

# Converting Knowledge into Capability: Assessment as Formation in The Age of AI

*Stephen Murgatroyd, PhD*  
*University of Alberta*

## Abstract

Assessment is usually treated as a report upon learning that has already happened. This paper argues that it is one of the mechanisms by which learning happens, and that generative artificial intelligence (GenAI) has made the distinction urgent. Capability is converted from knowledge through conceptual organization, conditionalization, enactment, feedback, metacognitive monitoring and evidence generation, and three of those six elements are things assessment does. A learner who delegates production of an assessed artifact has not principally misrepresented themselves: they have skipped the process through which their knowledge would have become usable. The formative loss precedes the evidentiary one, and experimental evidence now supports it. Knowledge is therefore the antecedent of capability rather than its competitor, and a capability economy is not a post-knowledge economy but one in which conversion has become the binding constraint. Through purposive documentary analysis across four jurisdictions and two design cases, the paper finds that assessment from kindergarten to doctorate is still predominantly knowledge-demonstration assessment, locked in place by seat-time frameworks, time-based funding and course-bound technology. The resulting evergreen paradox is a failure of formation rather than of measurement, and it compounds across a working life. The remedy proposed is system-level capability-based assessment at institutional, programme and learner levels.

**Keywords:** capability-based assessment; formative assessment; assessment for learning; capability conversion; feedback; programmatic assessment; educational purpose; generative artificial intelligence; competency-based education; lifelong learner progression

## 1. Introduction

Generative artificial intelligence (GenAI) has been discussed in education mainly as a problem of academic integrity, a framing reflected in the concentration of institutional effort on detection and enforcement (Dawson et al., 2024; Weber-Wulff et al., 2023). That framing locates the

difficulty in student conduct and implies the remedy lies in policing. But assessment works because the artifact a learner produces stands as credible evidence of what that learner knows and can do, and its value lies in the defensibility of the inference drawn from it rather than in the artifact itself (Messick, 1995; Sadler, 1989). The challenge is to be more explicit about what students should be able to do with their knowledge. GenAI severs the link at scale: 94 percent of United Kingdom undergraduates now report using it on assessed work (Stephenson & Armstrong, 2026), up from 88 percent a year earlier (Freeman, 2025).

Once that link is severed, the consequences travel in two directions. The more serious, for a journal concerned with learning, is inward: producing the artifact was part of how the capability was formed, so a learner who delegates production has skipped the process through which their knowledge would have become usable. Outwardly, a credential is a trust signal others read as evidence of underlying attributes (Arrow, 1973; Spence, 1973), and GenAI has made a large class of academic performances cheap to counterfeit, so those who read such signals have begun to adjust. The formative consequence is the less visible of the two and the more damaging.

The reframing of assessment integrity as a problem of validity is not original here. Dawson et al. (2024) argued that cheating is best understood as subsumed within validity, and that measures introduced to suppress it can degrade the validity they were meant to protect. This paper extends that argument in three directions: towards formation, since the artifact is part of the mechanism by which capability develops; towards system design, since a structural rather than task-level response (Corbin et al., 2025) must reach into what credentials certify; and towards distribution, since inferences are drawn by readers whose reading is already patterned by race, Indigeneity, gender and migration status.

The paper makes five connected claims. The first is that assessment is constitutive of capability rather than a report upon it, because enactment, feedback and evidence generation are simultaneously the mechanisms by which capability forms and is evidenced. Second, capability is derived from knowledge, so knowledge is its antecedent rather than its competitor. Third, the shift now underway is from a knowledge economy to a capability economy, in which the scarce asset is not what a person knows but what they can reliably do with it, making conversion rather than accumulation the binding constraint. Fourth is the evergreen paradox, restated as a paradox of formation, which compounds across a learning life. Fifth is that the remedy is system-level capability-based assessment, at institutional, programme and learner levels simultaneously.

The paper positions itself against two prevailing responses. Restriction removes AI from the assessment environment by returning to supervised examination, restoring interpretability at the cost of authenticity. Accommodation permits AI use and adjusts criteria at the margins while leaving the underlying logic intact, most developed in the AI Assessment Scale (Perkins et al., 2024), since reframed by its authors as a redesign framework (Perkins et al., 2025); scaling permission does not change what a task can evidence. Both treat GenAI as a contaminant of an otherwise sound system, whereas the argument here is that it has revealed a weakness predating it, namely assessment certifying the production of artifacts rather than the capability they were assumed to indicate (Boud, 2000; Gulikers et al., 2004).

## **2. Materials and Methods**

This is a conceptual paper supported by purposive documentary analysis and two illustrative design cases. The analytic procedure is set out below in a form that permits scrutiny and replication.

### ***2.1 Analytic approach***

The paper follows the conventions of conceptual and theory-building scholarship in education, in which the contribution lies in the construction and defence of an argument rather than in new empirical data. Documentary sources establish what assessment systems specify, fund and record, and test the argument against the bodies that regulate them.

### ***2.2 Documentary corpus***

Documents were selected purposively rather than exhaustively, to represent the range of positions rather than establish prevalence. Four jurisdictions were examined - Canada, Australia, the United Kingdom and the United States - chosen for the visibility of their reform activity. The corpus comprises qualifications frameworks and their reviews, quality assurance guidance on assessment in AI-integrated environments issued between 2023 and 2026, studies of competency-based education, workload and student aid instruments, and published capability frameworks. Empirical claims about the effects of GenAI on learning are drawn from the experimental and meta-analytic literature rather than the corpus.

### ***2.3 Analytic procedure***

Documents were read against three questions. What does this document treat as evidence of learning? What kind of capability does it name as valuable, and does it specify how that

capability is assessed? And what does it make structurally possible or impossible for an institution wishing to assess capability directly? Sections 7.2 and 7.4 then examine two design cases, establishing feasibility rather than effect.

### 3. How Capability Is Formed

#### *3.1 Knowledge as antecedent: the conversion mechanism*

To say that capability, rather than knowledge, is the scarce asset invites the conclusion that the disciplines transmitting knowledge can be reduced at no cost. That conclusion is false, and it is being acted upon in the programme closures documented in Section 6. Capability is not an alternative to knowledge; it is what knowledge becomes when converted.

The term capability is used here in preference to skill, and the distinction is conceptual rather than economic. A skill is a discrete, specifiable proficiency. A capability, in the human development tradition, is a person's effective freedom and capacity to act, converting resources including knowledge and skill into valued functioning under real conditions (Nussbaum, 2011; Sen, 1999). A skill can be demonstrated in isolation; a capability only in application, under ambiguity, constraint and consequence. That distinction makes this an argument about learning, since the conditions of a capability's demonstration are also those of its formation.

The Evergreen Model (The Productivity Project, 2026b) distinguishes four components differing in durability and function: evergreen knowledge, the theories, concepts and disciplinary ways of reasoning that retain explanatory value across changing contexts; evergreen skills, the portable cognitive, social and self-regulatory capabilities through which knowledge is mobilized in action; domain knowledge, the field-specific understanding required to interpret problems; and domain skills, the applied capabilities required for particular tasks. The relationship between the first two is directional. Critical thinking is not a module. It is what happens to statistical reasoning, causal inference or ethical theory when a person must use them repeatedly, under varied conditions, with feedback.

Conversion has a clear mechanism, and each element can fail independently. Conceptual organization connects ideas into interpretive structures rather than storing them as isolated facts. Conditionalization attaches knowledge to the circumstances of its use, so that the learner understands not only what a concept means but when it applies; knowledge that is not conditionalized remains inert, correctly held but unavailable in unfamiliar contexts (Bransford et al., 2000; National Academies of Sciences, Engineering, and Medicine, 2018; Whitehead,

1929). Enactment puts knowledge into action, feedback refines the performance, and metacognitive monitoring lets the learner judge whether their understanding suffices. Evidence generation makes the sequence observable, and transfer is a product of that sequence rather than of exposure to content (Bransford & Schwartz, 1999; Perkins & Salomon, 2012).

Two consequences follow from these observations. The first is that a system can hold abundant knowledge and produce little capability. Canada illustrates the gap sharply: it records the highest tertiary attainment rate in the OECD, roughly 65 percent of adults aged 25 to 64 against an OECD average of 41 percent (OECD, 2025), alongside labour productivity growth that has persistently lagged its comparators (The Productivity Project, 2026c). Attainment is not conversion. The second consequence is central to this paper. Assessment sits inside the conversion mechanism, since enactment, feedback and evidence generation are three of its six elements, and it is therefore one of the primary means by which capability is formed.

### ***3.2 Assessment inside the conversion mechanism***

This places the argument within the assessment for learning tradition rather than alongside it. Black and Wiliam (1998) established that assessment generating information used to adapt subsequent teaching produces substantial gains, largest for lower-attaining learners. Hattie and Timperley (2007) showed the effect depends on what feedback addresses, with feedback directed at process and self-regulation outperforming feedback directed at task or person. Nicol and Macfarlane-Dick (2006) reframed formative assessment as the development of self-regulation, and Carless and Boud (2018) extended this to feedback literacy. Each describes one element of the conversion mechanism: feedback refines enactment, self-regulation is metacognitive monitoring, and feedback literacy lets the loop close without an instructor.

The contribution here is to specify what that literature leaves implicit: what the feedback loop is converting. Formative assessment scholarship describes the mechanism with precision but treats its object as attainment against stated outcomes. The Evergreen architecture supplies a differentiated object, since the same loop applied to domain skill produces proficiency that depreciates while applied to evergreen knowledge it produces judgement that does not.

### ***3.3 Whose capability is read: the distributional condition***

The argument has so far treated the reading of a credential as though the reader responded only to its informational content. Field experiments in which otherwise identical applications differ only in the name attached find substantial and persistent differences in callback rates by racial and ethnic signal in the United States (Bertrand & Mullainathan, 2004; Gaddis, 2015), Canada

(Oreopoulos, 2011) and Australia (Booth et al., 2012). Gaddis (2015) is the most directly relevant, since it varied institutional selectivity alongside racial signal and found a more selective credential did not close the gap. Qualification-adjusted earnings differentials persist for Indigenous workers in Canada (Pendakur & Pendakur, 2011), and the wealth returns to a degree differ sharply by race in the United States (Hamilton et al., 2015).

Two key observations follow from this. The first strengthens the argument: if a credential is discounted on grounds unrelated to what its holder can do, evidence speaking directly to demonstrated capability is more valuable, not less. The second qualifies it: because such discrimination operates on signals other than the credential, the equity case for capability instruments is a hypothesis rather than a finding.

There is also a design warning. Capability assessment shifts weight from standardized instruments to human judgement, the mechanism through which patterned discounting operates. Behavioural descriptors written from a single cultural standpoint will read difference as deficit unless developed with the communities whose learners are judged against them, and the skill to develop such frameworks is not abundant.

### ***3.4 How capability evidence is read beyond the institution***

The possession of codified knowledge has ceased to be scarce, and therefore ceased to be informative as a signal, while the capacity to do something reliable with what is known has become scarcer (Brynjolfsson et al., 2023; Mehdi & Frenette, 2026). Skepticism about what a credential evidences long predates GenAI: Berg (1970) found employer ratings of productivity broadly similar across levels of attainment, Collins (2019) documented credential inflation as an arms race in which educational requirements rise without corresponding change in the work, and research on degree inflation reached the same conclusion from the employer side (Fuller & Raman, 2017). Two contemporary findings matter more. Employers do not currently distinguish graduates of competency-based programmes from those assessed conventionally, so evidence generated inside an institution does not reach the party it was meant to persuade (Glover & Garrison-Duncan, 2025; Sigelman et al., 2024). And earnings, the proxy most often used to establish the value of a qualification, measure something other than individual productivity (Darity & Underwood, 2021). If the standard proxy is itself contaminated, the case for capability evidence does not rest only on GenAI having degraded the signal.

### ***3.5 The capability economy is not the vocationalism argument***

Arguments for capability assessment are frequently read as arguments for narrowing education to employability. The position taken here is the reverse. The capabilities named throughout - judgement under uncertainty, ethical reasoning, communication across difference, learning how to learn, working with intelligent systems - are the historic claims of liberal education, and if evergreen skills are formed from evergreen knowledge, the disciplines supplying that knowledge are their raw material. A policy reducing provision in the liberal arts and pure sciences while declaring critical thinking most needed removes the antecedent of the outcome it has just prioritized. Policymakers have asked whether employers value critical thinking, and they clearly do, but that treats capability as something which can be demanded into existence. The question that matters is how it is formed.

Liberal education has historically claimed to develop these capabilities and has almost never assessed whether it does. On the account given above, that omission is serious, because a programme rich in evergreen knowledge that does not require enactment, supply structured feedback or generate evidence has converted nothing.

A more precise statement of scope comes from Biesta (2009), who argues that education serves three overlapping and sometimes conflicting functions: qualification, the acquisition of knowledge and skill required to do something; socialization, induction into the norms and ways of working of a practice; and subjectification, the emergence of the learner as someone able to think and act independently. The argument here belongs to the first and reaches into the second through the supervisory practices discussed in Section 8.2.

Biesta warns that a system attending only to qualification remains in the domain of training, a risk live for any framework organized around demonstrable capability (Biesta, 2022). Capability assessment is not thereby misconceived, but its reach must be stated directly: it can establish what a graduate is able to do and how they have been formed into a practice; it cannot establish who they have become. An institution assessing capability perfectly would still owe its learners an account of the third function, and that account is not supplied here.

## **4. Signal Collapse: What GenAI Does to Assessment**

### ***4.1 Assessment as evidence and as trusted signal***

Assessment performs two functions that are often conflated. It supports learning by informing the learner about the gap between current and desired performance (Sadler, 1989). And it generates evidence for others: the learner performs, the performance leaves a trace, the trace is

judged, and the record is read by someone who was not present. Validity is a property of the inference at the end of that chain, not of the instrument at its start (Messick, 1995).

#### *4.2 The severing of trace from process*

GenAI breaks the reliability of the trace and does so unevenly in an informative way. Tasks whose value lies in producing a text are most affected, because the artifact is precisely what a GenAI system produces. Tasks requiring performance under observation, defence of reasoning under challenge, or demonstration across varied contexts are far less affected, because the trace is not separable from the person producing it. A compulsory undergraduate course that moved from an AI-permissive take-home examination to an AI-restricted supervised one, with content, outcomes, criteria and examiner held constant, saw failure rates rise from between two and six percent to 18.4 percent (Brattli et al., 2026).

This is the central mechanism of the argument. It is not that students cheat. It is that the same assessment instrument, unchanged, now supports a weaker inference while the institution continues to make a strong one. A second, less discussed mechanism, even though the consequences are more damaging: a learner who delegates enactment has skipped the step at which their knowledge would have been conditionalized, so the loss is formative before it is evidentiary.

That proposition is no longer only theoretical. In a randomized field experiment with almost a thousand secondary mathematics students, access to an unrestricted GPT-4 interface during practice raised performance on practice problems by 48 percent; when access was removed for the subsequent examination, the same students performed 17 percent worse than those who had never had access (Bastani et al., 2025). A second arm, using a tutor configured to withhold solutions and supply teacher-designed hints, produced the practice gain without the loss. What harms learning is not the presence of the technology but the delegation of the step at which conversion occurs.

Read alongside the countervailing evidence, this becomes a stronger claim. A randomized trial in undergraduate physics found a pedagogically designed AI tutor produced more learning in less time than an active learning class built on the same principles (Kestin et al., 2025), and meta-analytic synthesis reports positive average effects on academic performance and higher-order thinking alongside a reduction in mental effort (Deng et al., 2025). That reduction is the

finding that matters here. Effort expended in enactment is not a cost good design should minimize; it is substantially where the learning is. What determines whether GenAI builds capability or forecloses it is task design, which assessment specifies.

#### *4.3 Why restriction and detection fail*

Restricting GenAI returns assessment to supervised conditions and has the merit of restoring the trace. But it certifies performance under conditions no graduate will encounter, cannot be applied to the extended iterative work carrying most of the value in professional education, and measures a narrower construct than the assessment it replaces (Brattli et al., 2026). Detection tools are unreliable; none have more than 80 percent accuracy, and they are all systematically biased against writers whose first language is not English (Liang et al., 2023; Weber-Wulff et al., 2023), and detection concedes the framing challenged here.

The alternative is to change what is assessed. If the assessment requires the learner to do something a GenAI system cannot do on their behalf - perform under observation, defend reasoning under challenge, demonstrate capability across varied contexts, or direct and correct an intelligent system and account for the result - then whether AI was used ceases to be the operative question. This is what the authentic assessment tradition has argued for three decades (Gulikers et al., 2004; Wiggins, 1998), and it restores the inference rather than policing the artifact.

#### *4.4 What students report and how regulators have responded*

Students are not confused about what has changed. In the most recent United Kingdom survey, 95 percent report using AI in at least one way and 94 percent in support of assessed work, while 65 percent say assessment has changed significantly in response (Stephenson & Armstrong, 2026). Institutional direction has not kept pace: only 36 percent feel encouraged to use these tools and only 38 percent are provided with them, and that access itself is unequally distributed. Students are being asked to judge when delegation forecloses their own learning without being taught the basis for such a judgement.

Quality assurance bodies have reached a similar conclusion. Australia's Tertiary Education Quality and Standards Agency has argued that assuring learning in an AI-integrated environment requires better evidence of learning processes - portfolios, reflective work, trace data and structured self-assessment - rather than reliance on terminal artifacts, and it names the

capabilities to be assured as evaluative judgement, critical thinking, ethical reasoning and metacognitive regulation (Tertiary Education Quality and Standards Agency, 2026). Those are the evergreen skills under another description, and they align with the account of evaluative judgement developed by Bearman et al. (2024). The agency's earlier guidance set out three reform pathways - programme-wide reform, unit-level assurance, and a hybrid - which is regulatory endorsement of the programme-level argument advanced below (Tertiary Education Quality and Standards Agency, 2025). The issue is thereby reframed from conduct policy to assurance of standards: an institution unable to evidence what its graduates can do has a quality problem rather than an integrity problem.

## **5. The Assessment Gap from Kindergarten to Doctorate**

If the argument to this point is accepted, the appropriate response is capability-based assessment. Education systems practise it rarely, and almost never in a form that survives contact with a formal credential. The obstacles are structural and load-bearing: not obstructions standing in the way of something better but supports carrying weight that would otherwise have to be carried by something new.

### ***5.1 The pattern from schooling to doctorate***

The pattern is consistent across the lifespan and differs only in its mechanism. In compulsory schooling, assessment is nominally outcomes-based, but the accountability layer above the classroom is dominated by standardized measurement of knowledge recall, because that construct scales cheaply, and classroom assessment converges on whatever accountability rewards. Schools are therefore where collaboration, persistence and ethical reasoning are most explicitly claimed as purposes and least likely to be assessed in any form leaving a durable record (Murgatroyd & Couture, 2026). In post-secondary education, competency-based provision has existed for decades in the United States (Nodine, 2016) yet remains a minority and usually partial practice, with the Higher Education Quality Council of Ontario finding programmes still operating within seat-time structures, so that credit is conferred at the end of a fourteen-week semester regardless of when a capability was demonstrated (Gallant et al., 2026). Graduate education is assumed exempt because the viva assesses capability, and it does, but it assesses a narrow band once.

### ***5.2 The structural lock-ins***

Four mechanisms hold the current infrastructure in place. The first is the qualifications framework: where a framework defines credentials by instructional credit hours, semesters or years, it defines the credential in terms of time rather than demonstrated capability, as the agreed Canadian national framework does (CMEC Canadian Degree Qualifications Framework, 2007). Such a framework does not prohibit capability assessment; it ensures capability assessment cannot by itself produce a credential.

The Australian case shows the obstacle is structural rather than attitudinal. The review of the Australian Qualifications Framework concluded that the volume of learning rests on increasingly outmoded assumptions and recommended that knowledge, skills and application be defined in terms of action, not time (Australian Qualifications Framework Review Panel, 2019). The Government accepted the higher education recommendations in full in December 2019, and the framework then continued substantially as before, because the time-based definition is load-bearing for student financing, institutional funding, recognition and regulation. The attitude changed - the structure did not.

The second is student financial aid. Where eligibility is calculated in periods of study, a learner who demonstrates competence early receives no benefit and one who requires longer exhausts eligibility (Ministry of Colleges, Universities, Research Excellence and Security, 2024). The same logic operates on the funding of provision: the Productivity Commission found that mid-career workers frequently do not need the full qualifications governments fund (Productivity Commission, 2020).

The third is academic workload, calculated through formulas built on time-based metrics with a weekly maximum (Ontario Public Service Employees Union, 2021), which cannot represent supervision whose load varies with learner progression, and which sits at the heart of collective agreements. The fourth is assessment technology: learning management systems built around weekly modules, gradebooks and one-time assessments handle competency mapping, iterative feedback and longitudinal tracking poorly (Gallant et al., 2026).

Together, these four mechanisms constitute a single, mutually reinforcing infrastructure. Capability assessment is not forbidden by any of them. It is made expensive, invisible and non-transferable by all of them at once.

## **6. The Capability Paradox**

The preceding sections argue that education assesses knowledge where it should assess capability. This section makes a sharper claim: that systems are drifting towards assessing the capabilities which depreciate fastest, and that the drift is a failure of formation rather than of measurement. Capabilities do not depreciate at a uniform rate. Some retain value across decades, occupations and technologies; others lose value as the tools, regulations or procedures they attach to change. The Evergreen Model names the first evergreen knowledge and skills and the second domain-specific knowledge and skills (The Productivity Project, 2026a). Statistical reasoning holds its value while proficiency in a particular statistical package does not, and the distinction is often stated loosely, conflating the obsolescence of tools with that of the reasoning which selects them.

### ***6.1 The paradox stated***

The capability paradox is this. Education, funding, accountability and labour market reward systems appear to be prioritizing faster-depreciating domain capabilities at precisely the moment when slower-depreciating capabilities are becoming more important for adaptation and renewal (The Productivity Project, 2026c). Stated at that level, it reads as a misallocation of assessment attention. The conversion analysis makes it worse, because the paradox operates in two layers, each sufficient on its own to prevent durable capability from forming. The first is antecedent: systems reduce provision in the disciplines supplying evergreen knowledge, and thereby the material from which evergreen skills would be made. In the United Kingdom alone over the last decade, more than 700 liberal arts and 100 pure science programmes have closed (McKenzie et al., 2026; Thornborough et al., 2026). The second is procedural: systems omit the conversion process itself.

Distinguishing the layers matters because a programme can fail on either dimension while succeeding on the other, and the two failures need different remedies: curricular breadth where evergreen knowledge is thin but conversion is strong, and assessment redesign where knowledge is abundant but conversion is weak. Both such institutions can publish the same graduate attribute statement in good faith, which is part of why a decade of attention has not shifted the position.

### ***6.2 The capability paradox across a learning life***

The capability paradox has so far been described as a property of programmes at a point in time. Its more serious form is temporal. A person leaving formal education without well-formed evergreen capability is not merely under-equipped for a first job. They are under-

equipped for every subsequent act of learning, because the capabilities that failed to form are the ones later learning depends upon: organizing unfamiliar material into interpretive structure, judging when existing understanding has stopped fitting, and seeking and using feedback without an instructor present. Formative assessment scholarship names the last of these as self-regulation and feedback literacy (Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006); they are the mechanism through which a person continues to convert after the institution stops doing it for them.

This produces a compounding asymmetry that neither the depreciation nor the lifelong learning literature has articulated. Domain capability depreciates and can be replaced, provided the person can still convert. Evergreen capabilities depreciate slowly, but where it was never formed it cannot easily be acquired later, because acquiring it requires precisely the conversion capacity it constitutes. The asymmetry is therefore not a trade-off between short-term and long-term value; it is a foreclosure. A system that assesses the perishable and neglects the durable spends the learner's one richly resourced formative opportunity on the component with the shortest shelf-life, and the learner discovers the omission where it is most expensive to repair, typically mid-career, without an institution, an assessor or time. Provision for later learning cannot compensate, because it assumes the very capacity it would need to supply.

Three implications follow from this analysis for assessment design. Durable capability must be assessed longitudinally rather than terminally, because its formation is cumulative and a single occasion cannot evidence it. Feedback must be designed to build the learner's capacity to generate and use feedback independently, a different objective from delivering accurate comments (Winstone & Carless, 2019). And the record must persist beyond the programme, because a capability whose value lies in portability across a working life cannot be evidenced by an instrument expiring at graduation. The unit of assessment is the person's capabilities over time, not the course.

Assessment is where the paradox becomes concrete. Institutions publish graduate attribute statements that are sincere, often carefully constructed, and with rare exceptions not assessed. The documentary analysis found capability language in mission statements and quality assurance submissions while assessment specifications, rubrics and transcripts addressed domain knowledge and procedural skill. The research literature reproduces the asymmetry: in the most comprehensive recent Canadian study, transversal competencies appear in a footnote while the body defines a competency as a job function (Gallant et al., 2026), a definition that

cannot capture capabilities whose value lies in portability across job functions not yet in existence.

### ***6.3 Why the durable is left unassessed***

Three reasons are often given for not assessing evergreen capabilities, and a fourth is rarely named. The first is measurement difficulty: durable capabilities are context-dependent, slow to develop, visible only in performance, and cannot be assessed through a single task. The second is accountability incentive: where funding is attached to short-horizon completion and earnings, institutions direct assessment effort accordingly, and Biesta (2009) names this drift, in which systems come to value what can be measured in place of measuring what they value. The third is professional confidence: many instructors do not believe they can assess judgement defensibly, and they are correct, since few have been trained to do so. The fourth is seldom stated. Durable capabilities are left unassessed partly because they are not being intentionally formed, and where capability statements are aspirational rather than designed into a sequence of enactment and feedback, the absence of assessment accurately reflects a curriculum never built to convert.

## **7. System-Level Capability-Based Assessment**

Changes in assessment intended to build, develop and strengthen capability cannot be implemented at a single point in a system. A capability that forms cumulatively, is evidenced longitudinally and must remain legible after graduation requires coordinated arrangements at three interdependent levels, examined below with two design cases.

### ***7.1 Three levels and how they fail separately***

The institutional level supplies the conditions under which capability assessment is administratively possible: the calendar on which assessment is available, the basis on which learners are funded and progress recorded, the workload instrument through which assessor time is counted, and the technology in which evidence is held. Its function is permissive, and its failure mode is that capability assessment becomes locally heroic, abandoned when its champions move on.

The programme level supplies the architecture of what is assessed and in what sequence: the capability set, the mastery descriptors, the distribution of enactment and feedback opportunities across courses and years, and their alignment with the disciplinary knowledge from which the

capabilities are formed. This is where the conversion mechanism is designed, and its failure mode is a graduate attribute statement that no assessment in the programme evidences.

The learner level supplies the person's own capacity to run the loop and a record they carry: feedback literacy and self-regulation on the one hand, and on the other an instrument in which capability claims, mastery levels and evidence persist beyond the institution. Its failure mode is the graduate who was assessed capably for four years, cannot say what they can do, and holds a document that does not say either.

The interdependence is asymmetric, which makes partial implementation worse than it looks. Institutional change without programme change produces flexible delivery of unchanged assessment; programme change without institutional change produces local heroism; either without learner-level change produces capability the institution can evidence and the learner cannot carry; and learner-level change alone, the free-standing badge or unanchored portfolio, produces a record with nothing behind it (Wheelahan & Moodie, 2022).

## ***7.2 A working instance: the Universities of Wisconsin Flexible Option***

The Flexible Option is a competency-based programme structure operated as a collaboration of the Universities of Wisconsin, offering associate and bachelor's degrees, certificates and a micro-credential across fields including business administration, information science, health sciences and nursing (University of Wisconsin System, 2026). It is examined because it is a rare instance in which arrangements at all three levels moved together, which makes both the achievement and the residual constraint legible.

At the institutional level, the university addressed three of the four lock-ins directly. The academic calendar was replaced by rolling twelve-week subscription periods beginning every month, so the interval between readiness and assessment is bounded by weeks rather than a semester. Tuition was decoupled from the number of courses attempted and charged as a flat subscription. Assessment was separated from instruction, so it is available when the learner is ready rather than when the teaching sequence concludes.

At the programme level, competencies are specified by faculty in consultation with industry, progression is by demonstrated mastery rather than elapsed time, and prior capability converts directly, so that an experienced nurse entering the RN-to-BSN pathway is not required to sit through instruction in what she already does competently. Assessment functions here as the mechanism of progression, which is the constitutive claim of this paper implemented as an operating rule.

At the learner level, each student is paired with a success coach acting as mentor and adviser across the whole programme rather than within a single course. A self-paced competency model without a formative infrastructure is a correspondence course with better software, and its predictable failure is that learners who cannot yet regulate their own progress stall silently. The coach is the institutional response to the self-regulation problem Nicol and Macfarlane-Dick (2006) identify, made structural rather than left to individual practice.

The example establishes that the lock-ins are not laws of nature. Three were altered within a public university system, at scale, under regional accreditation, awarding degrees on the same faculty-set standards as more conventional programmes. The objection that capability-based assessment cannot be reconciled with institutional quality assurance is answered by its existence.

### ***7.3 What the case does not establish***

The fourth lock-in did not move, and the accommodation it required is instructive. Federal student aid rules are denominated in time and require institutions to establish attendance and engagement, so the Flexible Option maintains an academic plan of study, publishes rules for establishing engagement within a subscription period, and defines conditions for repeating a competency set. The programme is competency-based internally and time-legible externally, because the funding rule it could not change compelled a layer of translation. United States direct assessment programmes that developed algorithms converting competency units into credit hours show the same phenomenon (Gallant et al., 2026): three levels moved, the fourth did not, and the cost was borne in machinery whose only function is to make capability legible to a rule about time.

Three limits follow. The subscription period is twelve weeks, so batching is reduced rather than removed; the credential remains denominated in credit hours, so what leaves the institution is again a measure of time; and there is no independent published evidence on whether Flexible Option graduates are judged more capable by employers.

### ***7.4 The learner-held record: capability passports***

The Flexible Option addresses the learner level through coaching, which builds the capacity to convert, but not through a record that persists. That is the function of the second design case.

A digital capability passport holds capability claims, mastery levels and supporting evidence in a form the learner controls and a third party can verify, sitting alongside rather than replacing

the credential. The instances drawn on here are unpublished instruments developed in consulting engagements, so what follows is a design synthesis rather than a finding. Four components recur. A capability architecture defines a set of capabilities organized into domains, in one instance twenty across six domains at four mastery levels. A mastery scale supplies performance descriptors at each level, the component most often done badly, since descriptors in comparative language produce unreliable judgements.

An evidence rule specifies what may count as evidence at each level, how many independent instances are required and who may verify. This distinguishes a passport from a self-report, and it is where GenAI is either handled or ignored, since admitting unsupervised written artifacts as sole evidence reproduces the validity problem the instrument was meant to solve. It is also where distributional risk is contained, because multiple instances, verification by more than one assessor and observable descriptors narrow the discretion through which patterned judgement operates. A portability layer then carries claims to third parties in verifiable form (Curtin University, 2024), and the micro-credential experience shows how quickly such instruments proliferate into incoherence when many issuers make non-comparable claims (Murgatroyd, 2025; Wheelahan & Moodie, 2022).

One design decision follows from the temporal argument above. In one instance the capability architecture was built independently of the host programme rather than mapped onto its courses. A passport mapped onto a curriculum certifies that curriculum; one built independently can receive evidence from any source, including work, volunteering and prior learning, which is what makes it usable across a working life.

What the two cases jointly establish is narrower than advocacy for either might suggest. Capability architectures can be specified at a granularity supporting reliable judgement, evidence rules robust to generative production can be written, institutional and programme arrangements can be rebuilt around mastery within an accredited public system, and portability is technically solved. What remains unestablished is effect.

## **8. Implications for Practice**

The transition from knowledge assessment to capability assessment cannot be achieved by asking instructors to assess differently within systems designed to prevent it. If assessment is constitutive of conversion rather than descriptive of its results, what follows comprises curriculum recommendations as much as measurement recommendations, and an institution

implementing them as a reporting exercise will change what it records without changing what it produces.

### ***8.1 Programme level: redesigning the assessment architecture***

The first requirement is architectural. Programmes currently assess at the level of the course, an administrative unit rather than the level of capability, and a capability develops across courses and years. Constructive alignment established that assessment, objectives and teaching activity must be designed together (Biggs, 1996); the extension required now is that alignment operate at programme level. Four design principles follow. Assess performance, not artifacts: the unit of evidence should be an observed or verifiable performance under conditions approximating those of use, including the presence of intelligent tools. Assess repeatedly and cumulatively, since the difference between a lucky performance and a capability is repetition across varied contexts (Sadler, 1989). Build assessment that outlasts the programme (Boud, 2000). And treat AI use as part of the assessed performance, since capability includes directing, interrogating and correcting intelligent systems and knowing when not to (Bearman et al., 2024).

A fifth principle governs the other four: assessment must be designed as the conversion mechanism rather than as its audit. Where the preceding principles are followed, the assessment system is itself performing the enactment, feedback and evidence work through which knowledge becomes capability, which is why assessment redesign undertaken without curriculum redesign will fail.

This is not an untried proposition. Programmatic assessment, developed in health professions education over two decades, is a whole-system design built on four features: purposeful longitudinal aggregation of many low-stakes data points, triangulation across modalities, a formative-first orientation in which individual data points are used principally for feedback rather than decisions, and committee-based synthesis of aggregated evidence for high-stakes progression decisions (Heeneman et al., 2021; van der Vleuten & Schuwirth, 2005; van der Vleuten et al., 2012). Each of the five principles above has an implemented counterpart there, and the accumulated evidence describes what such a system asks of an institution rather than what it promises (Govaerts et al., 2022). The argument here differs in extending the design to general education, where the capability set cannot be derived from a defined practice, and in treating the record as something that must leave the institution with the learner.

### ***8.2 The work of assessment: what capability assessment asks of faculty***

Instructor capability is the pivotal constraint and is consistently underestimated. Capability assessment asks instructors to make defensible judgements about performance rather than apply marking schemes to artifacts, a skill closer to clinical supervision than to conventional marking, and research finds faculty reticence widespread (Gallant et al., 2026). The reticence is rational: the request is that they exercise, and defend, a form of professional judgement most were never trained in and which their institutions have historically protected them from having to exercise.

Four kinds of work follow, each with a cost institutions consistently fail to budget. The first is the construction of behavioural performance descriptors, difficult technical writing and where most capability frameworks fail, since descriptors written in comparative language produce judgements that cannot be defended. The second is calibration practice, in which groups of assessors judge the same performances and reconcile their differences; inter-assessor reliability is the property on which the credibility of the whole system rests, and calibration is the mechanism through which systematic differences in how assessors judge particular groups become visible and correctable. The third is collective decision-making: in programmatic assessment no single data point carries a high-stakes consequence, and progression decisions are made by a committee synthesizing the aggregate (Heeneman et al., 2021). That is the principal safeguard against the idiosyncratic assessor and also the principal cost, since committee time is real time no workload formula built on contact hours can represent. The fourth is supervision and coaching capability, since capability develops through repeated performance with expert feedback under conditions of increasing complexity (Collins et al., 1989; Schön, 1983).

The implementation literature is candid about what this means. Programmatic assessment succeeds where a culture of holistic professional judgement has been deliberately built and fails where the instruments were installed without it, and the effort is substantial, sustained and cultural rather than procedural (Govaerts et al., 2022). Institutions that have attempted it describe faculty development as the largest single implementation cost. Any jurisdiction proposing capability-based assessment as a policy direction is therefore proposing a significant programme of academic staff development.

Calibration and committee synthesis are necessary but not sufficient, because assessor reliability concerns the accuracy of judgement rather than its formative effect. Instructors must also design feedback for the development of the learner's own evaluative capacity rather than the correction of the immediate artifact: feedback that corrects an artifact completes one enactment, while

feedback that builds the capacity to evaluate one's own work allows conversion to continue after the programme ends (Winstone & Carless, 2019). It is also the aspect of assessment least susceptible to being done at scale by machine.

### ***8.3 Institutional and system levels: policy and regulatory change***

The structural lock-ins are policy artifacts and can be changed by policy. Qualifications frameworks can be amended to permit credentials to be defined by demonstrated capability as well as by duration, and the Australian review shows a national panel will recommend precisely this when asked (Australian Qualifications Framework Review Panel, 2019). That case also shows acceptance is not implementation, so the useful policy question is which body holds the mandate, timetable and budget to rebuild the dependencies the existing definition carries. Student aid rules can be amended so eligibility is not consumed by time, workload instruments renegotiated to represent supervision and committee judgement, and quality assurance bodies required to seek evidence that graduates possess the capabilities institutions claim to develop.

The most powerful lever is already moving. When a regulator asks an institution to evidence the graduate attributes in its own published statements, the institution must either produce the evidence or amend the statement, and either outcome improves on a claim made annually and tested never. Intergovernmental funding agreements are a further lever, since an agreement specifying the outcomes for which money is transferred determines what jurisdictions collect evidence about, and such measures should be reported disaggregated by learner group.

One further point concerns funding rather than assessment. If evergreen capabilities are formed from evergreen knowledge, decisions that reduce provision in the liberal arts and fundamental sciences are capability decisions, whether or not they are recognized as such.

## **9. Limitations and Research Agenda**

Five limitations qualify the argument. The documentary corpus is purposive and small, drawn from four jurisdictions selected for the visibility of their reform activity, and supports claims about the range of positions and the mechanisms in play but not about prevalence.

The signalling argument is theoretical. The proposition that employer trust in knowledge-assessed credentials is weakening is consistent with the evidence assembled here but is not directly measured.

The distributional argument carries a limitation of its own. The audit and earnings evidence is drawn from several jurisdictions and studies conducted for other purposes, and none of it concerns capability records, so it establishes a scope condition rather than a finding.

The two design cases are of different evidentiary standing, and neither reports outcome data. The Flexible Option is publicly documented; the capability passport instances are not. System-level capability-based assessment is therefore theoretically warranted, demonstrably feasible in part, and empirically untested at the level of effect.

The fifth limitation is philosophical rather than methodological. The instruments recommended above make the person more legible, more comparable and more continuously assessed than the transcript they replace, and Biesta (2004, 2013) has argued that this direction of travel substitutes the language of learning for the question of educational purpose. The objection cannot be met by better instrument design, because it is an objection to instrumentation; it is registered as a boundary on what is claimed here.

Four research questions follow, and they are the questions this paper cannot answer.

Are capability judgements reliable across assessors and institutions, and equally reliable across learner groups? Inter-rater reliability studies at scale are the precondition for any national capability infrastructure, and their absence is the strongest current argument against proceeding.

Do employers read capability evidence, and does it change hiring decisions? This requires experimental comparison of employer response to conventional transcripts and to capability records for matched candidates, with candidate characteristics varied so the distributional question is answered alongside the efficacy question.

Which capabilities are in fact durable? The evergreen distinction is theoretically developed and empirically thin, and longitudinal tracking of capability value across occupational transitions would allow it to be tested rather than asserted.

Does capability-based assessment produce better learning, or only better evidence of learning? These are separable outcomes, and the field has an interest in not conflating them. The designs now used to study GenAI and learning, in which access to a tool is varied and performance measured after it is withdrawn (Bastani et al., 2025), offer a usable template.

## 10. Conclusions

GenAI has been read as a threat to academic integrity. It is better understood as a disclosure. It has revealed that much of what education assesses was only ever a proxy for capability, that the proxy has become unreliable, and that the systems built on it were carrying more weight than the evidence could bear. The response proposed here is to assess the thing itself, which requires accepting that assessing capability is more expensive, and that no redesign will succeed while qualifications frameworks define credentials as time, funding rules pay for time, workload agreements measure time, and technology records time.

Assessing capability rather than knowledge is not a demotion of knowledge. Capability is what knowledge becomes when a system does the work of converting it, and a jurisdiction that responds to the capability agenda by narrowing its disciplinary base has misread the argument. The capability economy is not a post-knowledge economy. It is one in which knowledge has stopped being sufficient without ceasing for a moment to be necessary.

The transition is therefore a systems redesign operating at three levels at once. Institutional arrangements must make capability assessment administratively possible, programme architecture must make conversion deliberate rather than incidental, and learners must acquire both the capacity to convert independently and a record that outlasts the institution. The Flexible Option shows the first two are achievable within an accredited public system, and that the level which does not move exacts its cost in translation machinery. Partial implementation is not merely slower; it is actively misleading.

Making capability legible is not the same as having it read fairly. Credentials are already discounted and amplified on grounds unrelated to what their holders can do, and a capability infrastructure built without disaggregated monitoring would produce a more precise record of capability and the same distribution of opportunity. The alternative to making this transition deliberately is having it made elsewhere: employers unable to trust credentials have begun building their own assessment, and skills-based hiring commitments have moved further in assessment practice than in advertised requirements (Sigelman et al., 2024). If education systems decline to assess what they claim to develop, others will assess it for them.

## **Data Availability Statement**

No new data were created in the documentary analysis reported here; all documents in the corpus are publicly available and cited in full. The two design cases discussed in Sections 7.2 and 7.4 are drawn respectively from publicly available institutional documentation and from unpublished instruments developed under consulting engagements which cannot be released.

## References

- Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics*, 2(3), 193-216.  
[https://doi.org/10.1016/0047-2727\(73\)90013-3](https://doi.org/10.1016/0047-2727(73)90013-3)
- Australian Qualifications Framework Review Panel. (2019). *Review of the Australian Qualifications Framework: Final report*. Department of Education, Skills and Employment.  
<https://www.education.gov.au/australian-qualifications-framework-review>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122.  
<https://doi.org/10.1073/pnas.2422633122>
- 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>
- Berg, I. (1970). *Education and jobs: The great training robbery*. Praeger.
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4), 991-1013. <https://doi.org/10.1257/0002828042002561>
- Biesta, G. J. J. (2004). Against learning: Reclaiming a language for education in an age of learning. *Nordisk Pedagogik*, 23, 70-82.
- Biesta, G. J. J. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, 21(1), 33-46. <https://doi.org/10.1007/s11092-008-9064-9>
- Biesta, G. J. J. (2013). *The beautiful risk of education*. Paradigm Publishers.
- Biesta, G. J. J. (2022). *World-centred education: A view for the present*. Routledge.
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347-364.  
<https://doi.org/10.1007/BF00138871>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. <https://doi.org/10.1080/0969595980050102>
- Booth, A. L., Leigh, A., & Varganova, E. (2012). Does ethnic discrimination vary across minority groups? Evidence from a field experiment. *Oxford Bulletin of Economics and Statistics*, 74(4), 547-573. <https://doi.org/10.1111/j.1468-0084.2011.00664.x>
- Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2), 151-167. <https://doi.org/10.1080/713695728>

- Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academy Press.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24, 61-100. <https://doi.org/10.2307/1167267>
- Brattli, H., Utne, A., & Lynch, M. (2026). Assessment validity in the age of GenAI: A natural experiment. *Informatics*, 13(4), Article 56. <https://doi.org/10.3390/informatics13040056>
- Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *GenAI at work* (NBER Working Paper No. 31161). National Bureau of Economic Research. <https://doi.org/10.3386/w31161>
- 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>
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Lawrence Erlbaum.
- Collins, R. (2019). *The credential society: An historical sociology of education and stratification*. Columbia University Press. (Original work published 1979)
- Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, 50(7), 1087-1097. <https://doi.org/10.1080/02602938.2025.2503964>
- Council of Ministers of Education, Canada. (2007). *Ministerial statement on quality assurance of degree education in Canada*. <https://www.cmec.ca/Publications/Lists/Publications/Attachments/95/QA-Statement-2007.en.pdf>
- Curtin University. (2024). *Competency passport*. Assessment 2030. <https://www.curtin.edu.au/assessment2030/assessment-design-studio/competency-passport/>
- Darity, W. A., Jr., & Underwood, M. (2021). *Reconsidering the relationship between higher education, earnings, and productivity*. Postsecondary Value Commission.
- 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>
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, Article 105224. <https://doi.org/10.1016/j.compedu.2024.105224>
- Freeman, J. (2025). *Student generative AI survey 2025* (HEPI Policy Note 61). Higher Education Policy Institute. <https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/>

- Fuller, J. B., & Raman, M. (2017). *Dismissed by degrees: How degree inflation is undermining U.S. competitiveness and hurting America's middle class*. Harvard Business School.
- Gaddis, S. M. (2015). Discrimination in the credential society: An audit study of race and college selectivity in the labor market. *Social Forces*, 93(4), 1451-1479. <https://doi.org/10.1093/sf/sou111>
- Gallant, L., Patel, J., Tishcoff, R., & Colyar, J. (2026). *Challenges and opportunities for competency-based education in Ontario*. Higher Education Quality Council of Ontario.
- Glover, H., & Garrison-Duncan, A. (2025). *The role of competency-based education in strengthening trust between employers and higher education*. Upskill America and Competency-Based Education Network.
- Govaerts, M., van der Vleuten, C., & Schut, S. (2022). Implementation of programmatic assessment: Challenges and lessons learned. *Education Sciences*, 12(10), Article 717. <https://doi.org/10.3390/educsci12100717>
- Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67-86. <https://doi.org/10.1007/BF02504676>
- Hamilton, D., Darity, W. A., Jr., Price, A. E., Sridharan, V., & Tippett, R. (2015). *Umbrellas don't make it rain: Why studying and working hard isn't enough for Black Americans*. The New School, Duke University, and the Insight Center for Community Economic Development.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Heeneman, S., de Jong, L. H., Dawson, L. J., Wilkinson, T. J., Ryan, A., Tait, G. R., Rice, N., Torre, D., Freeman, A., & van der Vleuten, C. P. M. (2021). Ottawa 2020 consensus statement for programmatic assessment - 1. Agreement on the principles. *Medical Teacher*, 43(10), 1139-1148. <https://doi.org/10.1080/0142159X.2021.1957088>
- Kestin, G., Miller, K., Klaes, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, Article 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- McKenzie, C., Roblero, P., & Cullen, R. (2026). *Staff numbers are falling in some cold spot areas: Mapping SHAPE academic staff briefing*. British Academy. <https://www.thebritishacademy.ac.uk/policy-and-research/british-academy-shape-observatory/higher-education-and-research/mapping-shape-academic-staff/staff-numbers-are-falling-in-some-cold-spot-areas/>
- Mehdi, T., & Frenette, M. (2026, January 28). Canadian employment trends in the era of generative artificial intelligence: Early evidence. *Economic and Social Reports*. Statistics Canada. <https://doi.org/10.25318/36280001202600100003-eng>

- 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>
- Ministry of Colleges, Universities, Research Excellence and Security. (2024). *How to maintain eligibility for OSAP and the Ontario Learn and Stay Grant*. Government of Ontario.
- Murgatroyd, S. (2025). Micro-credentials: Not working as advertised. In G. Durak & S. Çankaya (Eds.), *Integrating micro-credentials with AI in open education*. IGI Global.
- Murgatroyd, S., & Couture, J.-C. (2026). *AI unplugged: The hype and hope in education futures*. Routledge.
- National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press. <https://doi.org/10.17226/24783>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. <https://doi.org/10.1080/03075070600572090>
- Nodine, T. R. (2016). How did we get here? A brief history of competency-based higher education in the United States. *Competency-Based Education*, 1(1), 5-11. <https://doi.org/10.1002/cbe2.1004>
- Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Harvard University Press.
- OECD. (2025). *Education at a glance 2025: OECD indicators*. OECD Publishing. <https://doi.org/10.1787/1c0d9c79-en>
- Ontario Public Service Employees Union. (2021). *SWF - Standard Workload Formula*.
- Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3(4), 148-171. <https://doi.org/10.1257/pol.3.4.148>
- Pendakur, K., & Pendakur, R. (2011). Aboriginal income disparity in Canada. *Canadian Public Policy*, 37(1), 61-83. <https://doi.org/10.3138/cpp.37.1.61>
- Perkins, D. N., & Salomon, G. (2012). Knowledge to go: A motivational and dispositional view of transfer. *Educational Psychologist*, 47(3), 248-258. <https://doi.org/10.1080/00461520.2012.693354>
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS): A framework for ethical integration of GenAI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6), 49-66. <https://doi.org/10.53761/q3azde36>
- Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the Artificial Intelligence Assessment Scale (AIAS): A refined framework for educational assessment. *Journal of University Teaching and Learning Practice*, 22(7). <https://doi.org/10.53761/vgnaq267>

- Productivity Commission. (2020). *National Agreement for Skills and Workforce Development review* (Study report). Commonwealth of Australia.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144. <https://doi.org/10.1007/BF00117714>
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
- Sen, A. (1999). *Development as freedom*. Oxford University Press.
- Sigelman, M., Fuller, J., & Martin, A. (2024). *Skills-based hiring: The long road from pronouncements to practice*. Burning Glass Institute.
- Spence, M. (1973). Job market signalling. *The Quarterly Journal of Economics*, 87(3), 355-374. <https://doi.org/10.2307/1882010>
- Stephenson, R., & Armstrong, C. (2026). *Student generative artificial intelligence survey 2026* (HEPI Report 199). Higher Education Policy Institute. <https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/>
- Tertiary Education Quality and Standards Agency. (2025). *Enacting assessment reform in a time of artificial intelligence*. Australian Government. <https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/enacting-assessment-reform-time-artificial-intelligence>
- Tertiary Education Quality and Standards Agency. (2026). *Assuring quality learning in a gen AI-integrated future: The role of adaptive capabilities*. Australian Government. <https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assuring-quality-learning-gen-ai-integrated-future-role-adaptive-capabilities>
- The Productivity Project. (2026a, June 1). *Capability depreciation: Why capabilities endure, decline, or regain value* [Discussion paper]. LearningCITY.
- The Productivity Project. (2026b, June 1). *The Evergreen Model: Adaptive by design* [Discussion paper]. LearningCITY. <https://www.learningcity.ca/series-4>
- The Productivity Project. (2026c, June 1). *The Evergreen Paradox: Canada's shift towards short-term readiness* [Discussion paper]. LearningCITY.
- Thornborough, J., Roblero, P., & Cullen, R. (2026). *Pathways for early career SHAPE academics are in decline: Mapping SHAPE academic staff briefing*. British Academy.
- University of Wisconsin System. (2026). *UW Flexible Option: Competency-based degree and certificate programs*. <https://flex.wisconsin.edu/>
- van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, 39(3), 309-317. <https://doi.org/10.1111/j.1365-2929.2005.02094.x>

- 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>
- Wheelahan, L., & Moodie, G. (2022). Gig qualifications for the gig economy: Micro-credentials and the "hungry mile". *Higher Education*, 83(6), 1279-1295. <https://doi.org/10.1007/s10734-021-00742-3>
- Whitehead, A. N. (1929). *The aims of education and other essays*. Macmillan.
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.
- Winstone, N. E., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge. <https://doi.org/10.4324/9781351115940>