Woodworth, J., Kharbach, M., & Doe, C. (2026). Validity Architecture for AI Integrated Assessment (VAAI): A conceptual framework for defensible assessment practice. *Intersection: A Journal at the Intersection of Assessment and Learning, Early View*.

Validity Architecture for AI Integrated Assessment (VAAI): A Conceptual Framework for Defensible Assessment Practice

Johanathan Woodworth, Ph.D., Mohamed Kharbach, Ph.D., Christine Doe, Ph.D.

Author Note

Johanathan Woodworth, https://orcid.org/0000-0002-3656-3329
Mohamed Kharbach, https://orcid.org/0000-0001-7840-7474
Christine Doe, https://orcid.org/0000-0003-1538-6041
We have no conflicts of interest to disclose

Intersection: A Journal at the Intersection of Assessment and Learning
Early View

Abstract: Generative AI has weakened the link between submitted work and warranted claims about learner competence, creating an inference problem that detection and prohibition alone cannot resolve. This article introduces the Validity Architecture for AI Integrated Assessment (VAAI), a conceptual framework for redesigning assessment when AI assistance may shape student work. Grounded in argument-based validity and construct validity theory, VAAI treats AI-related assessment challenges as problems of inference, evidence, and interpretation rather than primarily as problems of academic integrity. For practitioners, it organizes assessment design around three questions. First, what competence is being assessed, under what conditions, for what decisions, and with what consequences? Second, what evidence is required to justify the claim? Third, could a student satisfy the assessment standard through AI-mediated strategies without demonstrating the intended competence? VAAI combines a consequence-calibrated specification guide across eight dimensions with an evidence model centered on provenance, adversarial rebuttal testing, and implementation review. Worked examples from nursing, counselling, and graduate capstone review illustrate how the same validity logic applies across disciplines and consequence levels. The article contributes a structured validity architecture for making interpretive claims, assistance boundaries, and evidentiary requirements more explicit, contestable, and defensible in higher education assessment practice.

Keywords: *assessment practice, generative AI, validity, assessment design, higher education*

Introduction

Generative artificial intelligence has not simply made academic misconduct easier. Across higher education contexts, it has fundamentally shifted what student work can demonstrate about ability (Eaton, 2023; Kaldaras et al., 2024; Swiecki et al., 2022). Whether the task is a first-year lab report, an undergraduate research essay, or a graduate capstone proposal, the same difficulty appears: the submitted artifact may no longer provide sufficient evidence of what the learner knows or can do (Gouseti et al., 2025; Southworth et al., 2023). The institutional consequences vary across courses, programs, and professional pathways, but the underlying validity problem does not.

Assessment practice has long relied on a basic premise: that submitted work could be attributed to the student, thereby supporting claims about competence (Eaton, 2023). GenAI unsettles that premise because it can contribute substantive conceptual and rhetorical work, not just surface correction or formatting support (Dawson et al., 2024; Luo, 2024; Miao & Holmes, 2023). Once competence claims depend on artifacts with unclear origins, mismatches between claims and evidence become harder to avoid (Kane, 2013; St-Onge et al., 2017). Many institutions have treated this as a compliance problem, responding with prohibition, detection, and sanction (Eaton, 2023; Moorhouse et al., 2023; Weber-Wulff et al., 2023). Yet those approaches are weakest where validity pressures are greatest, particularly in large-scale settings marked by low trust, uneven standards, and limited resources (Ross & Macleod, 2018). A surveillance response does not resolve the inferential problem; it simply relocates it. Even accurate detection cannot establish what assessment outcomes mean when AI use is permitted in part or in whole (Dawson et al., 2024; Messick, 1989; Weber-Wulff et al., 2023).

Assistance conditions can also alter the interpretive claim itself. In some cases, the claim is narrowed because competence is demonstrated under qualified conditions; in others, the construct being assessed changes because essential intellectual work has been delegated (AERA et al., 2014; Kane, 2013; Messick, 1989). AI-assisted and unassisted performances cannot be assumed to carry equivalent score meaning. That equivalence must be argued. GenAI should not be treated as a single undifferentiated category. Text generators, code assistants, multimodal systems, and domain-specific tools create different pressures on response processes, construct representation, and comparability (Dawson et al., 2024; Miao & Holmes, 2023). The relevant analytic focus is the relationship between tool affordances, the intended construct, and the conditions of performance.

This paper introduces the Validity Architecture for AI Integrated Assessment (VAAI), grounded in argument-based validity. VAAI is designed as a general validity architecture, developed at a level of abstraction that permits context-sensitive adaptation across educational settings. The discussion in this paper focuses on higher education assessment, where generative AI has intensified questions about the interpretation of student performance and the defensibility of assessment decisions, but the underlying logic is not restricted to that sector. The framework provides a structured basis for designing, testing, and revising assessments when AI mediation is present. Its central claim is that current assessment challenges should be understood less as problems of rule enforcement and more as problems of inference, evidence, and interpretive justification within routine assessment practice.

Three questions organize the framework: (1) what competence is being assessed, under what conditions, and for what decisions and consequences; (2) what evidence is required to justify the claim without becoming unduly intrusive; and (3) whether a student could satisfy the assessment standard through plausible AI-mediated strategies without demonstrating the intended competence (American Educational Research Association [AERA] et al., 2014; Kane, 2006, 2013; Messick, 1989; Ross & Macleod, 2018). In the validity literature, these correspond to the interpretive claim, the evidence model, and adversarial rebuttal testing, which together form the basis of VAAI.

For assessment practitioners, the same logic can be stated more directly: (1) name the learning claim, decide what evidence would justify it, (2) define what AI assistance remains compatible with the claim, align scoring and fairness with the evidence collected, and (3) identify when the task must be reviewed

or redesigned. The framework is technically derived, but its practical use is diagnostic. It helps instructors, program assessment committees, and assessment leaders ask whether a task still supports the conclusion and, in turn, the decisions they intend to form from student work.

Key Terms for Practitioners

Interpretive claim is the conclusion an instructor or program wants to draw from student work. For example, can a student evaluate evidence independently, apply professional judgement, or justify a design decision? The claim defines what the assessment is supposed to tell us.

Inference refers to the reasoning step that connects observed performance to that conclusion. In practical terms, it is the move from what the student submitted or did to what the assessor claims the student knows or can do.

Evidence model names the combination of product, process, justification, or supplementary evidence needed to support the claim. It could include a final artifact, a short rationale, an annotated outline, a revision log, a sampled oral defense, or another proportionate source of evidence.

Provenance refers to proportionate evidence about how the work was produced and what role the student played in producing it. It is not a demand for maximal surveillance; its purpose is to support the response process warrant required by the claim.

Assistance boundary identifies what AI may and may not do in relation to a given task, because some forms of support preserve the intended competence and others displace it. The boundary should follow from the assessment claim, not from a general institutional preference for permissiveness or restriction.

Stop rule is a pre-specified condition under which the assessment result should no longer be interpreted as intended without additional evidence, redesign, or reduced consequential weight. For example, if a student cannot explain the reasoning behind their submitted analysis, the result should not be treated as evidence of analytical competence.

Adversarial stress testing (also called adversarial rebuttal testing) is a design check in which instructors ask whether a plausible AI-supported strategy could meet the assessment criteria without the learner demonstrating the intended competence. The target is the assessment design, not the student.

Argument-Based Validity and the Inference Chain

An argument-based approach treats validity as the degree to which interpretations and uses of assessment results are warranted (Kane, 2006, 2013). The core idea is straightforward: every assessment involves a chain of reasoning. An instructor observes student performance, draws conclusions about competence, and makes decisions based on those conclusions. An interpretive argument maps that chain, and a validity argument tests whether each link holds up (Kane, 2006, 2013).

Under GenAI conditions, several links in that chain become fragile. Response processes shift when students use AI to generate or reshape their work. New sources of distortion emerge from unequal access, prompting skill, or platform quality. And fluent, polished outputs may mask the fact that a task did not force the student to demonstrate the intended competence (Messick, 1989, 1995). A validity lens gives educators a more precise language for identifying these problems without reducing them to academic integrity discourse (Eaton, 2023; Kaldaras et al., 2024; Swiecki et al., 2022).

Two concepts from the validity literature are particularly relevant here. Construct underrepresentation arises when a task omits essential features of the target competence. Construct-irrelevant variance arises when extraneous factors distort observed performance. In the first scenario, a task may narrow the skills being assessed; in the second, unwanted information or 'noise' is added, interfering with a teacher's ability to assess a student's true ability. In more severe cases, the observed product can no longer be credibly linked to the intended competence under the stated conditions, and interpretation becomes untenable. These are distinct problems requiring different responses (Messick, 1989, 1995).

Validity also encompasses consequences, not just measurement precision (Messick, 1989, 1995). When an assessment invites the outsourcing of core intellectual work or amplifies inequities through unequal access and false accusation, what looks like a policy issue is also a validity issue (Lane, 2014; Messick, 1989; Ross & Macleod, 2018). VAAI is designed to surface such failures before grades are assigned. That said, argument-based validity is not without criticism. Its dependence on professional judgement creates risks of circularity, and what counts as sufficient evidence is often contested (Kane, 2013). Those limitations are reasons to make warrants, rebuttals, and stopping conditions more explicit, particularly where AI complicates the evidence base. A more detailed treatment of argument-based validity and Messick's unified framework appears in Appendix A.

Alignment with Established Assessment Frameworks

VAAI builds on two foundational sources that practitioners may encounter in assessment work. Evidence-Centered Design (ECD) starts by naming the competence to be assessed, identifying the evidence that would indicate that competence, and designing tasks that can produce such evidence (Mislevy et al., 2003). VAAI adapts this logic to AI-mediated contexts by asking whether student work still supports the intended claim when GenAI assistance is available, permitted, unevenly accessed, or difficult to observe.

The *Standards for Educational and Psychological Testing* organizes validity evidence into five categories: test content, response processes, internal structure, relations to other variables, and consequences (AERA et al., 2014). VAAI maps these categories to GenAI-specific assessment concerns, including construct modelling, evidence of response processes, rater comparability, transfer evidence, washback, and equity-related consequences. A fuller discussion of how VAAI relates to ECD and the *Standards* appears in Appendix A.

Relationship to Existing Frameworks

VAAI builds on argument-based validity theory, the *Standards*, and established assessment traditions, including authentic assessment and assessment for learning (AERA et al., 2014; Biggs, 1996; Kane,

2006, 2013; Messick, 1989; Wiggins, 1998). These traditions have strengthened task authenticity, feedback processes, and program coherence. VAAI also draws on practitioner-facing assessment scholarship concerned with how assessment tasks shape learning, judgement, and program-level decision-making. Work on feedback literacy and evaluative judgement has emphasized that students must learn to appraise the quality of feedback, rather than merely receive corrections (Carless & Boud, 2018; Tai et al., 2018).

Research on assessment design in digital environments has argued that assessment must be organized around the kinds of evidence that can reasonably support learning claims under current conditions (Bearman et al., 2023; Dawson et al., 2024). Programmatic assessment adds a further consideration: consequential judgements are more defensible when they draw on multiple partial sources of evidence rather than on isolated performances treated as decisive (van der Vleuten et al., 2012). Perkins et al. (2024) have also proposed a scaled approach to integrating AI across assessment tasks, and Zhang and Tang (2025) provide recent evidence that AI-generated content tools are taken up differently across disciplinary learning contexts.

Under GenAI mediation, the remaining problem is not authenticity alone but whether the relationship between performance and competence has been specified well enough to sustain a valid inference (Kane, 2013). For example, an authentic policy brief may still fail as evidence of evaluative judgement if a student can use AI to select sources, synthesize positions, and generate recommendations without making the relevant decisions. VAAI responds with three extensions to existing frameworks. First, it makes the inference chain itself an object of design, requiring explicit articulation of how observed performance connects to competence claims under current conditions of AI availability (Kane, 2013; Messick, 1989). A professionally authentic task can still permit plausible AI-mediated performance that weakens interpretive validity.

Second, VAAI embeds adversarial stress testing as a routine design activity, operationalizing Kane's rebuttal logic. The target of the adversarial exercise is the assessment design, not the student; designers attempt to break the inference using available GenAI tools and develop a rebuttal catalogue grounded in actual tool affordances (Miao & Holmes, 2023). Third, it formalizes governance, making revalidation an expected responsibility rather than an improvised response to controversy (AERA et al., 2014; Kane, 2013).

A related distinction concerns assistance boundaries. VAAI derives them from interpretive claims rather than from policy preference or institutional posture (Kane, 2013; Messick, 1989). That logic allows one context to justify permissive arrangements and another to require tighter restrictions without inconsistency at the principle level (AERA et al., 2014). Where unaided individual performance is constitutive of the target construct, AI assistance collapses the construct rather than merely complicating its measurement. Once that AI limit is articulated within the interpretive argument, the boundary can be communicated as a feature of what is being assessed rather than as a presumption of misconduct (AERA et al., 2014; Kane, 2013; Messick, 1989; Ross & Macleod, 2018). Borderline cases remain common, and tool use and proxy use are better treated as points on a continuum of delegated intellectual work than as a binary.

VAAI operates at an abstraction level that provides shared validity logic for context-sensitive application without prescribing local operational details. A classroom teacher and a doctoral supervisor both need a defensible validity argument, but the documentation burden, institutional authority, and appropriateness of provenance requirements differ across educational contexts. What is claimed is a common architecture rather than demonstrated equivalence across sectors.

The VAAI Framework: Structure and Components

Framework Architecture

VAAI has two integrated components: (1) a framework for assessment design, evidence, feedback, and monitoring, and (2) a specification scale that indicates how far those components have been articulated, tested, and governed (AERA et al., 2014; Kane, 2013). This approach aligns with digital assessment scholarship that treats assessment design as an organizing problem: tasks, evidence, interpretation, and learning effects must be considered together rather than as separate administrative steps (Bearman et al., 2023). The specification scale provides a shared language for distinguishing what has merely been declared from what has been structured, defended, and audited. Two points are important before reading the scale. First, Level 4, Audited and Adaptive, is not the expected destination for every assessment. The appropriate level depends on the consequences of the interpretive decision: low-consequence classroom tasks may be well served at Level 1 or Level 2, while decisions about certification, progression, or public reporting typically warrant Level 3 or higher. Second, the scale should be read dimension by dimension rather than as a single overall score. A program may need a Level 3 specification for assistance boundaries while accepting Level 2 evidence documentation for a lower-weighted assignment. Table 1 presents the VAAI specification scale, where specification refers to the degree to which each assessment dimension has been articulated, evidenced, tested, and governed, rather than to developmental maturity or pedagogical sophistication.

Table 1

VAAI Specification Scale: Levels of Articulation, Evidence, Testing, and Governance Across Eight Assessment Dimensions

Dimension	Level 0: Unspecified	Level 1: Declared	Level 2: Structured	Level 3: Defensible	Level 4: Audited and Adaptive
A. Interpretive claim and intended use	No explicit claim or intended decision	Broad claim stated; consequences and uses remain vague	Claim bounded; decisions, consequences, universe of generalization specified	Full inference chain articulated; interpretation boundaries and stop rules explicit	Decision use audited; misuse triggers redesign or reduced consequential weight
B. Construct model and representation	Construct inferred from task only	Construct described but not partitioned or operationalized	Construct partitioned into content or reasoning, medium where relevant,	Construct stress tested; GenAI-sensitive skills classified; interactional	Construct representation monitored over time; drift checks include GenAI tool updates

Dimension	Level 0: Unspecified	Level 1: Declared	Level 2: Structured	Level 3: Defensible	Level 4: Audited and Adaptive
			epistemic responsibility; indicators mapped	competence included where relevant; rebuttal catalogue developed	
C. Evidence model and provenance sufficiency	Final product only	Some process artifacts requested, not inference aligned	Product, process, and justification evidence aligned to inferences; minimum provenance defined	Provenance tiers by consequence level; verification proportionate and non-punitive; privacy limits explicit	Provenance adequacy audited; data minimization and retention governed
D. Assistance boundaries and agency allocation	AI unregulated or banned without construct rationale	AI permissions stated without evidence rationale	Assistance boundaries derived from evidence needs; agency across actors specified	Non-delegable intellectual work defined; co-production paired with individual evidence; non-portability enforced	Assistance boundaries reviewed with context or tool changes; comparability rules enforced
E. Scoring, interpretation, and reporting	Impressionistic or misaligned scoring	Assessment criteria reward polish; interpretation limits absent	Assessment criteria aligned to construct where feasible; moderation planned; limits acknowledged	Rater training addresses GenAI effects; moderation implemented; interpretation limits enforced	Rater effects monitored; reporting prevents misuse; revalidation follows material change
F. Fairness, accessibility, and differential impact	No fairness analysis	Access noted without mitigation	Access and payoff analyzed; accommodations considered	Equity rebuttals tied to inferences; mitigation hierarchy applied; redesign or reduced consequential weight enacted	Subgroup and accommodation impacts monitored; inequity triggers enforced
G. Feedback, revision, and transfer	No feedback or misaligned feedback	Generic feedback; GenAI feedback unconstrained	Feedback purpose, timing, source specified; revision supported	AI feedback constrained to criteria; bias checks occur; transfer evidence supports extrapolation	Feedback and transfer outcomes monitored; washback informs iterative redesign
H. Monitoring, consequences, and governance	No monitoring or governance	Informal reflection only	Some monitoring; redesign ad hoc	Validity evidence linked to inferences; washback indicators tracked; redesign triggers specified	Governance assigns responsibility; revalidation thresholds exist; tool

Dimension	Level 0: Unspecified	Level 1: Declared	Level 2: Structured	Level 3: Defensible	Level 4: Audited and Adaptive
					change triggers review

Note. Dimension F's mitigation hierarchy refers to a sequenced approach to equity: first, redesign the assessment task to reduce differential burden; second, where the validity threat comes from AI assistance itself, tighten assistance boundaries or require more attributable evidence; third, provide targeted accommodations; fourth, reduce the weight of the assessment outcome if the preceding responses do not achieve proportionality. Each step requires active justification rather than default acceptance of inequitable conditions.

The specification scale is therefore a consequence-calibrated guide to assessment defensibility rather than an analytic rubric or a developmental maturity model. It makes the degree of specification, evidence, and governance visible. The eight dimensions were constructed deductively, drawing on Kane's (2006, 2013) argument-based validity framework, Messick's (1989, 1995) unified theory of construct validity, and the five validity evidence categories in the *Standards* (AERA et al., 2014), with additional input from ECD (Mislevy et al., 2003), assessment-for-learning scholarship (Carless & Boud, 2018; Tai et al., 2018), digital assessment design (Bearman et al., 2023), programmatic assessment (van der Vleuten et al., 2012), and emerging frameworks for AI-mediated assessment (Kaldaras et al., 2024; Perkins et al., 2024; Swiecki et al., 2022). Dimensions were included where they corresponded to a distinct inferential function in the validity argument, and overlapping functions were merged. The levels are ordinal descriptors of specification and defensibility rather than interval measures, and the dimensions are not assumed to carry equal weight across contexts. Further detail on the construction rationale appears in Appendix A.

Assessment Design Through the Inference Chain

VAAI posits an inference chain that begins with observed performance under specified conditions and proceeds through scoring, interpretation, and eventual use (Kane, 2013). Under GenAI mediation, the most vulnerable links are often response process warrants and extrapolation warrants (AERA et al., 2014; Kane, 2013). Even when intended learning outcomes are clearly stated, the evidence collected may not be sufficient to attribute performance to the student or to justify generalization beyond the immediate artifact (Kane, 2013; Messick, 1989).

A recurring failure is treating product quality as equivalent to competence. The artifact is a possible source of evidence; it is not the competence itself (Messick, 1989). In teacher education, for example, a student teacher may submit a polished unit plan with differentiated objectives, formative checkpoints, and scaffolded activities, yet, during practicum, fail to adapt when learners do not grasp a key concept. The pedagogical rationale was never demonstrated. Under AI-mediated conditions, the distance between artifact quality and attributable capability widens unless the division of labor is constrained or explicitly evidenced (Kane, 2013).

GenAI pressures should not be collapsed into a single validity category. Some cases involve construct underrepresentation because the task no longer elicits central features of the intended competence. Others involve construct-irrelevant variance because performance is shaped by differences in access,

prompting skill, platform quality, language coverage, or cultural alignment. In more severe cases, the score can no longer be interpreted as a credible representation of the intended attribute under the stated conditions (Messick, 1989, 1995). Tool heterogeneity sharpens the problem. A general-purpose text generator may threaten justificatory reasoning in ways that a code assistant proposing executable solutions does not, and multimodal or domain-specific tools alter task demands in other ways. The validity problem does not lie in AI as an abstract category, but in the interaction among tool capability, task design, and the competence being inferred. Stop rules make these vulnerabilities actionable: they are pre-specified conditions under which an assessment's interpretation is no longer defensible without redesign (Kane, 2013). Identifying them at the design stage, rather than after disputes arise, is one of VAAI's most practical features.

Evidence Models and Provenance Proportionality

Within VAAI, provenance is treated as evidence supporting inferences about response processes, not as a surveillance mechanism (AERA et al., 2014; Ross & Macleod, 2018). It refers to documentation sufficient to justify that observed performance is attributable as required by the interpretive claim (Kane, 2013), and its scope should remain proportionate to consequence level, accessibility requirements, and privacy limits (AERA et al., 2014; Ross & Macleod, 2018).

To reduce conceptual vagueness, VAAI distinguishes four forms of provenance:

1. Process provenance concerns traces of how the work was developed: planning notes, revision histories, or analytic logs.
2. Interactive provenance concerns evidence of the learner's engagement with prompts, feedback, or oral questioning tied to the warrant being tested.
3. Contextual provenance concerns the conditions under which the work was produced, including permitted resources, timing, collaboration rules, and assistance boundaries.
4. System provenance concerns the technical and procedural features of the tools involved, including version, access conditions, and affordances affecting comparability or interpretation.

These categories are analytically distinct, though they often overlap in practice. The central question is whether sufficient evidence has been gathered to make interpretations about student ability that are appropriate to the consequences and decisions being made (Messick, 1989). Some claims require little more than a bounded product and a declared assistance condition; others require sampled process evidence or supplementary defense. Examples are provided in Table 2.

This logic is consistent with programmatic assessment, where consequential judgements are strengthened by multiple partial evidence sources rather than by a single performance treated as decisive (van der Vleuten et al., 2012). In low- and medium-consequence contexts, modest supplementary evidence may suffice if it targets the most at-risk warrant. The issue is not the form of the artifact but whether it provides enough evidence to support the intended inference.

A further complication is that provenance artifacts themselves can be AI-generated, creating a second-order threat to the evidentiary chain. A student who fabricates a plausible revision history using GenAI, for example, can satisfy a process provenance requirement without having engaged in the documented

process. Revision histories, planning notes, reflective logs, and oral responses may all be fabricated or rehearsed with AI support (Dawson et al., 2024; Perkins et al., 2024). The implication is not that provenance requirements should be abandoned, but that the inferential value of any artifact type must be judged in relation to how easily it can be generated or manipulated under current tool conditions. Interactive provenance (oral defense, live justification, real-time walkthrough) tends to be more resistant than asynchronous written documentation, though it carries accessibility and feasibility costs of its own (Ross & Macleod, 2018; Swiecki et al., 2022).

Feasibility also must be treated as part of the evidentiary design. A sampled oral defense may be appropriate in some courses, but even a limited model creates real workload in large-enrolment settings. Short written justifications, rotating checkpoints, structured peer critique, or small-group walkthroughs may provide more proportionate alternatives where instructor time is constrained.

Table 2

Evidence and Provenance Options by Assessment Type Under GenAI Mediation

Assessment type	Primary claim risk under AI	Core evidence needed	Provenance options, proportional	Feasibility safeguards
Essay argument	Outsourced reasoning masked by fluent prose	Reasoning that links claims to evidence, source use, justificatory logic	Planning memo, annotated outline, decision trail for sources, brief oral defense on warrants	Sampled oral defenses; short, structured prompts; brief in-class written explanations
Literature synthesis	AI-generated synthesis without evaluative judgement	Inclusion rationale, comparison logic, epistemic stance	PRISMA-like selection notes, annotation set, contrast table authored by student	Pair work allowed but individual defense; accessibility alternatives
Data analysis report	AI-generated interpretation without method understanding	Method choice rationale, parameter choices, interpretation limits	Analysis log, code comments linked to decisions, oral walkthrough of key steps	Small-group defenses; checkpoint submissions
Programming task	Copied code with no debugging competence	Explainability of design, error handling, testing	Screen capture of debugging episode, commit history, test plan authored by student	Allow pseudocode alternatives; live check-in for a sample
Design artifact	Tool-generated output presented as student design	Rationale, iteration logic, constraints satisfaction	Versioned drafts, critique response log, rationale narrative tied to criteria	Focus on critique and redesign; studio critique formats
Language performance task	AI-generated fluency masking limited	Comprehension, pragmatic appropriateness, interactional repair, and	Recorded oral sample, brief live follow-up, annotated draft, self-explanation of revisions	Short, sampled interviews; alternative modalities for

Assessment type	Primary claim risk under AI	Core evidence needed	Provenance options, proportional	Feasibility safeguards
	communicative competence	explanation of language choices		accessibility; focus on targeted subskills

Assistance Boundaries and Agency Allocation

Assistance boundaries should not begin as policy declarations. Within VAAI, they follow from evidentiary requirements (Kane, 2013). Where the claim concerns individual competence in argumentation, the central question is which intellectual work cannot be delegated without undermining the inference (Messick, 1989). Where the claim concerns collaborative or professionally distributed work, co-production may form part of the construct, but agency still requires explicit specification and evaluation (Hutchins, 1995; Kane, 2013).

VAAI distinguishes between tool use and proxy use (Corbin et al., 2025; Miao & Holmes, 2023; Perkins et al., 2024). Consider an education student who uses GenAI to locate government reports and research on inclusive funding, then develops an independent argument about resource allocation. That student uses AI as a tool. A student who submits lightly edited GenAI output uses it as a proxy. Same task, same technology, but the evidentiary status differs because the intellectual work has shifted. The relevant question is not simply whether AI was used, but whether the assessment design can still support the claim that the student performed the relevant cognitive and analytical work (Kane, 2013). The distinction is heuristic rather than exhaustive; delegated work often falls along a continuum, with borderline cases involving partial drafting, iterative restructuring, suggested revisions, or selective support for reasoning.

Scoring, Interpretation Limits, and Reporting

Within VAAI, interpretation limits are built into the scoring design from the outset (Kane, 2013). Rubrics must align with the construct model, with explicit attention to which skills are sensitive to AI mediation. Surface fluency, unless it forms part of the construct, should not carry disproportionate weight. Where source evaluation, methodological judgement, or justificatory reasoning is central, those elements should be scored directly rather than inferred from polished output alone (Kane, 2013; Messick, 1989). A well-structured essay that reads fluently but lacks independent analytical reasoning should score lower than a rougher piece that demonstrates genuine critical engagement with the material.

Questions of reliability also arise here. Under AI-mediated conditions, consistency cannot be reduced to conventional psychometric precision. It also involves whether raters interpret AI-shaped performances comparably, whether moderation addresses tool-related distortions, and whether reporting preserves the limits of warranted interpretation. Reliability, in this sense, is partly technical and partly inferential.

Feedback, Revision, and the Inference Chain

Feedback occupies a structurally ambiguous position in AI-mediated assessment environments. Within VAAI, Dimension G (see Table 1) addresses feedback, revision, and transfer as sites where the

inference chain can be quietly rerouted. When feedback is designed to build learner capacity rather than to transmit corrective information, it can support evaluative judgement and contribute to transfer (Boud & Molloy, 2013; Carless & Boud, 2018; Tai et al., 2018). When AI systems mediate the revision cycle in ways that substitute for student reasoning rather than support it, the revised submission may appear stronger without the learner having done the intellectual work the assessment was intended to evidence.

Consider an assignment in which the intended learning outcome is for students to diagnose weaknesses in their own arguments. Identifying an unsupported claim, deciding what evidence would strengthen it, and revising the warrant are all part of the assessed intellectual work. If an AI system supplies the diagnosis, the evidence strategy, and the revised wording, the improved draft no longer provides clear evidence that the student has developed evaluative judgement.

The inferential problem here is specific. A student who receives detailed AI-generated feedback (criterion-by-criterion evaluations, suggested restructuring, rephrased arguments) and incorporates it into a revised draft has not necessarily demonstrated the evaluative judgement the revision cycle was supposed to develop (Carless & Boud, 2018; Messick, 1989; Tai et al., 2018). Feedback design scholarship makes the same point from a learning perspective: students need opportunities to generate, seek, interpret, and act on feedback, rather than merely receive corrections formulated by an external source (Boud & Molloy, 2013; Nicol, 2021). Where evaluative judgement is a target competence, AI-mediated feedback should be bounded in ways derived from the interpretive claim rather than from general policy (Kane, 2013). Permitting AI to flag surface-level errors while requiring the student to diagnose conceptual weaknesses and generate revisions independently may be defensible. Permitting full AI revision is not if the claim concerns the student's own capacity to evaluate and improve their work (Carless & Boud, 2018; Nicol, 2021).

Revision cycles also introduce additional complexity. A second submission raises a distinct question of attributability: what intellectual work did the learner perform between versions? Process provenance becomes especially important here. Revision logs, decision trails, or brief written justifications for changes can supply evidence that movement between drafts reflects the student's own analytical engagement (Nicol, 2021; Tai et al., 2018). Where such documentation is not proportionate to the level of consequence, a brief, structured reflection may serve as a viable alternative.

Implementation: A Practical Guide for Educators How to Use This Framework

VAAI is designed for assessment work and should be revisited whenever tasks, permitted uses of GenAI, consequences, or tool capabilities change (AERA et al., 2014; Kane, 2013; Miao & Holmes, 2023). It is not a one-time checklist. It is a design process that stays active as conditions shift. Table 3 presents the practical sequence for using VAAI: specify the claim, design the evidence and performance conditions, align interpretation and fairness, and establish governance for review. Table 4 then shows the extent of documentation and monitoring warranted at different levels of consequence.

To make the implementation logic explicit, VAAI can be read through four sequential stages that operationalize the eight dimensions in Table 1. The stages describe the order in which key validity

decisions are made; the dimensions identify the contexts in which those decisions must be articulated, tested, and governed. This sequence is intended to help practitioners use the framework without first mastering the full technical architecture.

Once the implementation sequence has been established, the next question is how much specification, documentation, and monitoring are warranted. Table 4 provides a consequence-calibrated guide for making that judgement across different decision contexts.

Table 3

VAAI Implementation Stages, Embedded Dimensions, and Stop Rule Checks

Stage	Embedded procedures linked to the eight dimensions	Stop rule check
Stage 1. Interpretive claim and construct specification	**A. Interpretive claim and intended use:** define the competence claim, decision, consequence level, universe of generalization, interpretation limits, and stop rules. **B. Construct model and representation:** specify the target construct, its components, intellectual work, relevant indicators, and AI-sensitive features.	**Stop Rule Check 1:** Is the claim sufficiently specified and the construct adequately represented? If no, return to **A. Interpretive claim and intended use**.
Stage 2. Evidence and performance condition design	**C. Evidence model and provenance sufficiency:** determine what product, process, and justification evidence is required; define minimum sufficient provenance; plan triangulation proportionate to consequence level. **D. Assistance boundaries and agency allocation:** specify permissible and impermissible AI use, locate non-delegable intellectual work, distinguish tool use from proxy use, and determine when stricter controls are warranted. Include adversarial rebuttal testing at this stage.	**Stop Rule Check 2:** Can the intended inference still be supported under the designed evidence and performance conditions? If no, return to **A. Interpretive claim and intended use**.
Stage 3. Interpretation, fairness, and feedback alignment	**E. Scoring, interpretation, and reporting:** align rubrics to the construct, address moderation and comparability, and specify reporting limits. **F. Fairness, accessibility, and differential impact:** review access, burden, language and cultural effects, accommodations, and mitigation hierarchy. **G. Feedback, revision, and transfer:** define the purpose, timing, and source of feedback; constrain AI-mediated feedback where necessary; consider revision and transfer conditions.	**Stop Rule Check 3:** Are scoring, fairness, and feedback conditions aligned with the claim and construct? If no, return to **A. Interpretive claim and intended use**.
Stage 4. Governance and revalidation	**H. Monitoring, consequences, and governance:** establish review triggers, monitor washback and subgroup effects, assign responsibility, define revalidation thresholds, and respond to tool, uptake, or administrative change.	**Stop Rule Check 4:** Does ongoing monitoring support continued interpretive adequacy? If no, return to **A. Interpretive claim and intended use**.

Note. The implementation sequence is ordered for design purposes but is not strictly linear. Later stages may require revision of earlier ones, particularly when adversarial testing, fairness analysis, or governance review indicates that the original claim, evidence model, or construct specification is no longer defensible.

Table 4

Consequence-Calibrated Specification Guide for VAAI Implementation

Consequence level	Decision examples	Minimum VAAI level	Required dimensions at Level 3 or higher	Documentation burden	Review frequency
Low consequence	Feedback only; no grade impact; self-assessment; draft work	Level 1 across dimensions	None required, but interpretive claim and construct model recommended	Minimal	When task changes significantly
Medium consequence	Unit grades; course outcomes; portfolio components; progression decisions	Level 2 for design dimensions; Level 1 for others	Interpretive claim; construct model; assistance boundaries	Moderate	Annually or with major tool updates
High consequence	Course pass or fail; program completion; certification; placement decisions; public reporting	Level 3 across dimensions	All dimensions with documented validity argument	Substantial	Each term or when any parameter changes
Very high consequence	Professional licensure; graduation requirements; system accountability; highly consequential admissions decisions	Level 3 to 4 across dimensions	Level 3 minimum for all, with audited governance for core dimensions	Comprehensive	Continuous monitoring with formal review cycles

Note. Specification, as used in this table, refers to the depth of articulation, evidence, testing, and documentation required for each validity dimension, calibrated to the potential consequences of misinterpreting student ability. Higher specification does not imply greater control over students; it requires greater explicitness in the interpretive argument.

Worked Examples Across Practitioner Contexts

The examples below translate VAAI into familiar higher-education assessment contexts. They are not templates to be copied wholesale. Each shows how the same validity logic can be adjusted when discipline, the level of consequence, and the expected form of judgement differ. Recent research also suggests that AI-generated content tools are not taken up uniformly across higher education and may affect learning outcomes differently across fields (Zhang & Tang, 2025).

Nursing: Care Plan Assignment

A care plan assignment might require students to use patient data, clinical priorities, and relevant evidence to justify an appropriate plan of care. A plausible AI-related risk is that a GenAI system produces a polished care plan while the student misses contraindications, prioritization errors, or patient-specific risks. Depending on the consequence, the evidence model may need more than the submitted plan alone: a brief rationale for the top priorities, a response to a change in the patient's condition, or a short oral walkthrough of one intervention may be enough. AI could be allowed for general background review, but not for generating the final prioritization or clinical rationale. A defensible stop rule would be triggered if the student cannot explain why the top two priorities shift when the patient's condition changes.

Counselling: Case Conceptualization

A case conceptualization assignment may support the claim that students can select and justify an intervention that is consistent with the client's context, ethical obligations, and counselling theory. GenAI can generate fluent theoretical language while missing relational nuance, cultural context, risk indicators, or ethical constraints. The evidence should include the written conceptualization, a brief justification of excluded alternatives, and a response to an ethical complication. AI use may be acceptable for reviewing general theory definitions, but the client-specific formulation and ethical reasoning remain non-delegable. If a student cannot explain how the client's context altered the choice of intervention, the submission should not count as sufficient evidence of counselling judgement.

Program Level: Capstone Review

A graduate capstone review poses a different problem. Consider a Master of Education program that uses final inquiry projects as evidence that graduates can integrate literature, contextual analysis, and professional judgement into a defensible educational intervention. AI-mediated production may weaken claims about synthesis, transfer, and independent judgement across the program. A proportionate VAAI response might require departments to sample capstones, check whether final scores reflect intended outcomes rather than surface polish, add a short student-authored rationale for major design decisions, and specify when results should be interpreted with caution. The stop rule may also operate at the program level: if multiple projects meet assessment criteria but provide little evidence of independent synthesis, capstone results should not be used to support program claims without redesign.

Stage 1: Interpretive Claim and Construct Specification

The first stage establishes what the assessment is meant to warrant. At minimum, the interpretive argument should identify the competence being inferred, the decision being made, the consequences attached to that decision, and the universe of generalization (Kane, 2013). Without that specification, GenAI rules become arbitrary rather than claim-based (Kane, 2013; Miao & Holmes, 2023).

A stop rule at this stage is the pre-specified condition under which the assessment should not be interpreted as originally intended without redesign, additional evidence, or reduced consequential weight (Kane, 2013). Consider an education faculty assignment requiring a 2,000-word reflective essay on inclusive classroom practices. If the claim is that students can critically evaluate their own teaching

decisions through a theoretical lens, the rebuttal condition is straightforward: if unedited GenAI output can meet the rubric threshold, the assessment has failed its validity test, and redesign is the appropriate response rather than intensified penalty structures (AERA et al., 2014; Messick, 1989). For an instructor, the stop rule functions as an advance decision about when evidence has become too weak to support the intended claim.

Stage 2: Evidence and Performance Condition Design

Once the claim and construct have been specified, the next task is designing the evidence model and the conditions of performance needed to support the inference. VAAI treats provenance as evidence supporting response process claims, not as a surveillance mechanism (AERA et al., 2014; Kane, 2013; Ross & Macleod, 2018). The issue is not how much documentation can be collected but what minimum combination of product, process, and justificatory evidence is sufficient for the claim at hand.

When product evidence is ambiguous under GenAI conditions, supplementary evidence may be needed. Suitable options include checkpoint submissions, annotated decision trails, one-minute papers, outlines, brief in-class written explanations, short justificatory prompts, sampled oral defenses, or small-group walkthroughs, depending on consequence level and local feasibility (Carless & Boud, 2018; Kaldaras et al., 2024; Kane, 2013; Swiecki et al., 2022; Tai et al., 2018; van der Vleuten et al., 2012). These are sampling strategies, not universal interrogation. Accessibility must be built in from the start, since methods that impose unnecessary differential burden weaken the validity argument they are supposed to support (Lane, 2014; Messick, 1989).

AI assistance boundaries are also designed at this stage and derived from evidentiary requirements rather than policy declarations (Kane, 2013). The central issue is which forms of assistance remain compatible with the claim and which displace essential intellectual work. The distinction between tool use and proxy use is useful here, though it is better treated as a heuristic than an absolute binary (Miao & Holmes, 2023; Perkins et al., 2024), particularly when GenAI contributes to partial drafting, restructuring, or revision rather than replacing the task wholesale.

Adversarial rebuttal testing functions as the stress test for this stage: designers ask whether plausible AI-supported strategies could satisfy the assessment criteria threshold without the learner demonstrating the intended competence (Kane, 2013; Messick, 1989). If so, the warrant is too weak. If a general-purpose text generator can produce a research methods assignment that, for instance, earns a B+ after light revision, the task may not be eliciting the competence that the assessment criteria presume. Adding a methods justification component may restore the warrant by requiring an attributable judgement rather than polished output alone.

In practice, this resembles an AI-mediated assessment-design review: the task is examined against the evidence it can realistically produce under current technological conditions (Bearman et al., 2023; Dawson et al., 2024). The protocol should remain proportionate to the level and scale of the consequences, with more systematic testing reserved for higher-consequence contexts and documented in ways appropriate to local conditions (AERA et al., 2014; Miao & Holmes, 2023).

Where the claim requires tighter control, stricter conditions may be warranted. A classroom management simulation in teacher education may require real-time judgement in response to unfolding student behavior. A foundational reading assessment may require the student to identify a miscue, explain the likely source of the error, and choose an appropriate instructional response. In such cases, GenAI involvement would collapse the construct if it generated the diagnosis or selected the instructional response, because the assessment would no longer require the student to perform the interpretive judgement being assessed (AERA et al., 2014; Kane, 2013). The relevant question is: what is the least restrictive condition that can still support the inference?

Stage 3: Interpretation, Fairness, and Feedback Alignment

After the evidence model and performance conditions have been established, the design must align scoring, fairness, and feedback with the intended claim. Assessment criteria, including rubrics where used, should align with the construct model rather than reward surface fluency by default (Kane, 2013; Messick, 1989). Where methodological judgement, source evaluation, or justificatory reasoning is central, those elements should be scored directly rather than inferred from polished output alone. When cohorts differ in tool access, assistance boundaries, or provenance requirements, score comparability may be restricted, and those limits should be stated rather than left implicit (AERA et al., 2014; Kane, 2013).

Differential access to paid tools is only one dimension of the fairness problem. Language coverage, cultural knowledge alignment, disability-related burden, platform familiarity, and local resource constraints can all produce construct-irrelevant variance (Lane, 2014; Messick, 1989; Miao & Holmes, 2023). VAAI addresses this through a mitigation hierarchy: redesign the task first to reduce differential burden; where the validity threat comes from AI assistance itself, tighten assistance boundaries or require more attributable evidence; provide targeted accommodations when necessary; and reduce the consequential weight of the outcome when none of those responses is sufficient. Where outcome weight cannot be reduced, as with outcomes tied to certification, stricter assistance boundaries or a redesigned evidence model become the operative remedies. That sequence is principled, though its application may be constrained in rigid or under-resourced settings.

Feedback and revision require specification too. In AI-mediated environments, feedback can support learning while also altering the evidentiary status of later performance. The issue is not whether feedback occurs but what kind, from whom, at what point, and toward what purpose. AI-mediated feedback should be constrained where it would substitute for evaluative judgement rather than support it, and revision opportunities should remain aligned to the underlying construct (Carless & Boud, 2018; Tai et al., 2018).

Stage 4: Governance and Revalidation

The final stage addresses how validity arguments are maintained over time. Governance is not itself a form of validity evidence; it is the structural condition under which validity arguments can remain adequate as tools, uptake patterns, and assessment environments change (Kane, 2013; Messick, 1989). Without that maintenance function, even well-designed assessments can lose interpretive credibility.

Revalidation should not be triggered only by an obvious tool change. Shifts in model capabilities, student use patterns, and local administration conditions can all affect the evidentiary basis for score interpretation. Review triggers should be specified in advance. Washback, subgroup effects, accommodation outcomes, and patterns of misinterpretation or misuse should be monitored as part of routine governance (AERA et al., 2014; Lane, 2014). Responsibility must be assigned explicitly. Validity work cannot rest solely with individual instructors or be delegated entirely to central administrative units. Shared ownership is necessary, particularly where provenance requirements or revalidation thresholds have consequences beyond a single course.

These four stages, presented sequentially, provide a usable implementation pathway. Iteration remains expected. Fairness analysis, adversarial testing, or governance review may still require a return to earlier stages when the original claim or construct specification can no longer be sustained.

Discussion

VAAI responds to a weakness in assessment practice that GenAI has made operationally difficult to ignore: higher education has often treated artifacts as workable proxies for cognition while leaving the warrants linking product to competence only partly articulated (Kane, 2013; Messick, 1989). Recent research suggests that this reliance now carries greater risk, as students increasingly offload cognitive work to AI in ways that may reduce the depth of processing and the metacognitive effort on which response-process warrants depend (Fan et al., 2025; Gerlich, 2025). Under AI-mediated conditions, implicit warrants require explicit formulation and ongoing scrutiny (AERA et al., 2014; Messick, 1995).

VAAI relocates interpretive argumentation from the margins of psychometric discourse to routine design work (Kane, 2013). In other words, assessment theory becomes operational. Equity functions as a validity constraint rather than a supplementary concern. Governance addresses not only policy but the maintenance of defensible score meaning as conditions shift (Lane, 2014; Ross & Macleod, 2018). The framework should not be treated as theoretically neutral or implementation-proof. Validity frameworks can be bureaucratized, reduced to procedural checklists, or applied in ways that intensify harm for marginalized students. Any account of VAAI that assumes institutional goodwill, stable expertise, or benign uptake would be too thin.

Reframing Assessment Problems as Inference Problems

VAAI extends contemporary assessment scholarship that prioritizes validity and assessment design over detection, given that student work can be shaped by AI assistance (Bearman et al., 2023; Dawson et al., 2024; Perkins & Roe, 2025). It asks whether the evidence collected under stated conditions can support the interpretations being made (AERA et al., 2014; Kane, 2013). This reframing aligns AI-mediated assessment challenges with long-standing debates in educational measurement concerning construct representation, construct-irrelevant variance, and the limits of proxy evidence (Kane, 2013; Messick, 1989). Consider a student whose essay is fluent, well-cited, and polished, yet who cannot explain their own analytical decisions in a follow-up conversation. The issue is evidentiary insufficiency, not simply misconduct; design attention shifts away from detection technologies and toward credible evidence of the intended competence under the permitted conditions of assistance (Dawson et al.,

2024; Weber-Wulff et al., 2023). Kalantzis and Cope (2025) suggest that AI is reshaping literacy and competence across disciplines. If so, redesign is an ongoing obligation, not a temporary adjustment.

Equity, Surveillance, and Construct-Irrelevant Variance

Within VAAI, equity is a validity requirement because differential payoffs across student groups constitute validity-relevant evidence (Lane, 2014; Messick, 1989). If scores systematically advantage students with greater linguistic capital, greater prior familiarity with GenAI tools, or access to paid platforms, the assessment may capture a construct the designer never intended (Messick, 1989; Miao & Holmes, 2023). These are not incidental effects. They are forms of construct-irrelevant variance that weaken the interpretive warrants on which the assessment depends (Lane, 2014; Messick, 1989). The link to differential item functioning is direct: when subgroup differences arise from the assessment design rather than from the target construct, this constitutes a validity failure that requires remediation at the design stage (Camilli & Shepard, 1994). The equity problem extends beyond paid versus free access. Differential system performance across languages, dialects, and culturally specific knowledge domains may distort the quality of responses and the value of assistance, creating particular risks for Indigenous, multilingual, and diaspora students. Equal access in formal terms may conceal unequal functional support.

Surveillance-based responses often intensify these problems (Ross & Macleod, 2018). Proctoring technologies and heavy demands for provenance can impose disproportionate burdens on students with disabilities, caregiving responsibilities, or precarious living conditions, and may distort response processes by rewarding concealment rather than meaningful engagement (Messick, 1989; Ross & Macleod, 2018). VAAI addresses this through the mitigation hierarchy embedded in Dimension F: redesign the task first to reduce differential burden; tighten assistance boundaries or require more attributable evidence where AI assistance is the source of the threat; provide targeted accommodations where necessary; reduce the consequential weight of the outcome where none of those responses is sufficient. This is an operational constraint, not a rhetorical commitment to fairness. It does assume some degree of institutional flexibility. Instructors working under rigid curricula, contingent employment, severe resource pressures, or highly standardized systems may lack the authority to redesign tasks, tighten boundaries, or reduce the weight of assessments, thereby limiting the framework's practical reach even where its logic remains sound.

Feasibility, Scale, and Proportionality

A predictable objection is that triangulated evidence and adversarial stress testing may be difficult to implement at scale. VAAI treats feasibility as part of the validity design rather than as an afterthought. Recent work on AI and assessment characterizes the problem as "wicked" precisely because educators face competing demands that cannot be resolved through a single stable solution: validity, scalability, equity, workload, and institutional policy constraints all pull in different directions (Corbin et al., 2026). A design that requires an extensive oral defense in a 600-student course may be theoretically attractive but practically indefensible if it produces delay, inequitable burden, or inconsistent administration. Proportionality is a validity condition, not merely an implementation preference.

Feasibility objections also tend to conflate increased workload with unfamiliarity (Kaldaras et al., 2024; Swiecki et al., 2022). Many practices recommended by VAAI, including short justificatory prompts, staged submissions, and brief oral explanations, are already used informally by instructors (Bearman et al., 2023). Programmatic assessment makes a related point: evidence need not be exhaustive to be useful, provided each source is interpreted for its limited purpose and combined cautiously as decisions become more consequential (van der Vleuten et al., 2012). Some recommendations remain genuinely costly, so selective implementation, shared departmental effort, or reduced procedural ambition may be necessary.

Governance and Risk of Performative Validity

Validity frameworks can be reduced to checklists that produce the appearance of rigor without its substance (Kane, 2013). VAAI is susceptible to that risk if institutions adopt its language without the assessment literacy needed for substantive implementation. Governance must therefore include the explicit assignment of responsibility for interpretive arguments and revalidation decisions, with shared ownership among instructors and administrative units, particularly where provenance requirements or revalidation thresholds have consequences beyond a single course.

Governance can also become a site of capture. Provenance requirements justified as inference support may be repurposed for surveillance, disciplinary escalation, or standardization pressures that disproportionately harm already marginalized students. Critical assessment scholarship has long warned that seemingly neutral assessment structures can reproduce inequity unless their normative assumptions are examined (Shepard, 2000). VAAI should therefore be understood not as a theoretically neutral framework but as one requiring explicit safeguards against institutional misuse, including data minimization, purpose limitation, defined retention periods, and restrictions on secondary use (Shavit et al., 2023).

An Empirical Research Agenda

VAAI is a structured conceptual proposal rather than an empirically validated model. Several research priorities follow. Because the scale was developed deductively rather than empirically, its content validity remains provisional. Research should examine whether practitioners treat the dimensions as conceptually distinct, whether some are regularly collapsed in use, and whether dimensions such as interpretive claim and construct model are separable in practice or better understood as overlapping parts of a broader architecture.

Adversarial rebuttal testing also requires direct study. VAAI treats it as a routine design activity, but its capacity to identify validity threats has not been established systematically. Comparative studies could examine its diagnostic reach relative to other forms of validity review and whether it helps distinguish construct underrepresentation, construct-irrelevant variance, and cases in which interpretation becomes untenable under stated conditions (Messick, 1989, 1995).

Evidence model redesign is a further priority. VAAI assumes that deriving assistance boundaries from interpretive claims, rather than policy preference, produces stronger alignment between evidence design and competence inference. Redesign studies comparing artifacts before and after VAAI

application, or contrasting VAAI-guided designs with those produced through alternative frameworks, could examine whether the framework improves the precision of interpretive claims and the fit between evidence models and intended constructs (AERA et al., 2014; Kane, 2013).

Usability requires the same level of scrutiny. Frameworks that exceed educators' cognitive or temporal resources fail through inaccessibility rather than conceptual weakness (Kaldaras et al., 2024). Studies in which educators apply VAAI to their own assessments, develop simplified practitioner versions, and report on usability and cognitive load would help identify where further scaffolding is needed across educational settings, including those where formal assessment literacy may be uneven. Student experience should also be examined: how learners interpret VAAI-governed environments, whether assistance boundaries are seen as intelligible and fair, and how such conditions shape trust and disclosure of tool use. Table 5 presents three illustrative research designs aligned with these priorities.

Table 5

Illustrative Research Designs for Empirical Evaluation of VAAI

Design	Description	Research question	Key outcomes	Contribution
Classroom redesign studies	Participants redesign the same assessment task twice: once using Kane's argument-based validity model and once using VAAI. Resulting designs are evaluated against predefined criteria for interpretive claim clarity, evidence model alignment, and AI boundary specification.	Tests whether VAAI produces stronger alignment between interpretive claims, evidence models, and AI assistance boundaries than argument-based validity applied without VAAI structure.	Assessment design artifacts; rubric alignment scores; clarity of interpretive claims; specificity of AI boundaries.	Evaluates whether VAAI improves practical assessment design under AI-mediated conditions.
Diagnostic validity studies	Participants analyze intentionally flawed assessments containing AI-related validity threats. They diagnose problems using either Kane's framework or VAAI.	Compares whether VAAI enables more accurate and differentiated identification of validity threats related to AI mediation.	Number and type of validity threats identified; depth of diagnostic reasoning; classification accuracy across construct underrepresentation, construct-irrelevant variance, and score non-interpretability.	Tests whether VAAI improves educators' ability to detect and categorize AI-induced validity problems.
Practitioner usability studies	Teachers apply VAAI to redesign their own classroom assessments and produce simplified	Evaluates usability and cognitive load when applying the framework in live	Practitioner adaptations; time-to-completion; perceived usability; framework simplifications and reductions.	Generates practitioner-facing versions of VAAI and tests

Design	Description	Research question	Key outcomes	Contribution
	practitioner versions of the framework suited to their teaching context.	teaching conditions without specialist measurement training.		feasibility under realistic classroom conditions.

Note. These study suggestions are illustrative rather than exhaustive. Complementary approaches might include longitudinal tracking of assessment redesign across faculty development programs, think-aloud protocols that examine how educators reason through the specification scale, design-based research on local adaptation, grounded theory studies of practitioner sense-making, and comparative institutional case studies that examine revalidation governance under differing policy conditions.

Conclusion

GenAI has forced educators across sectors to confront a long-standing fragility in assessment: the inferences underlying evaluative decisions have always depended on assumptions about how student work was produced (Kane, 2013; Messick, 1989), leading initial responses to AI use to focus on detection and to be grounded in conversations about academic integrity. For many years, the contingency in interpreting student ability from a product was obscured. GenAI has stripped it away (Eaton, 2023; Miao & Holmes, 2023), and the legitimacy of assessment now depends on reconstructing the interpretive logic that connects evidence to claims (Weber-Wulff et al., 2023). VAAI recenters assessment on inference accountability, with particular relevance for higher education under current conditions of AI mediation. Its practical contribution is to convert broad validity principles and applied assessment-design traditions into a usable sequence of design questions: what claim is being made, what evidence supports it, what assistance remains compatible with it, how scoring and fairness should be handled, and when revalidation is required. The framework is not intended to replace practitioner judgement with a technical protocol. It is meant to strengthen the evidentiary reasoning already present in feedback design, digital assessment, and program-level review. VAAI is less a fixed instrument than a disciplined way of asking whether assessment results still mean what educators intend.

Assessment validity can no longer remain an implicit assumption. In AI-mediated environments, it becomes an ongoing design responsibility for instructors, program leaders, and assessment practitioners. VAAI provides a structured framework for redesigning assessment as generative AI becomes part of ordinary teaching, coursework, and institutional assessment practice, especially in higher education.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bearman, M., Nieminen, J. H., & Ajjawi, R. (2023). Designing assessment in a digital world: An organising framework. *Assessment & Evaluation in Higher Education, 48*(3), 291–304. https://doi.org/10.1080/02602938.2022.2069674

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/BF00138871

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2026). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, *51*(4), 738–752. https://doi.org/10.1080/02602938.2025.2553340

Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025). 'Where's the line? It's an absurd line': Towards a framework for acceptable uses of AI in assessment. *Assessment & Evaluation in Higher Education, 50*(5), 705–717. https://doi.org/10.1080/02602938.2025.2456207

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

Eaton, S. E. (2023). Post-plagiarism: Transdisciplinary ethics and the future of academic integrity in the age of artificial intelligence. *International Journal for Educational Integrity, 19*, Article 23. https://doi.org/10.1007/s40979-023-00144-1

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Gouseti, A., James, F., Fallin, L., & Burden, K. (2025). The ethics of using AI in K-12 education: A systematic literature review. *Technology, Pedagogy and Education, 34*(2), 161–182. https://doi.org/10.1080/1475939X.2024.2428601

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Kalantzis, M., & Cope, B. (2025). Literacy in the time of artificial intelligence. *Reading Research Quarterly, 60*, e591. https://doi.org/10.1002/rrq.591

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education, 9*, Article 1399377. https://doi.org/10.3389/feduc.2024.1399377

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education, 49*(5), 651–664. https://doi.org/10.1080/02602938.2024.2309963

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. Educational Testing Service.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers & Education Open, 5*, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Perkins, M., & Roe, J. (2025). The end of assessment as we know it: GenAI, inequality and the future of knowing. In *AI and the future of education: Disruptions, dilemmas and directions* (pp. 76–80). UNESCO. https://doi.org/10.54675/KECK1261

Perkins, M., Roe, J., & Furze, L. (2024). *The AI Assessment Scale revisited: A framework for educational assessment*. arXiv. https://arxiv.org/abs/2412.09029

Ross, J., & Macleod, H. (2018). Surveillance, (dis)trust and teaching with plagiarism detection technology. *Proceedings of the International Conference on Networked Learning, 11,* 235–242. https://doi.org/10.54337/nlc.v11.8760

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023, December 14). *Practices for governing agentic AI systems*. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers & Education: Artificial Intelligence, 4*, Article 100127. https://doi.org/10.1016/j.caeai.2023.100127

St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education, 22*(4), 853–867.

https://doi.org/10.1007/s10459-016-9716-3

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers & Education: Artificial Intelligence, 3*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher, 34*(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, Article 26. https://doi.org/10.1007/s40979-023-00146-z

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Zhang, Y., & Tang, Q. (2025). Integrating AI-generated content tools in higher education: A comparative analysis of interdisciplinary learning outcomes. *Scientific Reports, 15*, Article 25802. https://doi.org/10.1038/s41598-025-10941-y

About the Authors

Johanathan Woodworth, Assistant Professor, Faculty of Education, Mount Saint Vincent University, johan.woodworth@msvu.ca

Med Kharbach, Instructor, Education, Mount Saint Vincent University, Med.kharbach@msvu.ca

Christine Doe, Associate Professor, Faculty of Education, Mount Saint Vincent University, christine.doe@msvu.ca

Appendix A: Theoretical Foundations

This appendix provides a fuller account of the validity scholarship on which VAAI draws. The main text presents these ideas in condensed form; the discussion here supports readers seeking a more detailed theoretical grounding.

Argument-Based Validity

Kane's (2006, 2013) argument-based approach treats validation as an evaluation of the interpretive argument connecting observed performance to claims and decisions. The interpretive argument specifies the inference chain: scoring, generalization, extrapolation, and decision or use. The validity argument evaluates whether the warrants and assumptions supporting each link are adequate and whether plausible rebuttals have been addressed.

Under GenAI conditions, the most vulnerable links are typically at scoring and generalization, where evidence of the response process becomes uncertain, and at extrapolation, where AI-mediated production may weaken the relationship between task performance and the target competence. VAAI treats these as design problems to be addressed through specification, evidence modelling, and adversarial testing.

Kane (2013) emphasizes that validation is ongoing and context-sensitive. The adequacy of a validity argument can deteriorate as conditions change, which is why VAAI builds revalidation triggers into its governance dimension. A recurring criticism is that argument-based validity depends on professional judgement, creating risks of circularity (Kane, 2013). VAAI addresses this partly through adversarial stress testing, which introduces structured self-challenge into the design process.

Messick's Unified Framework

Messick (1989, 1995) treats validity as a single integrated concept in which all evidence bears on construct validity. Two threats are especially relevant to AI-mediated assessment. Construct underrepresentation occurs when a task fails to include important features of the target construct. Construct-irrelevant variance occurs when extraneous factors distort scores. Under GenAI conditions, construct underrepresentation may arise when AI can complete a task without engaging the intended reasoning; construct-irrelevant variance may arise from differential access, prompting skill, or system performance across languages and cultural contexts.

Messick insists that validity encompasses consequences as well as evidence: the social effects of the use of assessment on teaching, learning, and equity. This is why VAAI treats equity as a validity constraint. When an assessment design produces systematically different outcomes for different student groups as a result of the design itself, the validity argument is weakened.

The *Standards*

The *Standards* (AERA et al., 2014) organize validity evidence into five categories: test content, response processes, internal structure, relations to other variables, and consequences of testing. Under GenAI conditions, evidence of the response process requires particular attention because AI can alter or replace the cognitive processes that assessments are designed to elicit. Consequential evidence must

address washback effects, equity implications, and the risk that surveillance measures produce their own forms of construct-irrelevant variance. *The Standards* also establish that the burden of evidence should be proportionate to the consequences of the decision, a principle VAAI operationalizes through the specification scale and the consequence-calibrated specification guide in Table 4.

Evidence-Centered Design

Evidence-Centered Design (Mislevy et al., 2003) provides a systematic framework organized around three coordinated models: a student model (which competences are assessed), an evidence model (which observable features count as evidence), and a task model (which tasks can elicit that evidence). ECD makes the reasoning from task to evidence to inference explicit at the design stage. VAAI extends this logic by asking what happens to the evidence model when AI tools can generate or substantially alter the observable features of student performance.

Construction Rationale for the Eight Dimensions

The eight VAAI dimensions were constructed deductively. Dimensions A and B (interpretive claim and construct model) specify what is being measured (see Table 1). Dimensions C and D (evidence model and assistance boundaries) establish the evidential basis. Dimensions E and F (scoring and fairness) address interpretation and use. Dimensions G and H (feedback and governance) address maintenance and revision of the validity argument over time. Each was included where it corresponded to a distinct inferential function; overlapping functions were merged.

The levels (Level 0 through Level 4) are ordinal descriptors of specification and defensibility, not interval measures. They are not assumed to carry equal weight across contexts. The deductive construction means content validity remains provisional; whether practitioners treat the dimensions as conceptually distinct and whether level descriptors are interpreted consistently are empirical questions addressed in the research agenda.