

When Is Checking Enough? Verification Sufficiency Calibration as a Missing Construct in Generative-AI-Supported Higher Education

Wong, Kaiyan^{1*} & Sun, Bojian¹

¹SEGi University, Malaysia

*Corresponding author email: wongkaiyan@segi.edu.my (Advisor: Wong, Kaiyan)

Received 02 September 2026, Revised 05 October 2026, Accepted 09 October 2026, Available online 11 October 2026

To link to this article: <https://doi.org/10.53797/ujssh.v5i2.18.2026>

Abstract: Generative artificial intelligence (GenAI) has made verification a central educational concern. Current research increasingly asks whether students check AI-generated claims, how well they verify them, whether their trust is calibrated, and whether verification supports learning. A less developed question concerns the termination of verification: how does a learner decide that the available evidence is sufficient and that checking can reasonably stop? This conceptual article introduces Verification Sufficiency Calibration (VSC), defined as the degree to which a learner's decision to stop verifying a GenAI-supported claim is proportionate to the evidential demands, uncertainty, consequences, and disciplinary standards of the task. A targeted integrative review of recent research on GenAI literacy, verification behaviour, epistemic vigilance, evaluative judgement, trust calibration, and epistemic dependence was conducted through 3 October 2026. Recent work has documented task-sensitive verification thresholds and explicit 'enough checking' rules; it has distinguished verification initiation, quality, success, and reliance calibration; and it has developed broader AI-literacy and assessment-literacy measures. However, the adequacy of the stopping decision itself remains under-specified as a distinct educational construct. We propose a VSC framework that distinguishes under-verification, calibrated verification, and over-verification using task-specific acceptable evidence bands rather than a universal number of checks. Six propositions specify expected relations among VSC, task stakes, accountability, time pressure, prior knowledge, source cues, and transfer. We further propose a performance-based assessment architecture combining instrumented tasks, evidence-state scoring, stopping rationales, reliance decisions, and delayed AI-withdrawal transfer. VSC extends, rather than replaces, GenAI literacy, epistemic vigilance, evaluative judgement, and trust calibration by identifying an educationally consequential regulatory decision that lies between verification activity and subsequent reliance. The framework shifts responsible-AI education from the maxim 'always verify AI' towards a more demanding competence: knowing what requires verification, how much evidence is enough, and when continued checking no longer improves epistemic responsibility.

Keywords: generative artificial intelligence, higher education, verification, AI literacy, epistemic agency

1. Introduction

Generative artificial intelligence has altered the epistemic environment of higher education. A student can now obtain an apparently coherent explanation, bibliography, interpretation, worked solution, or piece of academic prose within seconds. The educational problem goes beyond the fact that GenAI sometimes produces errors. Fluent responses can compress several activities that learners previously had to perform themselves: finding sources, comparing evidence, judging credibility, interpreting disagreement, constructing justification, and deciding when inquiry is complete.

Responsible GenAI use is accordingly framed more and more as a problem of human judgement rather than simple tool proficiency. Research on evaluative judgement argues that learners must remain able to judge the quality of outputs and processes instead of surrendering the role of arbiter to a machine (Bearman, Tai, Dawson, Boud, & Ajjaw, 2024). Research on human epistemic agency similarly emphasises adaptive epistemic stances in learner-GenAI interaction (Wu, Lee, Chai, & Tsai, 2025). More recent work describes epistemic dependence, a condition in which AI

*Corresponding author: wongkaiyan@segi.edu.my

<https://ujssh.com/> All right reserved.

becomes directive enough to displace the learner's own verification, interpretation, justification, or regulation (Du & Yuan, 2026).

Within this literature, verification has become a particularly important behaviour. Experimental evidence indicates that interventions can increase students' observable verification even when their stated intentions do not change, and that time pressure can suppress fact-checking (Vu, Cummings, & Park, 2026). Eye-tracking evidence from vocational education shows meaningful differences between GenAI-only, conventional web-search, and hybrid information-seeking conditions. In that study, the hybrid condition produced more source-focused attention and cross-verification, as well as better immediate near-transfer performance (Ma, Yue, & Li, 2026). Cross-context research has also linked AI literacy to prompting and verification behaviour, while showing that the pathways through which verification matters can differ across disciplinary environments (Shen, 2026).

These developments create a new measurement problem, because 'verification' is not one thing. Wei and Shang (2026) distinguish epistemic evaluation, verification initiation, process quality, verification success, reliance decisions, immediate performance, and independent learning. Their targeted searches produced 493 unique records. Yet among 14 priority empirical studies, they found none that jointly measured verification success and subsequent reliance calibration against an independently adjudicated standard; delayed retention and transfer were also uncommon.

A further distinction is needed. Even a student who initiates verification and performs it competently must eventually decide when to stop. Continuing indefinitely is impossible, and stopping immediately is often irresponsible. Verification thus contains an allocation problem: learners must decide how much checking a claim deserves, given constraints of time, knowledge, task importance, evidence accessibility, and the costs of error.

The empirical significance of this decision is beginning to surface. In a qualitative study of 20 pre-service science teachers, Hu, Yu, and Xu (2026) explicitly investigated 'verification stop rules' and identified hierarchical verification strategies with thresholds for 'enough checking'. These thresholds varied with perceived risk, familiarity, time pressure, and instructional responsibility. The finding shows that learners regulate how far verification should proceed, in addition to deciding whether to verify at all.

Yet the stopping phenomenon remains conceptually distributed across neighbouring constructs. GenAI-literacy frameworks include critical evaluation, and assessment-literacy scales include verification and evidence checking. Evaluative judgement concerns judging quality. Trust-calibration research asks whether reliance tracks system reliability, epistemic-vigilance research concerns protection against misinformation, and epistemic-dependence research asks whether learners retain responsibility. None of these makes the adequacy of the verification stopping point its primary object.

This article therefore develops Verification Sufficiency Calibration (VSC) as a candidate educational construct.

We define VSC as: the degree to which a learner's decision to stop verifying a GenAI-supported claim is proportionate to the task-specific evidential demand, residual uncertainty, consequences of error, and relevant disciplinary standard.

The contribution is deliberately narrower than a claim that stopping behaviour has never been observed. Hu et al. (2026) provide direct evidence of stopping rules, and recent conceptual work already recognises the problems created when AI determines what counts as 'good enough' (Du & Yuan, 2026). The contribution proposed here is to formalise the stopping decision as a separable educational target, specify its construct boundaries, define under- and over-verification, and provide a testable measurement architecture.

The article addresses three questions:

- RQ1. How do current GenAI-education frameworks conceptualise and measure verification, and where is the termination decision located within those models?
- RQ2. What distinguishes verification sufficiency calibration from verification frequency, verification success, trust calibration, AI literacy, evaluative judgement, and epistemic dependence?
- RQ3. How can VSC be operationalised and taught without reducing responsible verification to 'more checking is always better'?

2. Conceptual Background

2.1 From AI Literacy to Demonstrated Epistemic Performance

AI literacy is now a prominent educational construct, but its breadth also makes it difficult to measure. Park's (2025) systematic review of 51 empirical GenAI-literacy studies organised the field around knowing and understanding GenAI, using and applying it, evaluating and incorporating its outputs, ethical engagement, and attitudes. Such frameworks rightly treat critical evaluation as part of literacy. Evaluation, however, covers many actions and does not on its own specify when evidence gathering should end.

Measurement work has begun to move beyond broad self-perception. Jin, Martinez-Maldonado, Gašević, and Yan (2025) developed the 20-item Generative AI Literacy Assessment Test (GLAT) using responses from 355 higher-education students. Their performance-oriented measure showed satisfactory psychometric properties and predicted

GenAI task performance more strongly than self-rated proficiency. This matters for VSC, which concerns demonstrated decisions rather than students' beliefs that they 'know when to verify'.

Stockwell (2026) makes the underlying measurement issue explicit: self-reported AI literacy captures perceived competence, whereas claims about capability require performance evidence. The distinction is especially relevant to verification. Learners may confidently report that they 'always check reliable sources' yet behave very differently under time pressure or when an AI answer appears plausible.

A recent assessment-specific measure illustrates both progress and the remaining gap. Nie, Zhang, Lu, Zhang, and Zhang (2026) developed an 18-item Generative AI Assessment Literacy Scale with five dimensions, one of which is 'verification and evidence checking'. Its two validation samples comprised 486 and 798 students, respectively. The scale provides useful evidence that verification belongs within responsible assessment literacy. Its verification dimension, however, is self-reported and does not separately operationalise whether a learner's eventual stopping point is adequate for a particular claim.

Process-oriented work has also begun to track how students' GenAI literacy changes during authentic use, rather than relying only on one-off self-reports (Yan, Nakajima, & Sawada, 2025). VSC responds to two measurement needs at once: narrower construct specification and behavioural assessment.

2.2 Epistemic Vigilance

The theoretical roots of verification extend beyond AI. Sperber et al. (2010) conceptualised epistemic vigilance as the mechanisms by which people guard against acquiring misinformation through communication. Epistemic vigilance concerns the reliability of sources and the plausibility of communicated content.

GenAI intensifies this problem because it occupies an unusual communicative position. It may reproduce authoritative knowledge, synthesise contested material, fabricate support, or generate novel combinations, all in a similar rhetorical voice. Learners must therefore decide when vigilance is warranted.

Vigilance and sufficiency, however, are not equivalent. Vigilance asks whether and how credibility should be scrutinised; VSC asks when that scrutiny has generated enough warrant for the decision being made. A learner can be appropriately vigilant yet poorly calibrated in stopping. For example, a learner might notice that a claim needs checking but stop after confirming it on a derivative website that simply repeats the same unsupported assertion. This emphasis is compatible with the broader goal of developing apt epistemic performance in education (Barzilai & Chinn, 2018).

2.3 Evaluative Judgement and Epistemic Agency

Evaluative judgement is the capability to judge the quality of one's own and others' work. Bearman et al. (2024) argue that GenAI makes this capability increasingly important and identify the evaluation of AI outputs and processes as an educational priority. VSC is compatible with this approach but adds a regulatory layer: a learner may hold the standards needed to recognise strong evidence yet still allocate verification inefficiently.

Wu et al. (2025) argue for strengthening human epistemic agency in symbiotic learning relationships with GenAI, and VSC provides one behavioural expression of such agency. Agency is weakened when the learner never checks AI, but also when AI implicitly determines the point at which inquiry ends.

Du and Yuan's (2026) Epistemic Dependence in AI-Mediated Learning framework sharpens this distinction. Their analysis highlights situations in which learners accept AI outputs as epistemically directive, for example by allowing AI to settle what counts as adequate or true. VSC focuses on one specific decision within this broader phenomenon: whether the learner retains control over the evidential threshold for ending inquiry.

2.4 Trust and Appropriate Reliance

Research on trust in automation has long held that the desirable state, educationally or operationally, is appropriate reliance rather than maximal distrust. Lee and See (2004) showed that reliance decisions depend on both trust and contextual demands.

Recent GenAI research likewise distinguishes trust from calibrated reliance. Wei and Shang (2026) emphasise that a reliance decision should be interpreted against independently adjudicated output quality. The point is important: a person can verify an AI output and still make a poorly calibrated reliance decision.

VSC lies immediately upstream of this decision. A learner first accumulates or evaluates evidence, then decides that checking is sufficient, and then accepts, revises, qualifies, or rejects the AI-supported claim. A poor reliance decision can thus arise from at least two different sources:

- the learner stopped verification at an inappropriate point; or
- the learner had sufficient evidence but interpreted or used it poorly.

Conflating these pathways prevents precise educational diagnosis.

3. Method

3.1 Review Design

This article uses a targeted integrative review and a construct-development approach; it is not a systematic review or meta-analysis. Its purpose is to identify the strongest applicable neighbouring constructs, determine whether the termination of verification is already adequately represented, and, where a conceptual gap remains, construct a falsifiable educational framework.

The literature search and source verification were updated through 3 October 2026. Searches combined GenAI terms with terms representing the principal neighbouring constructs, including:

- 'generative AI' AND education AND verification
- ChatGPT AND 'verification behaviour'
- 'verification stopping rules'
- 'evidential sufficiency'
- 'verification quality'
- 'trust calibration'
- 'reliance calibration'
- 'epistemic vigilance'
- 'epistemic dependence'
- 'evaluative judgement' AND generative AI
- 'AI literacy assessment'
- 'verification and evidence checking'
- 'epistemic agency' AND generative AI

Searches drew on scholarly web search, publisher platforms, DOI and index records, PubMed where applicable, institutional repositories, and reference chaining. Recent literature (2023-2026) was prioritised for direct GenAI applications, while foundational research was retained where it defined a neighbouring theoretical construct.

3.2 Inclusion Logic

Studies were given priority when they contributed directly to at least one of the following:

- observable verification behaviour in GenAI-supported tasks;
- verification process quality or success;
- evidence standards or stopping rules;
- trust or reliance calibration;
- GenAI-literacy measurement;
- evaluative judgement or epistemic agency;
- epistemic dependence;
- transfer of verification-related learning;
- contextual variables likely to change reasonable verification thresholds.

The aim was conceptual coverage of the nearest-neighbour literature, not an exhaustive enumeration of every GenAI-education study.

3.3 Analytic Procedure

For each closely related study or framework, five questions were asked:

1. What is the target construct?
2. Is verification represented as intention, initiation, process, success, stopping, reliance, or learning?
3. Is the measure self-reported or performance-based?
4. Does the design identify a defensible criterion for 'enough' checking?
5. What task conditions change the expected amount or quality of verification?

A candidate gap was retained only if it remained distinguishable from the strongest neighbouring constructs.

3.4 Scope and Evidential Status

No new human participants were recruited, no student dataset was collected, and no empirical effect estimate is reported in this article. The proposed VSC construct and measurement architecture are therefore theoretical propositions that require empirical validation. The article does not claim that VSC has already shown factorial validity, predictive validity, instructional responsiveness, or cross-cultural invariance.

4. Synthesis: The Missing Decision Between Verification and Reliance

4.1 Current Work Increasingly Separates 'Checking' into Different Targets

A major advance in recent literature is the recognition that verification should not be represented as a single behaviour. Wei and Shang (2026) distinguish epistemic evaluation, verification initiation, process quality, verification success, reliance decisions, immediate task performance, and independent learning. Their framework guards against a common inferential error. Observing that a student opened external sources does not establish that verification succeeded, and successful verification does not establish that later reliance was calibrated.

The proposed VSC construct adds a complementary distinction:

evaluation -> verification initiation -> verification process -> evidence state -> sufficiency judgement/stop decision -> reliance action -> performance/learning

What this adds is the transition from an evidence state to the judgement that further verification is, or is not, warranted.

4.2 Verification Initiation Does Not Reveal Verification Sufficiency

Vu et al. (2026) provide experimental evidence that students can be induced to verify more, although their stated verification intentions do not necessarily shift accordingly. They also found that time pressure reduced fact-checking.

This study illustrates why counting verification actions is useful but incomplete. Suppose two students each consult one external source. One is checking a low-stakes definition whose accuracy is independently established by an authoritative reference. The other is checking a consequential medical or legal claim whose single accessible source merely repeats the AI output. Their verification counts are identical, but the adequacy of their stopping decisions is not.

Conversely, a student who checks six sources is not necessarily more epistemically competent than one who checks two. The six sources may be mutually dependent, irrelevant, or redundant, and more verification can also impose unnecessary cognitive and time costs.

VSC therefore rejects the assumption: more checking = better verification.

Instead: better verification = evidence accumulation and termination appropriately matched to the epistemic demands of the task.

4.3 Process Quality Does Not Fully Determine Stopping Quality

Ma et al. (2026) provide unusually rich process evidence by combining eye tracking, behavioural logs, questionnaires, and task scores. In their 96-participant experiment, the hybrid-search condition showed the highest proportion of attention to source evidence and the highest rate of cross-verification attempts. It also produced stronger immediate near-transfer performance than the alternative conditions.

These results indicate that attention to evidence and cross-checking are meaningful process variables. Process quality and stopping calibration nevertheless remain analytically separable. One learner may execute a sophisticated search strategy but stop before addressing a decisive contradiction. Another may keep performing high-quality searches after all task-relevant uncertainty has been adequately resolved.

VSC is therefore concerned with the fit between the terminal evidence state and the evidential requirement, rather than with the sheer volume of search behaviour.

4.4 Existing Qualitative Evidence Already Reveals Stopping Rules

The closest empirical neighbour to VSC is Hu et al. (2026). Their pre-service science teachers used hierarchical verification strategies and expressed explicit thresholds for 'enough checking'. The thresholds varied with the task, perceived risk, familiarity, time pressure, and responsibility towards future students.

This study changes the status of the research question. Learners already use stopping rules, so verification stopping is more than a hypothetical theoretical issue. The unresolved issue is whether those rules can be:

- defined as a distinct educational competence;
- evaluated against defensible task-specific standards;
- distinguished from general verification frequency;
- improved through instruction; and
- shown to transfer to new tasks and to settings in which AI support is unavailable.

VSC thus formalises an observed phenomenon; it does not invent an unsupported behavioural category.

4.5 Broader Literacy Measures Remain Too Coarse for This Decision

The GLAT provides evidence that GenAI literacy can be assessed through performance-oriented tasks rather than self-report alone (Jin et al., 2025), while GAA-LS embeds verification within a multidimensional assessment-literacy structure (Nie et al., 2026).

These instruments serve broader purposes. VSC is narrower and should not compete with them as a replacement 'AI literacy' scale; its value depends on avoiding construct inflation.

A learner can, for example:

- understand GenAI hallucination risk;
- know how to open reliable databases;
- distinguish scholarly from non-scholarly sources;
- report strong verification habits;

yet still have difficulty determining when the evidential case is sufficiently strong to stop. That specific failure is the target of VSC.

Table 1. Construct boundaries surrounding Verification Sufficiency Calibration

Construct	Central question	Typical evidence	What it does not by itself establish
GenAI literacy	Can the learner understand, use, evaluate, and ethically engage with GenAI?	Knowledge tests, performance tasks, self-report	Whether a particular verification stopping point was appropriate
Epistemic vigilance	Should this information/source be trusted or scrutinised?	Credibility evaluation, plausibility judgement	How much evidence is sufficient before inquiry stops
Verification initiation	Did the learner begin checking?	Clicks, searches, source visits	Whether checking was adequate
Verification process quality	How well was verification conducted?	Source diversity, relevance, cross-checking, process logs	Whether continued checking was still necessary
Verification success	Was the claim correctly verified/falsified?	Accuracy against adjudicated ground truth	Whether the learner allocated verification efficiently
Evaluative judgement	Can the learner judge quality?	Comparison and judgement tasks	How verification effort should be terminated under constraints
Trust/reliance calibration	Does reliance match actual system/output quality?	Accept/reject/revise behaviour relative to ground truth	Whether preceding evidence gathering stopped appropriately
Epistemic dependence	Has AI displaced learner epistemic responsibility?	Patterns of judgement displacement	A task-level metric of verification termination
Verification Sufficiency Calibration	Did the learner stop checking at a defensible point for this task?	Evidence-state trajectory, stopping point, rationale, task criterion, reliance action	Broader AI literacy or all forms of epistemic agency

5. Verification Sufficiency Calibration

5.1 Definition

We define Verification Sufficiency Calibration (VSC) as the degree to which the point at which a learner terminates verification corresponds to a defensible level of evidence for the claim, task, uncertainty, stakes, and disciplinary context at hand.

Four elements are important.

First, VSC is relational rather than a fixed personal trait. A learner can be well calibrated on familiar historical-source tasks and poorly calibrated on unfamiliar statistical claims.

Second, VSC is normative but contextual. There must be some defensible educational standard against which stopping can be judged, but that standard varies across tasks.

Third, VSC is behavioural and epistemic, not only motivational. A learner may be highly motivated to verify yet still stop poorly.

Fourth, VSC concerns both too little and too much verification.

5.2 Under-Verification

Under-verification occurs when a learner terminates checking before the evidence state is adequate for the decision being made.

Examples include:

- accepting an AI-generated citation after confirming only that a similarly titled webpage exists;

- treating one secondary summary as sufficient evidence for a contested claim;
- accepting a numerical result without examining the assumptions or source data when the assignment requires analytical justification;
- stopping after finding confirmatory evidence while ignoring a salient contradiction;
- using the presence of citations as a substitute for inspecting whether they support the claim.

Under-verification is not the same as no verification. A learner can perform several verification actions and still stop prematurely.

5.3 Over-Verification

Over-verification occurs when a learner continues checking substantially beyond the evidence required for a defensible educational judgement, without a corresponding reduction in decision-relevant uncertainty.

This category matters because educational advice often assumes that verification has only benefits. In real learning environments, however, attention and time are limited. Excessive checking may:

- reduce time available for analysis and synthesis;
- encourage unproductive certainty seeking;
- substitute information accumulation for judgement;
- make routine low-risk tasks unnecessarily costly;
- discourage independent decision making.

The ideal is appropriately calibrated verification, not maximal verification.

5.4 Calibrated Verification

Calibrated verification occurs when the learner reaches a defensible evidence threshold and stops within an acceptable range, rather than at a single universally 'correct' moment.

A range is preferable to a point because authentic academic inquiry often permits multiple reasonable verification pathways. Two students may consult different evidence while reaching equally defensible conclusions.

6. A Task-Specific Evidence-Band Model

6.1 Why a Universal Rule Is Unsuitable

Rules such as 'always check three sources' are pedagogically attractive but epistemically weak. Three dependent sources can provide less warrant than one authoritative primary record. Likewise, one source may be entirely sufficient to confirm a stable bibliographic fact but plainly insufficient to resolve a disputed causal claim.

VSC should therefore be judged against a task-specific acceptable evidence band.

For task i , let the educationally adjudicated acceptable terminal evidence range be:

$$[L_i, U_i] \quad (1)$$

where L_i represents the minimum evidence state at which stopping becomes defensible, and U_i represents a point beyond which additional checking ordinarily adds negligible decision-relevant value under the stated task conditions.

Let E_i denote the evidence state attained by the learner when verification stops.

Under-verification can then be represented as:

$$UV_i = \max(0, L_i - E_i) \quad (2)$$

and over-verification as:

$$OV_i = \max(0, E_i - U_i) \quad (3)$$

A simple calibration loss is:

$$CL_i = w_i(UV_i + OV_i) \quad (4)$$

where w_i is a task-specific consequence or importance weight. Higher VSC corresponds to lower calibration loss across a set of tasks.

6.2 Evidence State Is Not Source Count

E_i should not be interpreted as the number of webpages opened. It represents progress on an adjudicated evidence rubric. Depending on the discipline and task, an evidence state may incorporate:

- source authority;
- independence among sources;

- relevance to the exact claim;
- primary versus derivative evidence;
- methodological quality;
- triangulation;
- coverage of plausible counterevidence;
- consistency with disciplinary criteria;
- residual uncertainty;
- traceability from claim to supporting evidence.

A student could thus reach a higher evidence state with two well-chosen primary sources than with ten mutually derivative webpages.

6.3 Multiple Valid Pathways

The framework should avoid imposing a single prescribed search sequence. VSC is concerned with where inquiry ends and why; it does not require every learner to reproduce an expert's exact navigation path.

For assessment purposes, educators can specify:

1. the claim or problem;
2. the relevant evidential demands;
3. plausible evidence pathways;
4. minimum conditions that must be satisfied;
5. evidence that would materially change the decision;
6. the acceptable range for stopping.

This approach preserves authentic judgement while making assessment possible.

7. A Process Model of VSC

We propose the following process:

Task and claim characteristics -> initial uncertainty and perceived stakes -> verification initiation decision -> evidence acquisition and evaluation -> monitoring of residual uncertainty -> sufficiency judgement -> stop/continue decision -> reliance action: accept, revise, qualify, reject, or defer -> immediate performance and later independent transfer

The model contains several potential failure points.

- A student may fail to initiate verification.
- A student may verify using poor-quality evidence.
- A student may misread good evidence.
- A student may stop too early.
- A student may continue without information gain.
- A student may reach adequate evidence but still make a poor reliance decision.

Separating these points allows more precise pedagogical interventions.

8. Propositions

8.1 Proposition 1: Verification Quantity and VSC Will Be Empirically Distinguishable

Students who perform more verification actions will not necessarily show better VSC. Once differences in evidence quality and task requirements are taken into account, verification frequency should explain only part of the variance in stopping calibration.

This proposition directly rejects a 'more is better' model of responsible AI use.

8.2 Proposition 2: Higher Task Stakes and Accountability Will Raise Defensible Verification Thresholds

When the consequences of error are greater, L_i should generally increase.

Hu et al. (2026) provide initial qualitative support for this logic: verification thresholds were sensitive to perceived risk and to the responsibility associated with future teaching. Wei and Shang (2026) also identify task stakes and accountability as boundary conditions in verification and reliance research.

Educationally, this means that 'appropriate verification' for brainstorming is not identical to appropriate verification for assessment evidence, health information, research claims, or material presented to students.

8.3 Proposition 3: Time Pressure Will Increase Under-Verification When Learners Cannot Efficiently Prioritise Decisive Evidence

Vu et al. (2026) found that time pressure reduced fact-checking. VSC predicts a more specific effect: time pressure should increase premature stopping most strongly among learners who lack strategies for identifying high-information verification actions.

This implies an instructional interaction. Teaching learners merely to 'check more' may fail under time pressure, whereas teaching them to identify the next most decision-relevant check should be more robust.

8.4 Proposition 4: Apparent Source Credibility Can Lower Stopping Thresholds Prematurely

Verification is partly cue driven. Knor, Reiss, Möller, and Merten (2026), in a preregistered conjoint experiment involving 1,417 participants, found that high-credibility sources reduced the likelihood of verification in a generative-search context.

For education, this implies that citations, institutional-looking interfaces, polished language, or familiar source names may act as stopping cues even when the relevant claim has not been established.

VSC therefore requires learners to distinguish 'This looks credible' from 'I now possess sufficient evidence for the judgement required by this task.'

8.5 Proposition 5: Criteria-Based Stopping Instruction Will Outperform Generic Verification Exhortations on VSC

Generic warnings may increase caution or verification initiation, as the inoculation findings of Vu et al. (2026) suggest. VSC, however, predicts that instruction based on explicit evidential criteria will better calibrate when learners stop.

An intervention should therefore teach questions such as:

- What exactly must be established?
- What evidence would materially change my conclusion?
- Have I inspected a source that is independent of the AI output?
- Is the evidence primary enough for the claim?
- Have I addressed the most plausible counterevidence?
- What uncertainty remains?
- Would one more check be likely to change my decision?

The goal is not indiscriminately longer verification. A successful intervention should reduce both premature stopping and wasteful over-checking.

8.6 Proposition 6: VSC Should Predict Independent Transfer Beyond Same-Tool Task Performance

Immediate task success with AI access does not necessarily show durable competence. Wei and Shang (2026) found delayed retention and transfer to be uncommon in the studies they reviewed, and Ma et al. (2026) provide useful near-transfer evidence, but only within assigned information-seeking conditions.

A strong validation of VSC should therefore test whether learners can transfer sufficiency judgements to:

- unfamiliar topics;
- changed disciplines;
- different AI systems;
- tasks in which source access differs; and
- tasks completed without GenAI.

9. Measuring VSC

9.1 Why Self-Report Alone Is Insufficient

Items such as 'I know when I have checked enough sources' may measure confidence, metacognitive belief, or perceived competence. They cannot establish whether the learner actually stops at defensible points. This reflects the broader measurement critique that perceived and demonstrated AI competence should not be treated as interchangeable (Stockwell, 2026).

Self-report remains useful as a secondary measure. The gap between perceived and demonstrated VSC may itself be educationally important.

9.2 Proposed Performance-Based VSC Assessment

A VSC assessment could present learners with a series of GenAI-supported tasks in which they control the verification process.

Each task would contain:

1. an initial AI response;
2. a claim whose reliability varies;
3. optional evidence sources of different diagnostic value;
4. a cost associated with opening or examining additional evidence;
5. a 'stop verification' option;
6. a final reliance decision;
7. a written or verbal stopping rationale.

The system would record:

- whether verification was initiated;
- which evidence was selected;
- order of evidence acquisition;
- time spent;
- transitions among AI output and external evidence;
- contradictions noticed;
- evidence state at termination;
- stopping rationale;
- final reliance decision;
- confidence;
- final task accuracy.

9.3 Construct Validation

A validation programme should test whether VSC is empirically distinguishable from:

- GenAI literacy;
- domain knowledge;
- information literacy;
- verification frequency;
- source-evaluation skill;
- need for cognition;
- epistemic beliefs;
- self-regulated learning;
- trust in AI;
- reliance calibration.

Evidence of discriminant validity would require more than statistical factor separation; experimental manipulations should also produce theoretically discriminating patterns.

For example, generic verification reminders may raise verification initiation without substantially improving VSC. Criteria-based stopping instruction should improve VSC more specifically.

9.4 Reliability

Traditional internal consistency is not necessarily the main reliability target, because VSC is intentionally context-sensitive. Reliability should instead consider:

- generalisability across parallel task sets;
- stability within comparable evidence environments;
- rater agreement when evidence states require expert judgement;
- sensitivity to theoretically relevant contextual changes.

Perfect cross-task stability would not necessarily be desirable, since a learner who uses exactly the same stopping threshold for every task may be poorly calibrated.

9.5 Expert Adjudication

A central challenge is constructing $[L_i, U_i]$.

For each task, expert panels should identify:

- the knowledge claim;
- the decision the student must make;
- minimum acceptable evidence;
- decisive evidence;

- plausible misleading evidence;
- legitimate alternative evidence pathways;
- conditions under which continued checking remains warranted.

Where experts disagree, the assessment should preserve an interval or multiple defensible pathways rather than impose false precision.

10. Implications for Curriculum Design

10.1 Replace 'Verify Everything' with Risk-Sensitive Verification

'Always verify AI' is safe as a slogan but weak as a curriculum.

A more useful educational principle is that verification effort should increase with uncertainty, consequences, contestability, and the probability that an additional check will change the decision.

Students therefore need to learn verification triage.

A low-stakes brainstorming suggestion may require little checking. A quotation requires confirmation against the original source. A statistical claim may require access to underlying methods or data. A contested theoretical claim may require contrasting scholarly interpretations. A reference intended for publication may require bibliographic and substantive verification.

10.2 Teach Verification as Sequential Decision Making

Verification activities are often taught as checklists: inspect author, date, evidence, domain, citations, and so forth.

VSC suggests a sequential alternative:

Step 1: Specify the claim. What exactly am I trying to establish?

Step 2: Specify the consequence. What happens if this claim is wrong?

Step 3: Identify the highest-information check. Which next action would most change my judgement?

Step 4: Update uncertainty. What did this evidence resolve, and what remains uncertain?

Step 5: Apply a stopping criterion. Is remaining uncertainty tolerable for this particular use?

Step 6: State the epistemic status. Accept, reject, revise, qualify, or defer.

This routine teaches students that responsible inquiry includes ending inquiry responsibly.

10.3 Use Contrasting Cases

Instruction can make sufficiency visible through paired cases.

Case A: A student stops after one highly relevant primary source.

Case B: A student consults seven websites that all reproduce the same unsourced claim.

Students should decide which verification is stronger and explain why.

A second comparison can target over-verification.

Case C: A student continues searching for twenty minutes after an authoritative primary record has conclusively established a routine bibliographic fact.

This helps convey that verification quality cannot be reduced to verification quantity.

10.4 Require Stopping Rationales

Assignments can ask students to document what they checked and also to answer: 'Why was this evidence sufficient for the way you used the claim?'

This prompt shifts attention from procedural compliance to epistemic justification.

10.5 Design Assessment for Retained Judgement

Du and Yuan's epistemic-dependence framework stresses that AI integration should preserve learner responsibility, not merely regulate tool permission. VSC offers a concrete assessment target for this objective.

Assessment could therefore include AI-supported phases followed by:

- source-withdrawal tasks;
- AI-withdrawal tasks;
- altered claims;
- unfamiliar-domain tasks;
- oral defence of stopping decisions.

The aim is to determine whether the student, rather than the interface, controls the evidential threshold.

11. Implications for Teacher Education

VSC may be particularly important for teacher preparation.

Teachers do more than consume information for themselves. They turn information into explanations, assessment materials, examples, feedback, and instructional decisions that affect others.

Hu et al. (2026) found that pre-service science teachers' verification thresholds were shaped by instructional responsibility. This suggests that professional role can change what counts as sufficient checking.

Teacher education could therefore explicitly contrast:

- 'sufficient for my private brainstorming';
- 'sufficient for my assignment';
- 'sufficient to teach as factual content';
- 'sufficient to assess a student';
- 'sufficient to recommend as a learning resource'.

Such distinctions move AI literacy closer to professional epistemic responsibility.

12. Research Agenda

12.1Phase 1: Construct Elicitation

The first empirical phase should examine how students naturally express sufficiency judgements.

Think-aloud protocols can ask learners:

- Why did you continue?
- What evidence were you missing?
- What made you stop?
- What would make you reopen the question?
- What would you require if the task had greater consequences?

The purpose is to determine whether learners distinguish sufficiency in ways consistent with the proposed construct, not to 'prove' the framework.

12.2Phase 2: Task-Bank Development

Researchers should then build a task bank with systematically varied:

- output correctness;
- error subtlety;
- source traceability;
- controversy;
- stakes;
- time pressure;
- prior-knowledge demands;
- evidence redundancy;
- disciplinary norms.

Each task should be independently adjudicated to determine acceptable evidence pathways and stopping bands.

12.3Phase 3: Construct Validation

A multi-sample validation study should examine convergent and discriminant relations with existing AI-literacy, verification, trust, and epistemic-judgement measures.

A particularly informative test would compare:

- perceived VSC;
- demonstrated VSC;
- general GenAI literacy;
- verification count;
- verification success;
- reliance calibration.

A distinct VSC construct should contribute information not captured by those measures alone.

12.4Phase 4: Instructional Experiment

A useful randomised intervention could compare:

- business-as-usual AI-literacy instruction;
- generic verification warning;
- verification-procedure instruction;
- VSC instruction using explicit sufficiency criteria and stopping rationales.

Primary outcomes should include under-verification loss and over-verification loss rather than verification count alone. Secondary outcomes could include verification success, final decision accuracy, task completion time, reliance calibration, perceived cognitive load, confidence, and delayed transfer.

12.5 Phase 5: Tool-Withdrawal Transfer

Students should later complete tasks without access to the AI system used during training.

If VSC is an educational competence rather than an interface habit, trained students should make stronger evidence-allocation and stopping decisions even when GenAI is absent.

12.6 Phase 6: Cross-Disciplinary and Cross-Cultural Validation

Shen's (2026) cross-context findings suggest that verification can function differently across disciplinary environments. VSC should therefore not be validated in one discipline and then generalised immediately to higher education as a whole.

A future programme could begin in a Malaysian university context and deliberately sample several disciplinary areas before testing invariance in other institutions or national settings. Such work would require prospective ethics approval. It should be treated as a later empirical project, not as evidence already produced by the present conceptual article.

13. Discussion

13.1 Theoretical Contribution

The central argument of this article is that responsible GenAI use contains a stopping problem.

Existing literature already establishes the importance of critical evaluation, epistemic vigilance, evaluative judgement, verification behaviour, verification quality, reliance calibration, human epistemic agency, and avoiding epistemic dependence.

The proposed VSC framework does not compete with these ideas; it identifies a regulatory decision that connects them.

The theoretical sequence can be expressed as:

Need to verify -> How to verify -> What evidence has been obtained -> Is this evidence sufficient? -> How should I rely on the output?

The 'Is this evidence sufficient?' decision is the VSC problem.

13.2 Why This Matters Educationally

Without this distinction, educators may unintentionally reward visible activity instead of judgement.

A student may receive credit for:

- opening many sources;
- producing a long verification log;
- reporting distrust of AI;
- repeating that AI can hallucinate.

None of these behaviours shows that the learner knows when the available evidence is sufficient for the purpose at hand.

Education should therefore aim for epistemic proportionality. Learners should invest scrutiny in proportion to the claim and be able to justify both continuing an investigation and ending it.

13.3 Efficiency Is Not the Enemy of Epistemic Responsibility

Including over-verification may seem counterintuitive, because most current concern centres on uncritical acceptance.

Responsible learning, however, operates under finite resources, and academic expertise partly consists of knowing where uncertainty matters.

A mature learner does not maximise doubt. A mature learner distinguishes:

- claims that can reasonably be accepted;
- claims requiring direct source inspection;
- claims requiring triangulation;
- claims that remain uncertain after reasonable inquiry;
- claims for which the correct action is to qualify or defer judgement.

VSC thus integrates responsibility with efficiency instead of treating them as opposites.

13.4VSC Also Reframes Uncertainty

A stopping decision does not always produce 'true' or 'false'.

Sometimes verification should end with:

- 'evidence is mixed';
- 'the claim is plausible but insufficiently supported';
- 'the available sources cannot resolve this question';
- 'I will not use the claim in its current form'.

Recognising unresolved uncertainty may reflect better VSC than continuing to search until confirmatory material appears.

This feature helps guard against motivated verification, in which checking becomes a search for permission to accept a preferred answer.

14. Strongest Alternative Explanations and Boundary Conditions

14.1'VSC Is Simply Self-Regulated Learning'

Self-regulated learning includes planning, monitoring, strategy selection, and control. VSC can be located within regulation, but it specifies an unusually precise target: calibration of the threshold at which evidence checking ends. If VSC measures add no explanatory or predictive information beyond established self-regulation measures, the case for treating VSC separately would weaken.

This is an empirical falsification condition.

14.2'VSC Is Simply Verification Quality'

Verification quality concerns what happens during checking. VSC concerns whether the learner terminates at an appropriate evidence state.

The two should correlate, but they need not be identical. Strong process quality followed by premature stopping, and strong process quality followed by unnecessarily prolonged checking, are both theoretically possible.

If empirical studies show that a well-validated verification-quality measure fully accounts for VSC behaviour, the narrower construct would be unnecessary.

14.3'VSC Is Simply Reliance Calibration'

Reliance calibration evaluates whether the eventual reliance decision is appropriate given actual output quality.

A learner could reach adequate evidence but still misinterpret it, producing poor reliance calibration. Conversely, a learner could stop too early yet reach the correct reliance decision by chance.

The constructs therefore concern different transitions. Again, this distinction must be shown empirically rather than assumed.

14.4'Experts Cannot Define Enough Evidence'

This is the most serious methodological objection.

Many academic questions do not have one uniquely correct evidential threshold. VSC therefore uses an acceptable evidence band, permits multiple valid pathways, and makes task assumptions explicit.

For highly open-ended inquiry, quantitative VSC scoring may be inappropriate. The framework is best suited to settings in which educators can specify what evidence is needed to support a particular decision.

14.5'Over-Verification Could Reward Superficial Efficiency'

Poorly designed VSC assessment could indeed mistake fast decisions for efficient decisions.

Speed is therefore not itself evidence of calibration. A learner stops appropriately only when the terminal evidence state is defensible. Time is a secondary cost measure, not the criterion of quality.

15. Limitations

This article has several important limitations.

First, VSC is a proposed construct and has not yet undergone empirical construct validation. Its conceptual distinctions may prove less separable in observed learner behaviour than they appear in theory.

Second, the literature synthesis was targeted and integrative rather than systematic. Searches were designed to locate the strongest neighbouring frameworks and recent direct evidence, not to estimate prevalence or provide exhaustive bibliometric coverage. Claims of novelty are therefore deliberately limited to the literature reviewed through 3 October 2026.

Third, the normative definition of evidence sufficiency creates an adjudication challenge. Different disciplines legitimately use different evidential standards, and experts may disagree. Task-specific bands and transparent rubrics reduce this problem but do not eliminate it.

Fourth, under- and over-verification may be asymmetric. In high-consequence settings, the cost of premature stopping may substantially exceed the cost of additional checking. VSC should therefore not assume equal penalties.

Fifth, accessibility and equity require explicit attention. Students differ in database access, prior disciplinary knowledge, language proficiency, disability-related access needs, and time resources. A reasonable stopping threshold must be distinguished from an assumption that every learner has equal access to the same verification infrastructure.

Finally, the rapid development of GenAI systems may change the verification environment. Better source traceability, retrieval grounding, or provenance tools may alter which verification actions are necessary, although the higher-level sufficiency judgement is likely to remain educationally relevant.

16. Conclusion

Higher education has moved beyond asking whether students should trust generative AI. The emerging question is how students can retain responsibility for knowledge-related judgements while using systems that produce fast, fluent, and often persuasive answers.

Research has substantially improved understanding of GenAI literacy, evaluative judgement, epistemic agency, verification behaviour, verification success, trust, and reliance. Recent evidence also shows that learners already use different thresholds to decide when checking is 'enough' (Hu et al., 2026).

This article argues that the adequacy of that stopping decision deserves separate educational attention.

Verification Sufficiency Calibration describes whether learners stop checking at a defensible point, given the evidence required, residual uncertainty, task stakes, and disciplinary context. It distinguishes premature stopping from calibrated stopping and from unnecessary continued checking.

The educational objective is thus neither maximal scepticism nor maximal verification, but proportionate epistemic responsibility.

A student prepared for AI-mediated higher education should be able to answer four questions:

1. What exactly needs to be verified?
2. What evidence would meaningfully change my judgement?
3. Do I now have sufficient evidence for the way I intend to use this claim?
4. If not, should I continue, qualify the claim, reject it, or defer judgement?

The capacity to answer the third question (when is checking enough?) is the central competence that VSC places on the research agenda.

Ethics Statement

This article is a conceptual and integrative review and reports no newly recruited human participants, interventions, identifiable participant information, or newly collected participant data. No participant-level empirical study is represented as having been conducted in this manuscript. Any subsequent empirical validation of VSC involving students would require prospective ethics review and appropriate consent procedures.

Data Availability Statement

No new empirical dataset was generated for this conceptual article. The scholarly works used to develop the framework are identified in the reference list.

CRedit Authorship Contribution Statement

Wong Kaiyan: Conceptualization (50%), Methodology (50%), Investigation (50%), Formal analysis (50%), Writing - original draft (50%), Writing - review & editing (50%). Sun Bojian: Conceptualization (50%), Methodology (50%), Investigation (50%), Formal analysis (50%), Writing - original draft (50%), Writing - review & editing (50%). Both authors contributed equally to the work.

Funding

This research received no external funding.

Generative AI and AI-Assisted Technologies Disclosure

A generative-AI system was used to support targeted literature discovery, evidence mapping, construct comparison, and manuscript drafting. The manuscript was designed to preserve the evidential status of cited research and does not present AI-generated numerical results as empirical findings. Before journal submission, the named authors should independently verify every reference, quotation, factual claim, journal-specific disclosure, and final wording, and assume responsibility for the submitted version.

Conflict of Interest

The authors declare no conflicts of interest.

References

- Barzilai, S., & Chinn, C. A. (2018). On the goals of epistemic education: Promoting apt epistemic performance. *Journal of the Learning Sciences*, 27(3), 353-389. <https://doi.org/10.1080/10508406.2017.1392968>
- Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 49(6), 893-905. <https://doi.org/10.1080/02602938.2024.2335321>
- Du, Y., & Yuan, Y. (2026). Epistemic dependence in AI-mediated learning within higher education: A framework for student judgement and responsibility. *Assessment & Evaluation in Higher Education*. Advance online publication. <https://doi.org/10.1080/02602938.2026.2732250>
- Hu, H., Yu, S., & Xu, L. (2026). Evidence, trust, and objectivity with generative AI: A qualitative interview study of pre-service science teachers' truth-assessment practices. *Frontiers in Psychology*, 17, 1781772. <https://doi.org/10.3389/fpsyg.2026.1781772>
- Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence*, 9, 100436. <https://doi.org/10.1016/j.caeai.2025.100436>
- Knor, E. L., Reiss, M. V., Möller, J., & Merten, L. (2026). Determinants of verification behavior in generative search: Evidence from a conjoint experiment. *Computers in Human Behavior Reports*, 22, 101056. <https://doi.org/10.1016/j.chbr.2026.101056>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392
- Ma, J., Yue, H., & Li, X. (2026). Responsible use of generative AI in Chinese vocational education: Eye-tracking evidence on source verification and immediate near-transfer across three information-seeking conditions. *Frontiers in Psychology*, 17, 1946196. <https://doi.org/10.3389/fpsyg.2026.1946196>
- Nie, J., Zhang, Z., Lu, X., Zhang, Y., & Zhang, M. (2026). Development and validation of the generative AI assessment literacy scale for higher education students: Psychometric evidence and associations with feedback engagement and academic integrity. *Frontiers in Education*, 11, 1934632. <https://doi.org/10.3389/educ.2026.1934632>
- Park, J. (2025). A systematic literature review of generative artificial intelligence (GenAI) literacy in schools. *Computers and Education: Artificial Intelligence*, 9, 100487. <https://doi.org/10.1016/j.caeai.2025.100487>
- Shen, Q. (2026). Same AI, different pathways: Unpacking mechanisms of AI-mediated learning across discipline-institution contexts. *Computers and Education: Artificial Intelligence*, 11, 100601. <https://doi.org/10.1016/j.caeai.2026.100601>
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359-393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>
- Stockwell, G. (2026). Confidence is not competence: Self-report and the measurement of AI literacy. *Research Methods in Applied Linguistics*, 5(3), 100364. <https://doi.org/10.1016/j.rmal.2026.100364>
- Vu, C. B., Cummings, J. J., & Park, D. Y. (2026). Student engagement with ChatGPT for educational tasks: Effects of inoculation training on verification intentions and behavior. *Computers and Education Open*, 10, 100335. <https://doi.org/10.1016/j.cao.2026.100335>
- Wei, J., & Shang, Y. (2026). Beyond checking: Verification quality, reliance calibration, and learning in generative AI-assisted higher education. *Frontiers in Psychology*, 17, 1965371. <https://doi.org/10.3389/fpsyg.2026.1965371>
- Wu, J.-Y., Lee, Y.-H., Chai, C. S., & Tsai, C.-C. (2025). Strengthening human epistemic agency in the symbiotic learning partnership with generative artificial intelligence. *Educational Researcher*, 54(6), 358-368. <https://doi.org/10.3102/0013189X251333628>
- Yan, W., Nakajima, T., & Sawada, R. (2025). Beyond tool use: Tracking the evolution of generative AI literacy among university students through a process-oriented investigation. *Computers and Education: Artificial Intelligence*, 9, 100465. <https://doi.org/10.1016/j.caeai.2025.100465>