

A Multi-Agent Architecture for Valid, Reliable, and Scalable Skills Assessment

Megan N. Imundo¹, Kjorte Harra¹, Lesley Reilly¹, Cory Hammon¹, Lauren Zito¹,
Catrina Nieser¹, Betheny Gross^{1,2}

¹WGU Labs; ²Center on Reinventing Public Education

Abstract

Credentials often fail to capture skills developed outside formal education, limiting access to opportunity. We introduce Current Skills Validation, a multi-agent architecture that decomposes skills assessment into a set of specialized agents, supporting fine-grained adaptivity grounded in learning science. We describe its architecture, foundations, and measurement agenda.

1 Introduction

Access to educational and economic opportunity depends, in part, on how skills are measured. Conventional credentials certify where and how a person learned, such as the institution attended or the course completed, and they provide only an indirect indication of what that person can currently do. For adults who have developed skills on the job, while caregiving, during military service, or through self-directed study, this distinction is consequential, because relevant

competence may go unrecognized and stall advancement (Fuller et al., 2022). We frame the challenge using our here-to-there (H2T) model, in which a learner moves from a current state (“here”) toward a goal such as a credential, a role, or course credit (“there”). Accurate assessment of a learner’s “here” is a precondition for an efficient and equitable path to “there”. This is the challenge we are tackling with this work, as illustrated in Figure 1.

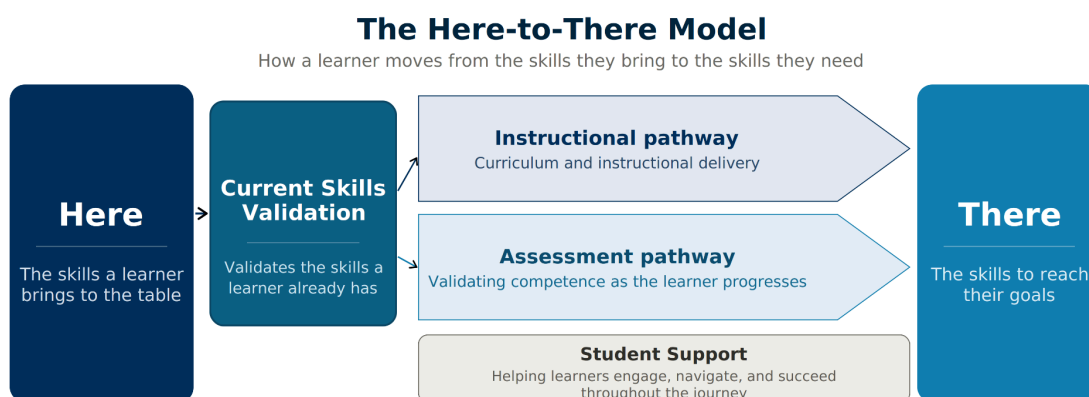


Figure 1: The here-to-there (H2T) model. The model begins with the “here”: the learner’s current skills. The learner can then use Current Skills Validation (CSV) to validate their skills, demonstrating evidence of competence to their school or employer. If the learner is seeking to grow in their skill level or expand their skills, their learning process can then be responsive to their starting state and, with additional data, be personalized to their goals, needs, and interests. The H2T model ends with the learner achieving their goal of skill acquisition. The learner can then exit this process (with the possibility of returning at a later time) or begin a new learning cycle from their updated starting point.

The Current Skills Validation (CSV) system concerns the “here”: the skills a learner currently possesses. Table 1 presents examples of when

advancement depends on validating skills not effectively captured by existing credentials.

Table 1: Personas of individuals who would benefit from the Current Skills Validation (CSV) system.

Name	Context	Skills validation supporting advancement
Carly	A medical assistant at a pharmaceutical company, trained on the job in IV insertion, infectious-disease management, and patient communication. She would like to apply to nursing school but is discouraged that her experience won't count toward the degree.	CSV validates her competence in these skills, which allows her to place out of the first two introductory courses and shorten her nursing program by approximately eight weeks.
Antonio	A director of people and talent at a large grocery chain who is seeking strong candidates for a leadership program. He is particularly interested in dispositional skills that credentials rarely capture.	Antonio generates a framework for a manager role centered on dispositional skills such as active listening. Candidates complete role-play and reflection items, and he reviews their system placements during selection.
Fatima	A full-time retail worker. She first enrolled at a state university at 18 and struggled academically before leaving. She now wants to complete a degree, but she begins with low academic confidence and a sense that expectations were never made clear. The emotional stakes are high, heightening the impact of early setbacks.	CSV makes expectations explicit from the outset and provides low-stakes, repeatable checkpoints through which Fatima can accumulate early evidence of competence. Seeing concrete progress helps her build confidence and reduces the isolation of gauging her own standing without feedback, well before any high-stakes assessment.

Several developments have begun to separate opportunity from credentials, including skills-based hiring, learning and employment records, and the expansion of apprenticeships (Goger, 2025; Fuller et al., 2022; Hills McBeth, 2025). For these approaches to succeed, experiential and non-traditional learning must be supported by validation. Recognition of prior learning, for example, has been associated with higher rates of credential completion among adult learners (Klein-Collins et al., 2020), but such recognition is defensible only when the underlying skills have been validated. Otherwise, substantial percentages of institutions may fall back on traditional proxies (e.g., the credentialing institution's reputation), re-establishing the original barrier (Sigelman et al., 2024). What is needed is validation that scales across many contexts, addresses both technical skills and dispositions, and supports the growth of skills through formative assessment as well as higher-stakes judgments through summative assessment.

This paper makes three contributions. First, we present a multi-agent architecture for skills assessment that decomposes the task into specialized agents, each of which can be inspected separately for quality. This architecture has been implemented as a software prototype that is undergoing benchmarking and iteration.

Second, we ground the design in the science of learning, mapping five item types onto the principles that motivate their use. Third, we describe the measurement framework that the architecture necessitates, including validity and reliability questions (e.g., Messick, 1995).

2 Background

Traditional standardized assessment was designed around fixed forms drawn from pre-calibrated item banks and optimized for ranking examinees on a latent trait (Lord & Novick, 1968). This paradigm is not well suited to non-traditional learners, for at least two reasons. First, it tends to privilege constructs and formats that are convenient to standardize, which often excludes the applied and dispositional competencies that matter most in work settings. Second, it generally treats assessment as a single, high-stakes event and offers limited capacity to support learning over repeated occasions.

Two lines of work loosened these constraints. Computerized adaptive testing (CAT) (Weiss & Kingsbury, 1984) improved efficiency by selecting items to match an evolving latent ability estimate; it nonetheless remains anchored to a single latent construct and to a pre-calibrated item pool that is expensive to build and maintain. Intelligent tutoring systems took a complementary path, maintaining a running

Large language models (LLMs) alter the economics of assessment. They can generate items on demand, score open-ended responses, and engage learners in interactive tasks such as role-play, making it feasible to assess applied and dispositional skills repeatedly and at scale. These same capabilities, however, complicate several assumptions on which conventional

3 The CSV Architecture

CSV coordinates a set of specialized agents under an orchestration layer, locating validation at three points: the skills framework, the generated assessment, and the evaluation of learner responses (depicted in Figure 2).

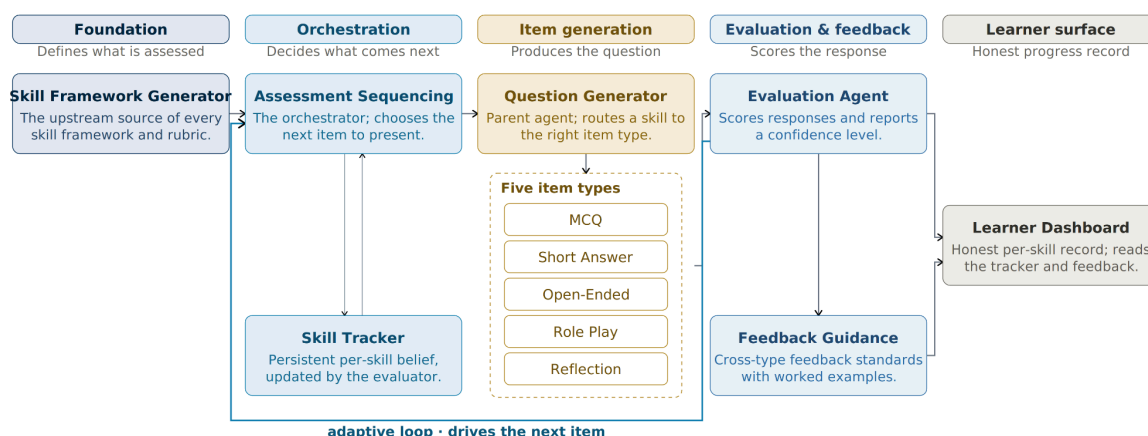


Figure 2: The Current Skills Validation (CSV) multi-agent skills-validation architecture.

The pipeline begins with a target and produces a structured skills framework that specifies the constituent competencies, their relationships, and the performance levels that define progress. The framework is generated in stages: the system first decides whether factors such as geography are relevant, then retrieves source material from O*NET (National Center for O*NET Development, n.d.) and CareerOneStop (U.S. Department of Labor, Employment and Training Administration, n.d.), and a web search over credentialing bodies. Next, it generates technical and dispositional skills, which can be edited by the user. Finally, it generates a three-level rubric for each skill. At present, the target is a job role, but the same approach extends naturally to other targets (e.g., courses), and users can upload their own frameworks. Because every subsequent item and judgment inherits the framework's claims

3.2 Item Generator

Item generation is carried out by several agents arranged in a parent-child structure. A design parent validates the requested combination of skill, Bloom's level (e.g., Remember; Bloom et al., 1956), and item type against a Bloom-by-type fit matrix. An example of this matrix is the rule that open-ended and role-play items are permitted only at the Apply level and above. The design parent then composes the skill targets, such as pairing a technical with a dispositional skill for the interactive formats. After, the design parent assembles a design brief and routes the request to a child agent specialized for a single item format. Because items are generated just-in-time, the system can tailor difficulty and content closely to the learner.

Generated items pass through three pre-delivery gates: strict schema constraints (e.g.,

exactly three multiple-choice options); deterministic content validators (e.g., every multiple-choice distractor must follow from a named misconception); and, for multiple choice, a separate LLM quality judge that screens each set for Bloom’s level mismatch.

For the interactive formats, the designer defines the prompt structure and conversation flow, while a runtime agent dynamically conducts the dialogue with the learner.

3.3 Evaluator

The evaluator agent scores learner responses against the framework using a three-level rubric (Not Competent, Approaching Competency, and Competent) and produces a calibrated confidence score for each judgment, with confidence bands tied to evidence strength. A separate evaluator configuration is used for each item type, pairing a type-specific prompt and output schema. For the interactive formats, evaluation proceeds in multiple passes: signal detection, per-skill scoring, and a holistic pass.

3.4 Orchestration and adaptivity

An orchestration layer sequences the agents and adapts the assessment to the individual learner. Sequencing is hybrid: the model first narrows the number of legal candidates using pattern signals such as consecutive Not Competent placements. The model then selects the next skill, item type, and cognitive level.

In its formative mode, CSV facilitates repeatable, low-stakes practice: it adjusts the

challenge level to keep the learner within a productive range (e.g., via scaffolding), provides feedback aligned with the science of learning, and treats each encounter as an opportunity for learning. In summative mode, CSV assembles less scaffolded evidence appropriate for higher-stakes judgments, with the shift from formative to summative occurring from a combination of system recommendation and user choice.

4 Learning Science Foundations

The agent decomposition makes principled, fine-grained adaptivity possible, and the science of learning informs how that adaptivity is used. Two commitments are central to the design. First, assessment is treated as an occasion for learning as well as for measurement, given that the act of retrieval can itself strengthen memory (Bjork & Bjork, 1992; Karpicke & Roediger, 2008). Second, the level of difficulty is calibrated to the learner rather than minimized, because difficulties that are challenging but attainable tend to produce more durable learning than does easier practice (Bjork & Bjork, 2011). In addition, repeatable checkpoints distribute practice over time, drawing on the spacing effect (Cepeda et al., 2006), and each encounter concludes with feedback intended to advance the learner’s understanding (Black & Wilam, 1998; Hattie & Timperley, 2007). Table 2 maps the five item types onto principles which motivate them.

Table 2: Item types and their learning science foundations.

Item type	Description	Sample learning science foundation
Multiple choice	Selected-response items that each evaluate one or two skills. As a collection, they provide efficient, broad coverage of knowledge and dispositions.	Dreyfus model of skill acquisition (Dreyfus, 2004): write distractors around real mistakes; trap novices, not experts; state the misconception.
Short answer	Brief scenario-based items that require the learner to respond in a few words.	Transfer of learning (Barnett & Ceci, 2002): constant skill, varied scenario.
Open-ended	Multi-turn (5–12) assessor dialogues in which a runtime agent conducts a scenario, eliciting application to a novel case.	Cognitive load theory (Sweller, 1988): working memory has limited capacity, so avoid unnecessary scenario complexity.
Role-play	Interactive, scenario-based performance of applied and dispositional skills with real-time personalization from the CSV system. Users experience role-play as a series of 10–20 turns between themselves and the agent.	Situated cognition (Lave & Wenger, 1991): knowledge and skills are tied to the contexts in which they are practiced, so role-play should situate the learner in the professional context rather than asking them to describe it from outside.

Item type	Description	Sample learning science foundation
Reflection	Prompts in which the learner articulates reasoning and assesses their own understanding, aligned with Gibbs' (1988) reflective cycle.	Metacognition (Flavell, 1979): awareness of one's cognitive processes, knowledge, and strategies. Reflective practice applies metacognition to professional experience.

The five formats serve complementary functions. Multiple-choice and short-answer items provide an efficient means of establishing breadth of knowledge, and their value depends on disciplined item construction (Haladyna et al., 2002), whereas open-ended, role-play, and reflection items are better suited to applied and dispositional competencies, and their combinations, which credentials most often omit. These latter formats draw on experiential traditions of learning (Kolb, 1984) and on the role of self-explanation and reflection in consolidating understanding (Chi et al., 1989). Because the architecture can combine formats and adapt within each, it can pursue both broad coverage and deeper evidence within a single, learner-specific sequence.

5 Strengths, Tensions, and Opportunities

5.1 Integrating formative and summative assessment

Separating validation into distinct agents allows one system to serve purposes usually divided across separate instruments. The same framework, generation, and evaluation components support adaptive formative practice while also producing the less scaffolded evidence for higher-stakes contexts. The tension is maintaining the differing evidentiary standards of the two modes within one architecture.

5.2 Measurement across every agent

Validity and reliability must now be established along the entire pipeline, from the Skills Framework Generator through the Evaluator. A defensible competency claim requires that the framework adequately represents the construct, that the generated items align with the framework, that scoring is reliable across repetitions, and that automated judgments agree with human experts at an acceptable rate. Each of these links contributes to the overall validity and reliability argument, and a weakness at any point can propagate to later stages.

Drawing from educational measurement principles, these links constitute a single validity argument, in which each inference—from construct definition, to item alignment, to score interpretation—is backed by explicit evidence and remains open to challenge (Kane, 2013; Messick, 1995). Agreement between automated and human scoring must be quantified using established evaluation criteria (Doewes & Pechenizkiy, 2021), and subgroup comparisons are needed so that neither item generation nor scoring introduces construct-irrelevant variance that would disadvantage particular groups of learners. This requirement represents one of the principal costs of a multi-agent architecture, in that the relevant standards must be satisfied at several points. This is in addition to other types of validity that may be based on the overall performance of the assessment system. For example, predictive validity (i.e., the extent to which a learner's score forecasts future performance), has been a focal point for standardized assessment (CollegeBoard, n.d.), and is a critical concern for a system like CSV which seeks to support workforce development.

5.3 Current and future development of CSV

As of October 2026, CSV has been built as a prototype and is undergoing iteration and testing. All agents in the architecture have been implemented, and refinement of the architecture has thus far centered on improving individual agent performance; for example, incorporating deterministic models trained on datasets labeled by learning science experts to evaluate and improve the quality of generated items.

We are also working to develop and disseminate our measurement framework. Our additional paper, "Off the assembly line: Applying manufacturing benchmarks to real-time assessment generation," offers one approach to testing reliability for an LLM-based evaluator agent. We plan to continue publishing works-in-progress at <https://www.wgulabs.org>.

6 Conclusion

We have introduced CSV, a multi-agent architecture that treats skills assessment as a coordinated, inspectable, and learning-grounded generating system intended to widen access to opportunity. Decomposition into specialized agents supports fine-grained adaptivity, informed by the science of learning, across five item types and two operating modes, while allowing each stage to be examined separately. The same design raises the evidentiary requirements for the system, in that validity and reliability must be established for a generating process, along an entire pipeline of agents, and across modalities. We regard these requirements as the appropriate research agenda for assessment that recognizes skills wherever they have been developed.

References

- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Erlbaum.
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world* (pp. 56–64). Worth.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. <https://escholarship.org/uc/item/3rr6q10c>
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182. [https://doi.org/10.1016/0364-0213\(89\)90002-5](https://doi.org/10.1016/0364-0213(89)90002-5)
- CollegeBoard. (n.d.). SAT validity. <https://research.collegeboard.org/reports/sat-suite/validity>
- Doewes, A., & Pechenizkiy, M. (2021). On the limitations of human-computer agreement in automated essay scoring. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 475–480). International Educational Data Mining Society. <https://files.eric.ed.gov/fulltext/ED615602.pdf>
- Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology & Society*, 24(3), 177–181. <https://doi.org/10.1177/0270467604264992>
- Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19, 243–266. <https://doi.org/10.1007/s11257-009-9063-7>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fuller, J. B., Langer, C., Nitschke, J., O’Kane, L., Sigelman, M., & Taska, B. (2022). *The emerging degree reset: How the shift to skills-based hiring holds the keys to growing the U.S. workforce at a time of talent shortage*. Burning Glass Institute. <https://www.burningglassinstitute.org/research/the-emerging-degree-reset>
- Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. Further Education Unit, Oxford Polytechnic.
- Goger, A. (2025, January 7). *There’s more to skills-based hiring than just removing degree requirements*. Brookings. <https://www.brookings.edu/articles/theres-more-to-skills-based-hiring-than-just-removing-degree-requirements/>
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hills McBeth, C. (2025, September 10). *Unlocking opportunity with WGU’s Achievement Wallet*. Western Governors University. <https://www.wgu.edu/blog/unlocking-opportunity-with-wgus-achievement-wallet2509.html>
- 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>
- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968.
<https://doi.org/10.1126/science.1152408>
- Klein-Collins, R., Taylor, J., Bishop, C., Bransberger, P., Lane, P., & Leibbrandt, S. (2020). *The PLA boost: Results from a 72-institution targeted study of prior learning assessment and adult student outcomes*. Council for Adult and Experiential Learning (CAEL); Western Interstate Commission for Higher Education (WICHE).
<https://files.eric.ed.gov/fulltext/ED608563.pdf>
- Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50(9), 741–749.
<https://doi.org/10.1037/0003-066X.50.9.741>
- National Center for O*NET Development. (n.d.). *O*NET OnLine*. U.S. Department of Labor, Employment and Training Administration.
<https://www.onetonline.org/>
- Sigelman, M., Fuller, J., & Martin, A. (2024). *Skills-based hiring: The long road from pronouncements to practice*. Burning Glass Institute & Harvard Business School.
<https://www.burningglassinstitute.org/research/skills-based-hiring-2024>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
[https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- U.S. Department of Labor, Employment and Training Administration. (n.d.). *CareerOneStop*.
<https://www.careeronestop.org/>
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375.
<https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>