

For example, a learner using dictation may produce text with different characteristics from a learner typing directly. Similarly, learners working in a language other than their primary language may express the same idea differently.

There is therefore a risk that similar changes in learner-authored work could receive different system classifications because of language or input method rather than because the underlying artifacts differ in the way the classifier is intended to detect.

Language and input method will be considered when interpreting classification results.

9.3 Absence of a Trace Is Not Absence of Thought

Cuestora can record only deliberative activity that becomes visible through interaction with the system.

A learner may think, question, reconsider, or reach a conclusion without recording those activities in Cuestora.

Low ThoughtBlock counts or limited revision activity will therefore not be interpreted as evidence that a learner did little thinking.

Instead, low observable activity will be treated as information about how much of the learner's activity the instrument was able to capture.

9.4 No Control Group

The study has no control or comparison group.

Changes observed over six weeks may result from many factors, including continued exposure to the task, maturation, school-calendar effects, increased familiarity with Cuestora, or other experiences outside the study.

The study therefore cannot establish that Cuestora caused any observed change.

9.5 Human Coding and Investigator Involvement

The investigator has a direct interest in the outcome of the study.

For the H5 instrument-validation analysis, the sampled ThoughtBlock revision pairs will therefore be coded by independent reviewers who are blinded as described in Section 4.2.

The investigator will not serve as one of the independent coders used to establish human inter-rater reliability.

Any investigator involvement in resolving coding questions or administering the coding process will be documented.

9.6 Differences Between Study Sites

The three study sites represent different educational settings and learner populations.

The expected number of participants at each site is too small to support strong conclusions about differences between the populations.

Site-level results will therefore be descriptive. The study will not claim that observed differences between sites represent wider differences between the populations from which those participants were recruited.

9.7 Attrition and Survivorship

H2–H4 require the relevant activity to be observable during both the baseline and late study periods.

Participants who stop using the system before Weeks 5–6 may therefore be unable to contribute to some longitudinal analyses.

This creates a risk that the later-period results disproportionately represent learners who found Cuestora easier, more useful, or more engaging to use.

Retention and eligibility will therefore be reported alongside the longitudinal results.

Findings from H2–H4 will be described as applying to participants who had the required observable activity in both periods, not to every participant who initially joined the study.

9.8 Procedural Familiarity

Learners may become better at using Cuestora simply because they become familiar with its interface and interaction model.

For example, a learner may revise more often in Week 6 because they have learned how and when to use the revision function rather than because their deliberative practice has changed.

The current study cannot separate these explanations.

Increasing familiarity with the system will therefore remain an alternative explanation when interpreting changes observed over the six-week period.

10. Study Timeline

Phase	Planned Period
Calibration	August–September 2026 (provisional)
Consent and onboarding	September 2026 (provisional)
Measurement	Six weeks; planned Fall 2026
Seven-day follow-up	Seven days immediately following Week 6
Continued observation	To end of participating school's term
Analysis	Following completion of measurement and follow-up
Findings/preprint	Following analysis

10.1 Seven-Day Outcome Follow-Up

Some study measures require a seven-day observation period after an event occurs.

For example, the study examines whether a ThoughtBlock is revised within seven days of commitment and whether a flagged tension is followed by revision within seven days.

Events that originate during Week 6 may therefore require observation beyond the end of the six-week measurement period.

For this reason, a seven-day follow-up period will take place immediately after Week 6.

This period does **not** extend the six-week measurement window. Only outcomes linked to qualifying events that originated during the six-week period will be counted.

New ThoughtBlocks, tensions, or other events originating during the follow-up period will not enter the confirmatory H1–H4 analyses.

10.2 Activity After the Confirmatory Study Period

Participants may continue using Cuestora after the seven-day follow-up period until the end of the school term.

Any data examined from this later period will be treated as exploratory or qualitative. It will not be used to test the preregistered hypotheses.

The distinction between the six-week measurement period, seven-day outcome follow-up, and later exploratory observation will be maintained in the study dataset and findings.

Appendix A - Transformation Codebook v0.1

Applied to before-and-after ThoughtBlock revision pairs. The codebook will be frozen before the measurement window begins.

The codes are descriptive, not evaluative. No code indicates that a revision is correct, better, more intelligent, or evidence of improved understanding.

Unit of Classification

The unit of classification is a **before-and-after ThoughtBlock revision pair**:

Before: the committed ThoughtBlock immediately before revision.

After: the resulting revised ThoughtBlock.

Each pair receives one primary transformation code.

Transformation Codes

Code	Transformation	Measurement Level
T01	General → Specific	Level 3
T02	Supporting Evidence Added	Level 3
T03	Example Added	Level 3
T04	Definition or Explanation Added	Level 3
T05	Claim Qualified	Level 3
T06	Claim Split	Level 2
T07	Claims Merged	Level 2
T08	Source or Citation Added	Level 2
T09	Counterargument or Objection Added	Level 3
T10	Assumption Made Explicit	Level 3
T11	Wording Only	Level 3
T12	Claim Modified / Position Changed	Level 3
T13	Claim Withdrawn	Level 3
T99	Unknown / Other	—

Measurement-Level Meaning

Level 2 — Observable Transformation

The change can be identified directly from the artifact structure or visible textual features without needing to judge the meaning of the learner's argument.

Examples include:

- one ThoughtBlock becoming two;
- two ThoughtBlocks becoming one; and
- a citation or source reference appearing.

Level 3 — Interpretive Classification

The classification requires judgement about the meaning of the before-and-after artifacts.

Examples include:

- determining that a claim became more specific;
- identifying that evidence was introduced;
- deciding that a claim was qualified;
- identifying a counterargument;
- determining that an assumption was made explicit; or
- deciding that wording changed without changing the proposition.

Level 3 classifications describe changes in the learner-authored artifact. They are not direct measurements of the learner's internal cognitive state.

Code Definitions

To1 — General → Specific

A broad statement becomes more specific through a narrower subject, population, circumstance, referent, or context.

To2 — Supporting Evidence Added

Material presented as support for an existing claim is introduced.

To3 — Example Added

A concrete example or illustration is added.

To4 — Definition or Explanation Added

A term, idea, or statement is further defined or explained.

T05 — Claim Qualified

The strength or scope of a claim is reduced or limited through language such as "some," "may," "often," or "under certain conditions."

T06 — Claim Split

Content previously contained in one ThoughtBlock becomes two or more ThoughtBlocks.

T07 — Claims Merged

Content previously contained in two or more ThoughtBlocks becomes one ThoughtBlock.

T08 — Source or Citation Added

A citation, source reference, link, or attribution is added.

T09 — Counterargument or Objection Added

An opposing position, limitation, objection, or challenge to an existing claim is introduced.

T10 — Assumption Made Explicit

A condition or assumption that was previously unstated becomes explicit.

T11 — Wording Only

Wording, grammar, spelling, formatting, or sentence structure changes without an identifiable change in the proposition being expressed.

T12 — Claim Modified / Position Changed

The central proposition changes in a way that cannot be described only as narrowing, qualification, added support, or wording change.

T13 — Claim Withdrawn

A previously stated claim is removed, rejected, or explicitly no longer maintained.

T99 — Unknown / Other

The revision cannot be classified confidently using the existing codes, or several transformations occur and no single primary transformation can reasonably be identified.

Coding Rule

Each revision pair receives **one primary code**.

Where more than one transformation appears to have occurred, the reviewer will select the code that best describes the main change.

If no single code clearly represents the main transformation, T99 will be used.

Reviewers will classify only what is visible in the before-and-after artifacts. They will not infer the learner's intention or internal thought process.

Relationship to H3

For the preregistered H3 analysis:

Cosmetic transformation

- T11 — Wording Only

Non-cosmetic transformation

- T01–T10
- T12–T13

Unclassified

- T99 — Unknown / Other

T99 cases will not be counted as either cosmetic or non-cosmetic in the primary H3 calculation.

These categories describe transformations in learner-authored artifacts only. They do not indicate whether the learner's reasoning, understanding, or ability improved.

Codebook Freeze

Records classified as T99 will be retained and may be reviewed after the study as candidates for Transformation Codebook v0.2.

Transformation Codebook v0.1 will not be changed during the measurement window.

Adjudication Procedure

The two independent reviewers will first classify each sampled revision pair separately. They will not see each other's classifications or the system-generated classification during this stage.

Where both reviewers assign the same code, that code will become the human reference classification for that revision pair.

Where the two reviewers assign different codes, the revision pair will be independently classified by a third reviewer. The third reviewer will not be shown the classifications assigned by the first two reviewers or the system-generated classification.

The third reviewer will classify the revision pair using Transformation Codebook v0.1 and may assign any code in the codebook, including T99 — Unknown / Other.

The third reviewer's classification will become the human reference classification for that pair.

Human inter-rater reliability for H5a will be calculated from the two original reviewers' independent classifications **before adjudication**. The adjudicated human reference classifications will be used only for the system–human comparison in H5b.

The investigator will not determine the reference classification for disputed pairs.