

Stealth Assessments in Technology-Rich Learning Environments: A Systematic Review of Implementation, Modeling, and Psychometric Quality

Seyedahmad Rahimi¹, Salah Esmailigoujar¹, Deniz Ercan¹, & Valerie J. Shute²

1. College of Education, School of Teaching and Learning,
University of Florida, Gainesville, FL, USA

2. Anne Spencer Daves College of Education, Health, and Human Sciences,
Florida State University, Tallahassee, FL, USA

Abstract

Advances in digital technologies, data logging, and computational modeling increasingly enable unobtrusive, continuous assessment as learners interact with technology-rich environments. Stealth assessment reflects this shift by embedding evidence-based assessment within learning activities to support real-time inference, feedback, and personalization. Following PRISMA guidelines, this systematic review synthesized 161 publications representing 164 study entries published between 2004 and 2026 to examine: (1) where and how stealth assessment has been implemented across environments, populations, competencies, and purposes; and (2) how studies modeled evidence identification and accumulation in relation to psychometric quality, fidelity to evidence-centered design, and validity evidence. Results show expansion beyond games into simulations, intelligent tutoring systems, virtual reality, and applied domains, but continued concentration in K–12 contexts and cognitive or disciplinary competencies. Modeling sophistication varied widely but was not associated with stronger psychometric quality reporting or greater fidelity. Validity evidence was commonly reported, whereas reliability and fairness were rarely addressed.

Keywords: Stealth assessment; evidence-centered design; psychometric quality; technology-rich learning environments; systematic review

Introduction

Assessment plays a central role in estimating learners' knowledge, skills, and other attributes. Yet assessment has long been treated as an interruption to learning rather than an integral part of it. The traditional *teach-stop-test* cycle consumes instructional time and fragments learning processes (Shute & Ventura, 2013). This reflects a broader reliance on summative assessment, or

assessment of learning, which is typically administered after instruction for grading, ranking, and accountability purposes (Harlen & James, 1997; Shute & Rahimi, 2017). Although summative assessments can provide reliable information for high-stakes decisions, they often arrive too late to meaningfully support learning while it is occurring.

In contrast, formative assessment, or *assessment for learning*, is intended to support learning during instruction by providing actionable evidence about learners' progress (Black & Wiliam, 1998; Wiliam, 2011). Research consistently shows that timely, diagnostic feedback tied to instructional decisions can improve learning outcomes (Black & Wiliam, 2009). However, implementing formative assessment continuously and at scale remains difficult because interpreting evidence and delivering individualized feedback are labor-intensive processes, especially in diverse classrooms with limited resources (Rahimi & Shute, 2024).

These challenges are closely tied to the limitations of the one-size-fits-all model of education, where learners progress through the same instruction at the same pace despite differences in prior knowledge and learning trajectories. Personalized learning offers an alternative by adapting instruction dynamically to learners' evolving needs (Rahimi & Shute, 2025; Walkington & Bernacki, 2020). However, personalization depends on continuous, accurate estimates of learner competence during learning rather than after instruction ends.

This need aligns with foundational learning theories. Flow theory suggests learners are most engaged when task difficulty matches skill level (Csikszentmihalyi, 1990), while Vygotsky's zone of proximal development emphasizes support just beyond a learner's independent competence (Vygotsky, 1978). Both require real-time information about learner performance, which traditional assessments cannot easily provide. Stealth assessment (Shute,

2011; 2023) was developed to address this need by embedding continuous, unobtrusive assessment directly within learning processes.

Stealth Assessment is Formative and Personalized

As originally defined (Shute, 2011), stealth assessment embeds evidence-based assessment within authentic learner activities to support real-time inference and adaptive feedback. “Stealth” refers not to deception, but to the unobtrusive collection of evidence in technology-rich, formative environments. Rather than functioning as a separate test, stealth assessment captures process data (e.g., actions, decisions, strategies) during meaningful tasks such as games and simulations to infer competencies including problem solving, creativity, persistence, and collaboration. By continuously updating learner models, stealth assessment supports personalized feedback, scaffolding, and task difficulty adjustment. Its integration within interactive contexts enhances authenticity by allowing learners to demonstrate competencies through situated action across repeated observations. When communicated clearly, these assessment data can also support self-regulated learning by making progress visible and actionable. In this way, stealth assessment positions assessment as a resource for learning rather than an interruption to it.

Common Misconceptions about Stealth Assessment

Stealth assessment is often misunderstood in several ways related to its scope, conceptual meaning, design process, theoretical grounding, and methodology. First, stealth assessment is sometimes treated as limited to digital games. Although early work often focused on game-based learning environments, stealth assessment has also been implemented in simulations, intelligent tutoring systems, virtual reality environments, and open-ended digital learning environments. For example, ALUSIM used workplace simulation tasks, a skill–task Q-matrix, and a cognitive

diagnostic model to infer competency profiles (Bley, 2017). In LEONARDO, students' drawing and writing artifacts in a digital science notebook were linked to competency models through an evidence-centered design framework (Smith et al., 2016), and Rahimi, Smith et al. (2024) demonstrated stealth assessment of creativity in the programming and music-remixing environment *EarSketch*. Stealth assessment has also been applied in VR-based environments, such as virtual chemistry labs where learners' actions were interpreted as evidence of higher-order thinking skills (Nowlan et al., 2023). So, stealth assessment is not tied to a specific platform or domain; rather, it is a formative assessment approach that can be applied across disciplinary knowledge, higher-order thinking, technical skills, and socio-emotional constructs.

A second misconception concerns learner awareness. Stealth assessment is sometimes interpreted as covert or deceptive. This interpretation confuses unintrusive assessment with secrecy. As discussed earlier, the term “stealth” refers to the degree to which assessment is seamlessly embedded within a learning activity, not about hiding assessment from learners (Rahimi et al., 2026). In other words, stealth assessment is achieved when learners are not pulled out of the learning experience to respond to conventional test items; instead, their meaningful actions within the environment generate evidence of developing competencies. At the same time, learners should understand that the system is using their interactions to assess competencies for formative purposes, such as providing feedback, personalization, and learning support. Such transparency is not contrary to stealth assessment; rather, it is central to ethical and effective stealth assessment design.

Finally, stealth assessment is sometimes viewed as a retrospective process in which data are collected during gameplay and analyzed only after the experience ends. In reality, stealth assessment is designed to identify and accumulate evidence during interaction, enabling dynamic

updates to learner models and supporting feedback, adaptation, and fine-grained tracking of learning as it unfolds (Shute & Ke, 2012; Shute & Ventura, 2013). Therefore, not every embedded, automated, or game-based assessment qualifies as stealth assessment. To be considered stealth assessment, the assessment must be intentionally designed around principled links among tasks, evidence, and competencies, with evidence identification and accumulation supporting defensible inferences about learner knowledge, skills, or abilities.

Stealth Assessment is more than “Just Learning Analytics”

As data-driven approaches to education have proliferated, stealth assessment has sometimes been conflated with learning analytics or educational data mining. Learning analytics and educational data mining have made important contributions by enabling retrospective analyses of large-scale learning data to identify patterns and predict outcomes (Long & Siemens, 2011; Baker & Inventado, 2014). However, these approaches often analyze data post hoc and focus on pattern discovery rather than intentional assessment design and ongoing development.

Stealth assessment differs fundamentally in its theory-driven, design-first orientation. Rather than beginning with available data and searching for patterns, stealth assessment begins by specifying a competency model and its claims (i.e., what learners with different levels of ability can do in the future), identifying what evidence would support those claims, and designing tasks and environments to elicit that evidence (Rahimi et al., 2026). This orientation is enabled through evidence-centered design (ECD), a psychometrics-driven framework for constructing assessments around defensible validity arguments (Mislevy et al., 2003).

The ECD framework includes four core models: the Competency Model (*what is being assessed?*), Evidence Model (*what are the indicators for the competency of interest?*), Task Model (*what are the task types and features that can help elicit those indicators?*), and Assembly

Model (*how many tasks are needed and in what order to comprehensively assess the competency of interest?*) (Mislevy et al., 2003; Rahimi & Shute, 2025). These models ensure that stealth assessment systems are not ad hoc analytic pipelines but principled assessment systems grounded in theory and evidence. ECD also provides a four-process implementation architecture (i.e., evidence capture, evidence identification, evidence accumulation, and activity selection) that supports real-time inference and adaptation in dynamic learning environments (Almond et al., 2002; Rahimi et al., 2026). Without such grounding, analytics-driven approaches risk becoming opaque “black boxes” that are difficult to interpret or defend from a measurement perspective (Wise & Shaffer, 2015). Thus, while learning analytics emphasize discovering patterns in existing data, stealth assessment emphasizes designing environments so that the right data are generated to support defensible assessment inferences.

Psychometric Qualities of Well-Designed Stealth Assessments

For stealth assessment to function as a rigorous assessment methodology, it must address validity, reliability, and fairness (Rahimi & Shute, 2024) issues. These are not merely technical criteria; they are also social and ethical values because assessment inferences can shape feedback, opportunities, and educational decisions (Messick, 1994).

Validity refers to the degree to which evidence and theory support the interpretation and use of assessment results (American Educational Research Association et al., 2014).

Contemporary validity theory frames validation as an argument supported by multiple sources of evidence rather than a single statistic (Kane, 2006; Messick, 1994). For stealth assessment, this means that validity begins during design and is strengthened through iterative alignment among competency models, evidence rules, task designs, evidence identification and accumulation processes, and intended uses (Rahimi et al., 2026). In the current paper, we characterize validity

evidence as follows. Construct validity includes content, convergent, and discriminant evidence; criterion validity includes concurrent and predictive evidence; and face validity and ecological validity were recorded when those were explicitly discussed in research publications. These categories allow us to examine whether studies supported claims about construct alignment, relationships with external measures, prediction of outcomes, stakeholder credibility, and generalization to authentic contexts (American Educational Research Association et al., 2014; Kane, 2006; Messick, 1994; Suchy et al., 2024).

Reliability concerns the consistency and stability of assessment inferences, which is especially important in stealth assessment because process data can be noisy, contextual, and longitudinal. Task assembly and evidence accumulation across repeated observations are therefore central to supporting reliable inferences (Rahimi et al., 2024).

Finally, *fairness* concerns the extent to which assessment systems avoid construct-irrelevant variance that may disadvantage particular learners (Dorans & Cook, 2016; Holland & Wainer, 1993). In stealth assessment, fairness extends beyond traditional item bias to include interface design, prior gaming experience, access to technology, and other contextual factors; therefore, fairness should be addressed through deliberate design, stakeholder involvement, and validation rather than only post hoc analysis (Rahimi et al., 2023).

Beyond ensuring that these assessments are equitable for all learners, it is equally vital to understand the specific constructs being measured across such diverse digital environments. To this end, the following section details the goals and focus of our review, outlining the various competencies—from disciplinary knowledge to socio-emotional skills—that necessitate a robust classification framework.

Goals and Focus of the Current Review

To meaningfully analyze the body of stealth assessment research in this review, we began by organizing what each study assessed into coherent categories revealing field-level patterns. This step was necessary because stealth assessment studies cover a wide range of targets, ranging from disciplinary knowledge (e.g., mathematics, physics, genetics) to complex cognitive and socio-emotional competencies (e.g., problem solving, creativity, collaboration, self-regulation). Currently, no single competency framework captures such breadth. For example, 21st-century skills frameworks emphasize transferable competencies but often lack sufficient granularity for discipline-specific knowledge, while domain-centered taxonomies rarely capture interpersonal and intrapersonal competencies central to modern learning goals. Therefore, we adopted a multi-framework coding approach to classify the competencies assessed across studies (National Research Council, 2012).

Specifically, we coded each stealth assessment study's target competencies using three complementary frameworks: the National Research Council's (NRC) 21st-century competency framework, the Partnership for 21st Century Learning (P21) framework, and the World Economic Forum (WEF) *Future of Jobs* skills framework (see Appendix A in the supplementary materials for the complete explanation of each skill per framework). We used NRC as the primary organizing structure because it provides a research-based classification into cognitive, intrapersonal, and interpersonal domains (National Research Council, 2012). We additionally used P21 to align our synthesis with widely recognized educational discourse around the "4Cs" (e.g., creativity, critical thinking, communication, collaboration), and related learning and life skills, including learning and innovation skills information/media/technology skills, and life and career skills (Partnership for 21st Century Learning, 2019).

The P21 Framework defines *Core Subjects* as traditional disciplinary domains (i.e., English/reading/language arts, world languages, arts, mathematics, economics, science, geography, history, and civics/government; Partnership for 21st Century Learning, 2019). However, many stealth assessment studies in our systematic review targeted competencies that support academic learning but do not map cleanly onto any single P21 category, such as spatial skills, executive functions, fluid intelligence, and other foundational cognitive abilities. Because the P21 framework does not explicitly specify where such cognitive competencies should be classified, we expanded our operationalization of the P21 Core Subjects category to include foundational academic cognitive competencies when they were assessed as learning outcomes or as central predictors of academic performance. This decision allowed us to maintain a consistent and interpretable coding scheme across the full range of constructs assessed in stealth assessment research, while still aligning with P21's emphasis on core academic learning domains (Partnership for 21st Century Learning, 2019).

Finally, we used the WEF framework to contextualize stealth assessment research in relation to future-ready workforce competencies such as analytical thinking, creative thinking, resilience and flexibility, leadership and social influence, technology literacy, and lifelong learning (World Economic Forum, 2025). Collectively, these frameworks enabled a multi-level synthesis that supports both educational and workforce-oriented interpretations of stealth assessment research. This multi-framework approach also enabled us to compare trends across studies that targeted domain-specific learning outcomes versus domain-general competencies, and to identify underexplored competency areas in the stealth assessment literature.

The Current Study

Since its formal articulation (Shute, 2011), stealth assessment has been adopted across a growing range of domains, competencies, and technology-rich learning environments, including educational games, simulations, and immersive platforms (Rahimi & Shute, 2024). Advances in learning sciences, computational psychometrics, machine learning, multimodal analytics, and generative AI have further expanded what kinds of learner evidence can be captured and how that evidence can be modeled (Rahimi et al., 2026). As the field has grown, however, implementations have varied widely in fidelity, psychometric rigor, and interpretation. Some studies implement stealth assessment as an ECD-driven formative assessment methodology that explicitly links competencies, tasks, evidence, and inference, whereas others use the term more loosely to describe embedded analytics or post hoc modeling of interaction data. These variations, along with persistent misconceptions about stealth assessment, risk diluting the methodological integrity of the approach (Rahimi et al., 2026).

Despite growing interest, the field lacks a comprehensive synthesis of stealth assessment as an assessment methodology. Following PRISMA guidelines (Page et al., 2021), this review examines implementation trends, evidence identification and accumulation approaches, psychometric quality, validity evidence, and fidelity to ECD principles. It also aims to clarify methodological expectations across educational technology, learning sciences, learning analytics, AI in education, and assessment science. This review addresses the following research questions:

- RQ 1) What are the publication trends and implementation characteristics of stealth assessment research, including fields of study, learning environments, target populations, competencies assessed, study purposes, and assessment implementation status?*
- RQ 2) How do stealth assessment studies model evidence identification and evidence accumulation, and how are these modeling approaches related to reported psychometric*

qualities (i.e., validity, reliability, and fairness) and fidelity to stealth assessment principles?

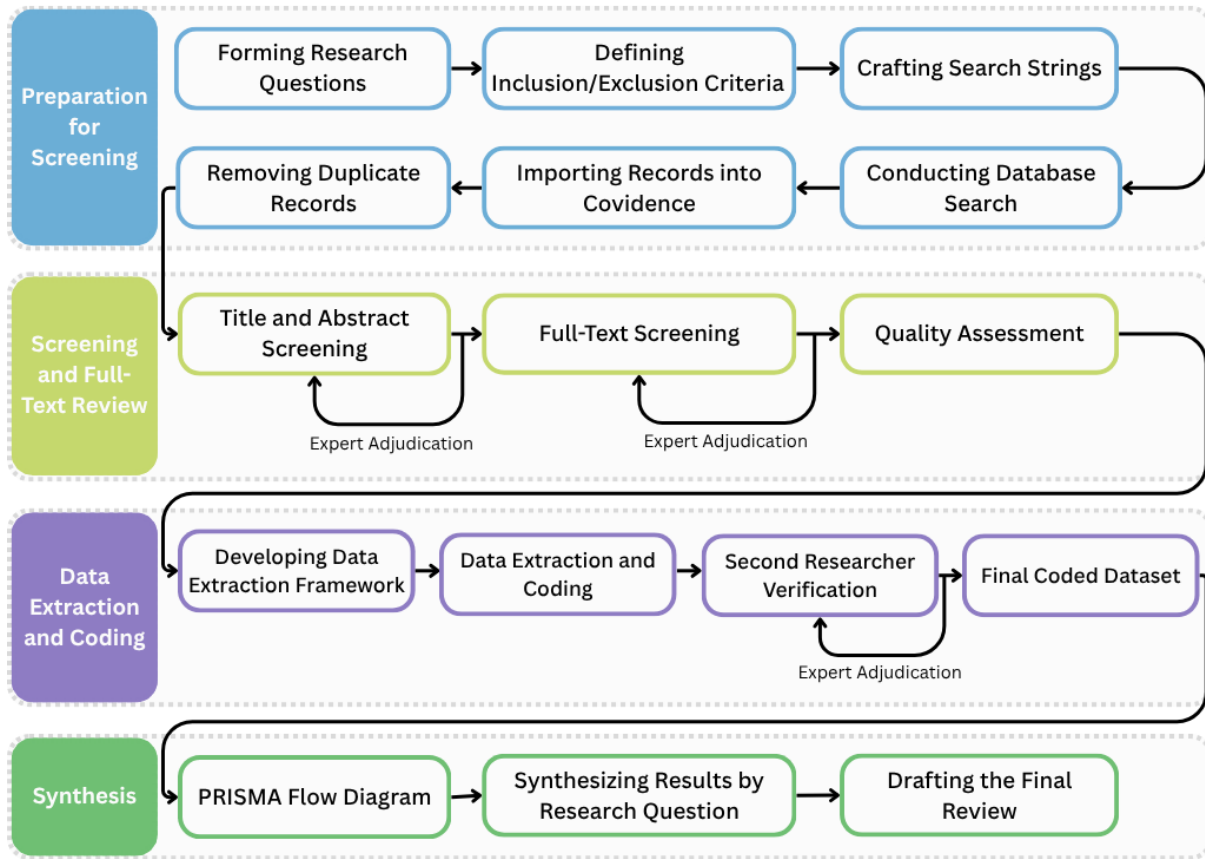
This review contributes to the literature in two ways. First, it maps where and how stealth assessment has been implemented across learning environments, populations, competencies, and purposes. Second, it introduces a conceptual framework for understanding stealth assessment pipelines through the distinction between evidence identification and evidence accumulation, as well as a continuum of modeling sophistication.

Method

This systematic review followed PRISMA guidelines (Page et al., 2021) to guide study identification, screening, eligibility assessment, and synthesis (Figure 1).

Figure 1

Overview of the Systematic Review Workflow



Search Strategy

The database search aimed to capture the broad range of terms used to describe stealth assessment approaches grounded in ECD. We used the following Boolean search string: (“stealth assessment” OR “game-based assessment” OR “embedded assessment” OR “evidence-based assessment” OR “computer-based assessment for learning” OR “evidence-centered design”) AND (“game*” OR “online learning” OR “virtual reality” OR “technology*” OR “computer supported” OR “simulation” OR “digital”). This search was applied across five databases: Scopus, Web of Science, ScienceDirect, ERIC, and Google Scholar. The search strategy was adapted for each database to account for differences in syntax and indexing rules (see Appendix B in the supplementary materials for the exact search strings used). All records were imported

into Covidence (Veritas Health Innovation, n.d.). We used Covidence to automatically remove duplicate records and handle the initial screening process.

Screening and Eligibility Assessment

Study selection followed a two-stage screening process. First, at least two researchers screened the titles and abstracts to remove clearly irrelevant records. In case of conflicts, the researchers met and attempted to resolve their conflicts. If the researchers were unable to resolve their conflicts, a third researcher (a senior researcher with stealth assessment expertise) arbitrated.

During the title and abstract screening phase, when potentially relevant studies were identified but full-text versions were not accessible through institutional subscriptions or public repositories, we contacted the corresponding authors of those studies via email and requested full-text copies. Studies for which full texts could not be obtained were excluded at the full-text screening stage. Next, full texts were reviewed to determine final eligibility. At this stage, we followed a similar process as the first stage. That is, initial review was conducted by two researchers, and a third stealth assessment expert arbitrated in cases of conflict. We employed this process on a weekly basis during the screening stage and over time, the two initial researcher assistants became aligned with the expert, leading to progressively fewer conflicts as the screening process unfolded. Table 1 summarizes the inclusion and exclusion criteria. We included English-language studies published between 2000 and 2026, beginning in 2000 to capture early work preceding the formal use of “stealth assessment” around 2004 and foundational ECD publications around 2003.

Table 1

Inclusion and Exclusion Criteria

Indicator	Inclusion Criteria	Exclusion Criteria
-----------	--------------------	--------------------

Language	English	Non-English
Publication Date	2000– May 2026 (inclusive)	Before 2000
Manuscript Type	Peer-reviewed journal articles, conference proceedings, dissertations, book chapters	Non-empirical reports, reviews, meta-analyses, books
Relevance to Stealth Assessment /ECD	Explicitly mentioned stealth assessment or ECD, or clearly implemented Stealth Assessment/ECD steps to assess Knowledge, skill, ability in a Technology rich environment.	Did not mention Stealth Assessment or ECD and did not implement related assessment practices; Used/mentioned ECD but the assessment was not stealth (i.e., learners answered questions) or in real time.
Assessment Context	Technology-rich environments (e.g., games, simulations, virtual reality, digital learning environments)	Non-digital or paper-based contexts
Conceptual Duplicates	Studies report unique implementations, tools, or empirical analyses, even if originating from the same research project.	Studies originate from the same project or conceptually redundant with another publication by the same author(s) describing the same SA tool or empirical analyses
Education Level / Target Audience	No restrictions	
Research Method	Quantitative studies using system-generated or log data (e.g., experimental, quasi-experimental, correlational, validation studies)	Qualitative perception studies or observational studies mislabeled as stealth assessment

We included peer-reviewed journal articles, conference proceedings, dissertations, and book chapters, and excluded non-empirical reports, reviews, meta-analyses, and books. Eligible studies had to explicitly mention stealth assessment or ECD, or clearly implement stealth assessment/ECD processes to assess competencies in a technology-rich environment. We

excluded studies that relied on explicit test questions rather than integrated assessment, were not conducted in real time, or used non-digital/paper-based contexts. We also prioritized quantitative studies using system-generated or log data and excluded qualitative perception or observational studies mislabeled as stealth assessment. Conceptually redundant publications from the same project were excluded unless they contributed unique tools, implementations, or analyses.

Quality Assessment

Methodological quality was assessed using an adapted Critical Appraisal Skills Programme checklist (Singh, 2013), supplemented with criteria from Pluye et al. (2009). The checklist focused on quantitative rigor, including research aims, methodological fit, recruitment, sampling justification, measurement justification, control of confounds, clarity of findings, and generalizability. Two researchers independently rated each study as high (>70%), medium (40–70%), or low (<40%) quality based on the proportion of criteria met. Discrepancies were resolved by averaging scores. All included studies met at least 40% of the criteria, so none were removed for low quality.

Data Extraction and Coding Framework

A structured data extraction protocol captured descriptive and analytical study characteristics aligned with the two research questions, including disciplinary background, assessment context, target competency, sample characteristics, modeling approaches, fidelity, validation practices, and psychometric evidence. Table 2 summarizes the extracted variables.

Table 2
Extracted Variables and Coding Dimensions

RQs	Category	Variables Extracted
-----	----------	---------------------

RQ1	Study Characteristics	Authors' field of study (e.g., CS/AI, Learning Sciences, Educational Technology, Health, Bioengineering) operationalized as the publication journal's field and authors' department, target audience sample and sample size
	Assessment Context	Designed a game/learning environment from scratch for stealth assessment or used an existing game/learning environment to integrate a stealth assessment; type (game, intelligent tutoring system, VR); name of the system
	Target Construct	Competency assessed among selected studies (e.g., problem solving, collaboration, persistence, domain-specific knowledge)
	Stealth Assessment Status	Stage of stealth assessment development: design/prototyping only; validation study; full assessment cycle; or operational/summative deployment.
RQ2	Modeling Method	Statistical modeling method (e.g., Aggregation, Bayesian networks, regression, IRT, machine learning, deep learning).
	Stealth Assessment Fidelity	Degree to which assessment is integrated in gameplay and aligned with ECD, coded as low, medium, or high based on the explicitness of competency–evidence links, use of ECD components, and presence of validation evidence.
	Validity Type	Construct validity (content, convergent, discriminant); Criterion validity (concurrent, predictive); Face and ecological validity (if explicitly discussed), Additional validation.
	Validity Metrics Reported	Methods used to establish validity, such as correlations with external measures, factor analysis, expert review, hypothesis testing, model fit indices, or ML performance metrics (e.g., accuracy, F1); explicitly coded as none if not reported.
	Psychometric Qualities (PQ) Reported	Evidence for validity, reliability, and fairness, coded individually and as a composite index (0 = no PQ; 1 = only 1 PQ; 2 = two PQs; and 3 = validity, reliability and fairness reported).

Reliability of Coding

To ensure coding reliability, a second researcher independently reviewed all extracted variables for each study. Agreement was examined at the variable level to identify discrepancies in interpretation, classification, or recording. Disagreements were resolved through structured, iterative discussion, revisiting the original texts and coding framework; unresolved cases were

adjudicated by a third expert researcher. The final consensus codes formed the authoritative dataset which helped reduce coder bias and improve consistency for complex dimensions such as fidelity and validity subtypes.

Results

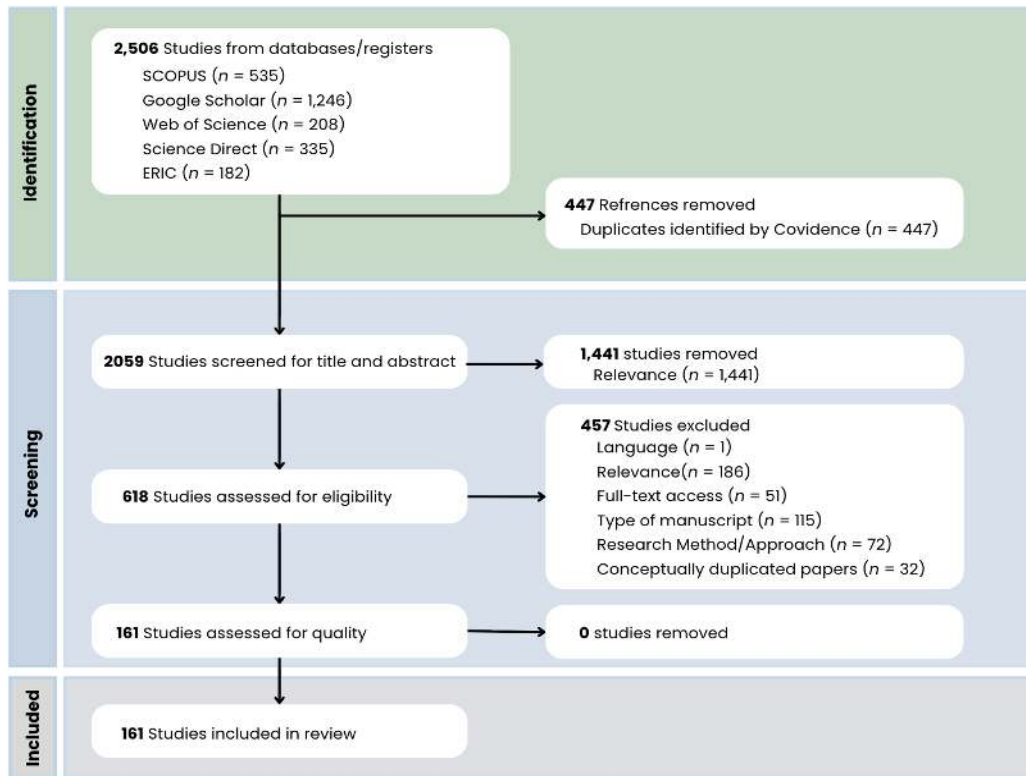
Our search focused on empirical, peer-reviewed studies published between 2000 and 2026 (inclusive) that examined stealth assessment and/or evidence-centered design in technology-rich learning environments. The studies that we included provide insight into how stealth assessment has been designed, validated, and used to support learning across contexts. In the following sections we investigate each research question, starting with the first multi-faceted research question: *What are the publication trends and implementation characteristics of stealth assessment research, including fields of study, learning environments, target populations, competencies assessed, study purposes, and assessment implementation status?*

RQ 1: Publication Trends, Contexts, and Implementation of Stealth Assessment

This systematic review included a total of 164 studies from 161 publications (see Appendix C in the supplementary materials for the complete table of included studies). Journal articles constituted the majority ($n = 83$) of source material. Conference-related publications, including proceedings and conference papers, accounted for ($n = 39$). Book chapters comprised ($n = 8$). Dissertations and theses, including PhD, doctoral, and master's theses, totaled ($n = 11$). Figure 2 shows the PRISMA flow diagram (see Figure 2) in which we documented the number of records identified, screened, excluded, and included at each stage of the review.

Figure 2

PRISMA Flow Diagram Showing the Process for Identifying and Including Relevant Studies



Note. Two publications reported multiple studies, resulting in 164 study entries from 161 sources.

Addressing RQ1, we examined publication trends and implementation characteristics, including learning environments, disciplinary fields, study purposes, target populations, implementation status, and competencies assessed across multiple frameworks.

Learning Environments and Implementation Status

To further contextualize stealth assessment implementations, we examined the types of digital learning environments in which stealth assessment was integrated, as well as whether the environment was designed ($n = 90$) from scratch specifically for stealth assessment or an existing environment was used ($n = 56$). Among studies designing new environments, stealth assessment was integrated in games ($n = 74$), Simulations ($n = 7$), Virtual Reality environments ($n = 4$), and Intelligent Tutoring Systems ($n = 5$). Among studies using existing environments ($n = 56$), stealth assessment was implemented in games ($n = 49$), Virtual Reality environments ($n = 1$), and

Intelligent Tutoring Systems ($n = 6$), with no instances in Simulations ($n = 0$). Because games were the most common learning environment, we further categorized studies by game genre.

Digital Game Genres. We categorized digital games by their dominant gameplay structure and mechanics. Because no single genre taxonomy is universally accepted, we used an integrative classification approach informed by prior genre frameworks and industry conventions, including action, adventure, role-playing, strategy, simulation-based, sandbox, serious, puzzle, and multiplayer game formats where applicable (Entertainment Software Association [ESA], 2019; Sevin & DeCamp, 2020).

Puzzle games were the most frequent genre ($n = 48$), followed by simulation games ($n = 23$), adventure games ($n = 14$), strategy games ($n = 11$), role-playing games ($n = 9$), action games ($n = 7$), sandbox games ($n = 5$), and explicitly labeled serious games ($n = 4$). Puzzle games, such as *Physics Playground*, *Portal 2*, and *Shadowspect*, typically supported stealth assessment through structured spatial, conceptual, or problem-solving challenges (Ke & Shute, 2015; Kim & Shute, 2015; Ruipérez-Valiente et al., 2022; Shute et al., 2020). Simulation games, such as *Geniventure*, *Earthquake Rebuild*, and *Operation ARA*, modeled scientific or real-world systems and generated evidence through learners' decisions, experiments, and system-level reasoning (Bounajim et al., 2020; Forsyth et al., 2020; Henderson et al., 2020; Ke & Shute, 2015; Smith, 2014). Less common genres broadened the evidentiary possibilities: adventure and role-playing games supported narrative inquiry and role-based decision making (Akram et al., 2018; Bertling et al., 2015; Gupta et al., 2021; Min et al., 2017; Reese et al., 2015; Sabourin, 2013), strategy games emphasized planning and resource management (Basu et al., 2020; Simons et al., 2021; Udeozor et al., 2021; Udeozor et al., 2024), action games captured speed, accuracy, and response patterns (Chen et al., 2020; Cui et al., 2019; Levy et al., 2016; Walsh & Bokhove, 2021), and

sandbox or serious games supported open-ended construction or purpose-built instructional assessment (Abbasi & Kazi, 2020; Caballero-Hernández et al., 2023; Chin et al., 2016; Kiili & Ketamo, 2017; McCashin et al., 2019; Westera et al., 2020).

Overall, the genre distribution shows that stealth assessment has been most often implemented in structured puzzle and simulation games, where learner actions can be clearly linked to disciplinary knowledge, reasoning, and problem-solving indicators. However, its use in adventure, strategy, role-playing, action, sandbox, and serious games shows its flexibility for capturing more open-ended, procedural, narrative, and decision-based evidence.

Simulations. Simulation-based environments ($n = 7$) represented a smaller but distinct subset of stealth assessment implementations and were typically designed to approximate real-world systems or procedural tasks. For example, *NetPASS (Networking Performance and Simulation System)* (Bauer et al., 2003) simulated network configuration tasks, enabling assessment through the analysis of learner decision sequences and troubleshooting strategies. *Energy3D* (Xie et al., 2018) provided a simulation-based platform for designing energy-efficient buildings, where stealth assessment could be embedded through monitoring design revisions and optimization behaviors. Similarly, the *Driving Simulation (Emergency Vehicle Training)* (Kleitman et al., 2022) modeled emergency response scenarios, allowing performance metrics such as route selection and reaction timing to serve as behavioral indicators. Across these environments, stealth assessment relied on capturing procedural actions within system-driven simulations that closely mirrored applied or professional contexts.

Intelligent Tutoring Systems (ITS). ITS were employed in 11 different studies and leveraged existing learner modeling infrastructures to embed stealth assessment within instructional interactions. For instance, *SimScientists* (Timms, 2012) integrated simulation-based

inquiry tasks within a tutoring framework, enabling automated inference of conceptual understanding based on response patterns and progression data. *Packet Tracer* (Rupp et al., 2012), commonly used for networking education, allowed assessment of learners' problem-solving approaches through logged configuration steps. More recently, the *Generalized Intelligent Framework for Tutoring (GIFT)* (Spain et al., 2021) provided an extensible architecture for integrating adaptive instruction and embedded assessment logic. In these ITS environments, stealth assessment was typically operationalized through continuous tracking of learner responses, hint usage, mastery updates, and step-level performance, capitalizing on the systems' existing data-driven feedback mechanisms.

Virtual reality (VR). VR environments ($n = 5$) represented the smallest proportion of implementations but offered immersive contexts with expanded behavioral data streams. For example, *AEMIN* (de-Juan-Ripoll et al., 2021) used a VR-based scenario to examine behavioral and physiological responses during gameplay. The *Carleton Virtual 3DVLE* (Nowlan, 2022) provided a virtual learning environment in which spatial navigation and interaction patterns could be analyzed. Additionally, the *Virtual Reality Organizational Environment (VROE)* (Parra Vargas et al., 2022) integrated assessment within immersive organizational simulations, allowing tracking of movement trajectories, object manipulation, and interaction timing. These VR-based implementations illustrate how stealth assessment can leverage embodied interaction data and spatial behavior metrics, although such applications remain comparatively limited within the broader literature.

Field of Study Distribution, Purpose, Target Population, and Status

To identify the disciplinary orientation of the reviewed studies, we examined the primary expertise areas of the authors, including departmental affiliations, research backgrounds, and

publication venues. We coded each study into one of four field-of-study categories: Educational Technology / Learning Sciences (Ed Tech / LS), Computer Science / Artificial Intelligence in Education (CS / AI in Ed), Applied Domains (e.g., Bioengineering, Information Engineering, Electric Engineering, and Applied Mathematics), and Psychometrics.

The largest proportion of studies were situated within Educational Technology / Learning Sciences ($n = 86$), where research typically emphasized instructional design, learning processes, and classroom implementation (e.g., Shute et al., 2013; Sabourin, 2013; Bang et al., 2023). These studies often focused on how stealth assessment could be integrated within learning environments to support formative feedback and pedagogical decision-making. A substantial number of studies were also grounded in Computer Science / AI in Education ($n = 39$), where the emphasis centered on computational modeling, machine learning approaches, and algorithmic inference (e.g., Kim et al., 2023; Ruipérez-Valiente et al., 2022; Udeozor et al., 2024). In these cases, stealth assessment was examined primarily as a technical modeling problem involving data-driven optimization and automated inference. A smaller subset of studies fell within Applied Domains ($n = 15$), including health, bioengineering, business, and organizational contexts (e.g., Westera et al., 2020; Sánchez Castro et al., 2024), where stealth assessment was implemented within domain-specific training or professional environments. Finally, a distinct group of studies were categorized as Psychometrics-related ($n = 6$), focusing on measurement validity, evidence models, and construct validation (e.g., DiCerbo, 2014; Rupp et al., 2012). These studies foregrounded the theoretical and statistical foundations of stealth assessment rather than its instructional or technical implementation. Overall, the distribution suggests that stealth assessment research has been shaped primarily by educational and computational perspectives,

with psychometric and applied-domain contributions providing complementary measurement and contextual grounding.

We categorized the reviewed studies into three groups based on their intended purpose: (1) Application of stealth assessment on knowledge and skills assessment, (2) Application of stealth assessment on hard-to-measure competencies, and (3) Automation/optimization of ECD models. The first category includes studies that assess students' knowledge or discrete skills, typically through structured or objective formats. The second category includes studies that evaluate higher-order competencies such as creativity, problem-solving, collaboration, or other complex skills. The third category includes studies that focus on improving the efficiency, personalization, or scalability of learning and teaching processes using AI (e.g., content generation, feedback automation, or system optimization).

In terms of primary purposes, 64 of the 164 studies (43.8%) focused on application of stealth assessment on hard-to-measure competencies, 52 (35.6%) on the application of stealth assessment to knowledge/skills acquisition, and 30 (20.5%) on the automation/optimization of ECD models. Additionally, 51 studies (34.9%) reported a secondary purpose, indicating that a notable portion of the literature extends beyond a single primary goal.

When primary purpose alone was considered, studies most often focused on applying stealth assessment to assess hard-to-measure competencies. However, when combining both primary and secondary purposes, the distribution reflects the broader scope of the field. Specifically, 20.8% of the studies addressed optimization/automation (either as a primary or secondary purpose), 42.6% addressed knowledge/skill assessment, and 36.5 % addressed hard-to-measure skills assessment. This combined view highlights that even when a purpose is not the main focus, it is still meaningfully represented across the studies.

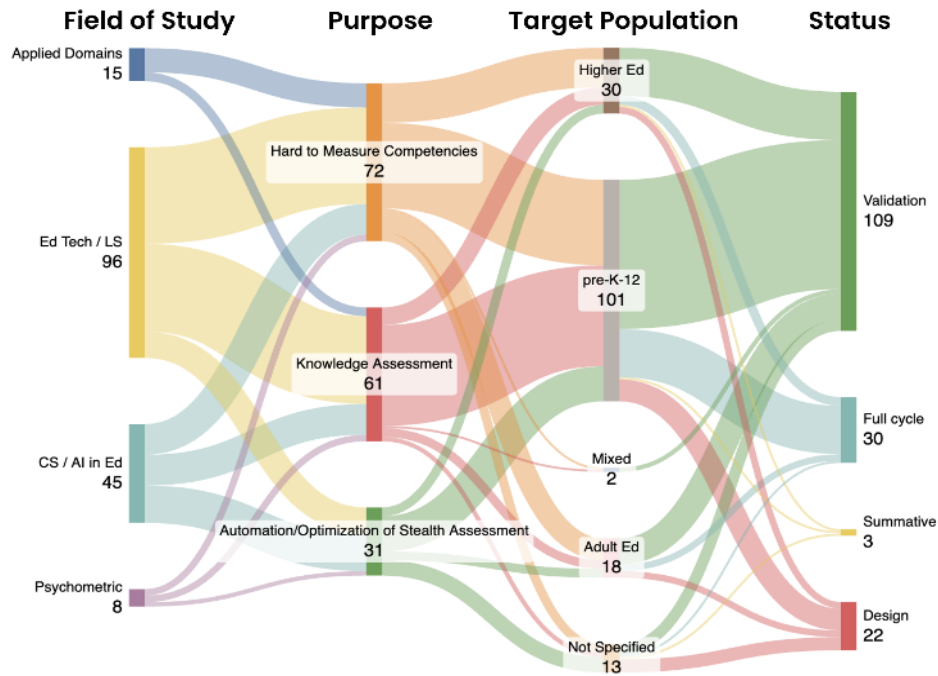
Additionally, we coded each study according to the primary target population reported in the sample description. Categories included pre-K–12, Higher Education, Adult & Continuing Education, Mixed, and Not Specified. We coded each study according to the implementation status of stealth assessment using four mutually exclusive categories: Design, Validation, Full Cycle, and Summative Assessment (see Table 3).

Table 3
Status of Included Studies

Status	Definition
Design	Studies that developed or implemented a stealth assessment framework but did not report empirical validation.
Validation	Studies that designed the assessment, collected empirical data, and conducted validation analyses (e.g., convergent or external validity evidence).
Full Cycle	Studies that designed and validated the stealth assessment and subsequently used it to support personalized learning processes such as adaptivity or feedback.
Summative Assessment	Studies that used stealth assessment solely for summative scoring purposes without providing adaptive support or formative feedback.

The Sankey diagram (see Figure 3) illustrates how 164 studies distribute across educational contexts, topical domains, and research phases. Most studies originate in pre-K–12 settings ($n = 101$), followed by higher education ($n = 30$) and adult education ($n = 18$), mixed ($n = 2$), and unspecified contexts ($n = 13$). Across all domains, the dominant research phase is validation ($n = 109$), indicating a strong emphasis on testing and evaluating stealth assessments rather than building them end-to-end. Fewer studies engage in full-cycle development ($n = 30$) or design-focused work ($n = 22$), and a small number address summative assessment ($n = 3$).

Figure 3
Flow of Studies Across Educational Levels, Research Domains, and Development Phases



Note: Ed Tech = Educational Technology, LS = Learning Sciences, CS = Computer Science, AI in Ed = Artificial Intelligence in Education, Higher Ed = Higher Education, Adult Ed = Adult Education

The flows highlight that pre-K–12 research heavily feeds into Ed Tech/LS and validation efforts, while CS/AI contributions are more evenly split between validation and full-cycle work. Overall, the figure reveals a field that is application-driven, evaluation-heavy, and centered largely on pre-K–12 educational contexts.

To characterize the competencies assessed in the reviewed studies, we examined each target competency through three complementary frameworks: the National Research Council (NRC), Partnership for 21st Century Learning (P21), and World Economic Forum (WEF) frameworks. Using these frameworks allowed us to capture both educationally grounded competency domains (e.g., cognitive, intrapersonal, and interpersonal skills) and broader future-oriented competencies relevant to workforce readiness and lifelong learning (see Appendix A in the supplementary materials for a full description of each skill per framework).

Within the NRC framework, the majority of competencies were classified under the Cognitive domain ($n = 117$), followed by Intrapersonal competencies ($n = 23$) and Interpersonal

competencies ($n = 12$). Cognitive competencies included domain-specific and higher-order areas such as *Mathematics, Physics, Chemistry, Geometric spatial reasoning, executive functioning, and creativity*. These findings indicate that stealth assessment research has predominantly focused on knowledge application, reasoning, and problem-solving constructs. In contrast, Intrapersonal competencies reflected affective and self-regulatory constructs such as *experienced stress, core psychological needs (e.g., self-esteem, distress regulation)*, and aspects of *self-regulated learning*. Interpersonal competencies were comparatively limited, which reflected constructs related to *collaboration, leadership styles, or management skills*. Overall, the NRC distribution suggests a strong cognitive emphasis, with comparatively less attention to socio-emotional and interpersonal domains.

Under the P21 framework, competencies were primarily concentrated in Core Subjects ($n = 77$) and the 4Cs ($n = 38$), followed by Life & Career Skills ($n = 30$) and ICT literacy ($n = 18$). Core Subject competencies captured disciplinary knowledge and applied STEM constructs, including *Science Education, Mathematics, weather phenomena, reading comprehension, Chinese history, and Geometry Skills*. The 4Cs category included higher-order constructs such as *creativity, critical thinking, collaboration, and communication*. Life & Career Skills included socio-emotional and adaptive constructs such as: *experienced Stress, risk taking, decision making processes, fire evacuation skills and psychological resilience*. ICT-related competencies captured digital or technological interaction skills such as *object-oriented programming skills, data literacy, and computer networks*. This distribution demonstrates that stealth assessment implementations largely target disciplinary knowledge and higher-order cognitive skills, while workforce-oriented adaptability and digital fluency constructs are present but less dominant.

Mapping to the WEF framework revealed a more selective distribution. The majority of competencies aligned with Analytical Thinking ($n = 71$). Smaller proportions were categorized under Resilience ($n = 14$), Technology Literacy ($n = 13$), Creative Thinking ($n = 13$), Leadership ($n = 13$), and Lifelong Learning ($n = 10$). Analytical Thinking captured competencies such as *problem solving*, *spatial reasoning*, and *computational thinking*, reinforcing the dominance of cognitively-oriented constructs within stealth assessment research. Resilience-aligned competencies included affective and stress-related constructs such as *experienced stress* and *psychological resilience*. Leadership (*e.g., managerial skills, collaboration skills*), Creative Thinking (*e.g. creativity, creative problem solving*), and Technology Literacy (*e.g., digital literacy, computational thinking*), and Lifelong Learning (*e.g., basic life support maneuvers, bullying awareness*) were comparatively less represented, indicating limited focus on organizational, innovation-oriented, or digital fluency competencies within the reviewed studies.

Based on these results, stealth assessment research, across all three frameworks, cognitively-oriented constructs (particularly those aligned with disciplinary knowledge and analytical reasoning) dominate the landscape. While socio-emotional constructs such as stress regulation and psychological needs are increasingly assessed, leadership, interpersonal, and broader workforce competencies remain comparatively underrepresented. This pattern suggests that stealth assessment research has achieved methodological sophistication in modeling cognitive and analytical constructs, yet its expansion into interpersonal and leadership domains remains limited. The findings therefore, highlight both the strengths and conceptual boundaries of current stealth assessment implementations in technology-rich learning environments.

We now focus on our second research question: *How do stealth assessment studies model evidence identification and evidence accumulation, and how are these modeling approaches*

related to reported psychometric qualities (i.e., validity, reliability, and fairness) and fidelity to stealth assessment principles?

RQ2: Evidence Modeling, Psychometric Quality and Fidelity, and Validity Evidence

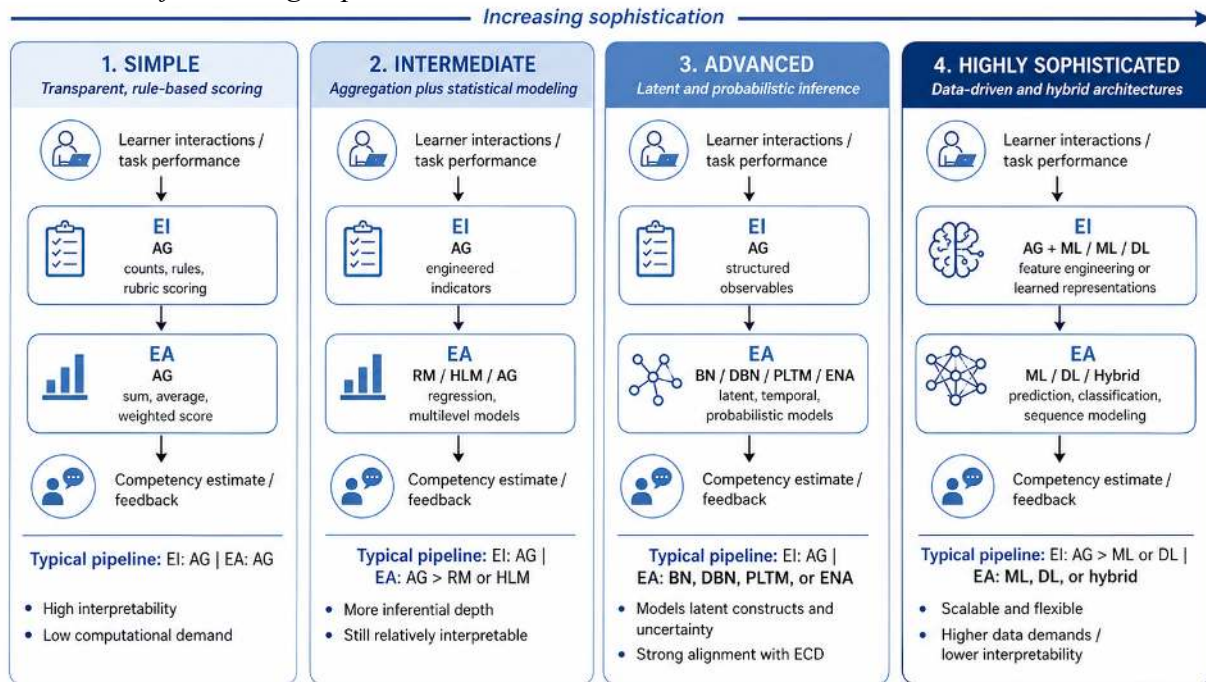
Addressing RQ2, we examined how stealth assessment studies modeled evidence and addressed psychometric rigor. Specifically, we analyzed how studies operationalized Evidence Identification (EI) and Evidence Accumulation (EA), categorized EI–EA pipelines by modeling sophistication, and examined whether these modeling approaches were associated with psychometric quality and fidelity to stealth assessment/ECD principles. We then focused on validity evidence, including the types of validity reported, the analytic approaches used to support validity claims, and the metrics used to document validity evidence.

Modeling Sophistication in Evidence Identification and Evidence Accumulation Processes

To examine how stealth assessment studies operationalize evidence, we categorized Evidence Identification (EI) and Evidence Accumulation (EA) pipelines into four levels of modeling sophistication: Simple, Intermediate, Advanced, and Highly sophisticated (Figure 4). This categorization was based on the most complex modeling technique employed within each study.

Pipelines were organized along a spectrum of modeling complexity while also accounting for pipeline depth, defined as the number of sequential modeling steps. Across the 164 studies reviewed, most explicitly separated EI and evidence accumulation EA processes ($n = 147$, 89.6%), consistent with the conceptual structure of stealth assessment and ECD. In these studies, EI and EA functioned as distinct stages: EI transformed raw learner-interaction data, such as gameplay logs, time on task, action counts, or scored behaviors, into measurable indicators, whereas EA combined or modeled those indicators to estimate latent competency states. By

contrast, a smaller set of studies used integrated pipelines ($n = 14$, 8.5%), in which raw interaction data were fed directly into a model, such as a machine-learning classifier, deep-learning model, or LLM-based scoring system, that simultaneously performed evidence extraction and prediction. A limited number used hybrid pipelines ($n = 3$, 1.8%), where EI and EA were partially distinguishable but not fully separated, such as pipelines that began with feature engineering and then passed those features into models that further transformed them while generating predictions.

Figure 4*Four Levels of Modeling Sophistication in Stealth Assessment*

Note. EI = Evidence Identification; EA = Evidence Accumulation; AG = Aggregation; RM = Regression Models; HLM = Hierarchical Linear Modeling; BN = Bayesian Network; DBN = Dynamic Bayesian Network; PLTM = Psychometric Latent Trait Models; ENA = Epistemic Network Analysis; ML = Machine Learning; DL = Deep Learning; ECD = Evidence-Centered Design.

Aggregation was by far the dominant EI approach. Most studies included aggregation somewhere in the EI process ($n = 153$, 93.3%), and in most cases aggregation was the first EI step ($n = 147$, 89.6%). This indicates that EI most often relied on rule-based, rubric-based, counting, or feature-engineering approaches to convert gameplay traces into indicators. A smaller but notable subset of studies used more computationally intensive EI approaches,

including machine learning ($n = 18$, 11.0%) and deep learning ($n = 14$, 8.5%), to identify evidence from more complex learner data, such as gameplay clusters, natural-language responses, automated speech scoring, and detector-based interpretations of gameplay behavior. Only a few studies used regression in EI ($n = 3$, 1.8%) or temporal/state-based modeling such as DBN-type approaches ($n = 1$, 0.6%).

EA showed greater methodological diversity. Although aggregation remained the most common EA approach ($n = 76$, 46.3%), many studies used more complex accumulation models, including machine learning ($n = 35$, 21.3%), Bayesian networks ($n = 33$, 20.1%), regression models ($n = 20$, 12.2%), and deep learning ($n = 18$, 11.0%). Less common EA approaches included dynamic Bayesian or temporal state models, such as DBN, BKT, or HMM ($n = 7$, 4.3%), psychometric latent trait models, such as IRT, Rasch, or DCM ($n = 7$, 4.3%), factor analytic models ($n = 4$, 2.4%), epistemic network analysis ($n = 3$, 1.8%), and hierarchical or multilevel models ($n = 2$, 1.2%). Because some studies used multi-step pipelines, these modeling categories were not mutually exclusive.

In terms of overall modeling sophistication, highly sophisticated pipelines were the largest category ($n = 58$, 35.4%), reflecting the use of machine learning, deep learning, LLMs, reinforcement learning, automated detectors, or multimodal computational models in either EI or EA. Simple pipelines, in which EI and EA were primarily aggregation-based, accounted for 50 studies (30.5%). Advanced pipelines, typically involving Bayesian networks, dynamic Bayesian models, HMMs, ENA, or psychometric latent-trait models, accounted for 38 studies (23.2%). Intermediate pipelines, usually involving aggregation-based EI followed by statistical modeling such as regression, factor analysis, or hierarchical modeling, accounted for 18 studies (11.0%).

Pipeline depth further clarifies how modeling complexity was implemented. Most studies used a one-step modeling structure per EI or EA ($n = 108$, 65.9%), while fewer used two-step pipelines ($n = 49$, 29.9%) or three-step pipelines ($n = 7$, 4.3%). Two- and three-step pipelines typically involved sequential transformations, such as aggregation followed by Bayesian modeling, machine-learning-based evidence detection followed by aggregation, or multiple modeling stages before final competency inference. These findings suggest that stealth assessment pipelines typically begin with rule-based EI, while EA has become increasingly diverse through probabilistic and AI-based models.

Simple Modeling. At the simplest level, pipelines were aggregation-oriented, with EI and EA relying primarily on rule-based or deterministic scoring approaches ($n = 50$; 30.5%). The dominant form was a fully aggregation-based pipeline, in which both EI and EA were coded as AG ($n = 45$; 27.4%). In these studies, EI typically involved transforming gameplay logs into observable indicators through counts, frequencies, durations, correctness scores, or rubric-based scoring rules. EA then combined these indicators into competency or performance scores through summation, averaging, proportions, or weighted scoring. A small number of studies in the simple category used aggregation-based EI but did not specify a distinct EA model, suggesting that evidence identification and score production were treated as part of the same rule-based scoring process.

For example, Smith et al. (2019) designed a stealth assessment within *Variant: Limits*, a calculus learning game in which students solve puzzle-based tasks focused on limits and continuity concepts as they progress through multiple levels (Zones). Learners engage with increasingly complex problems aligned with specific calculus competencies (e.g., one-sided limits, limit laws). The interactions generate gameplay data such as attempts, actions, and time

on tasks. In this context, EI is implemented through rule-based scoring rubrics that map gameplay actions to competency indicators, while EA consists of aggregating these scored indicators (e.g., ratios of points earned to points possible) to estimate students' mastery of each competency. Similarly, Kiili et al. (2015) and Kiili and Ketamo (2017) examined learning in a digital game environment involving problem-solving and decision-making tasks, where students progressed through levels by applying domain knowledge. EI involved transforming gameplay data, such as task completion, correctness, and time spent, into structured indicators, while EA aggregated these indicators into overall performance measures through summation or weighted scoring.

A typical simple pipeline therefore follows: EI: gameplay logs → rule-based indicators; EA: indicators → aggregated score. These pipelines are predominantly single-step and highly interpretable. Their transparency makes them easy to implement and communicate, but they offer limited capacity to model uncertainty, capture complex relationships among indicators, or infer latent competency states probabilistically.

Intermediate Modeling. Intermediate pipelines incorporated statistical modeling in EA, while EI typically remained aggregation-based ($n = 18$; 11.0%). In these studies, EI usually transformed raw gameplay logs into structured indicators, such as counts, scores, durations, or engineered performance features. EA then applied statistical models to estimate, explain, or validate competency-related constructs. The most common intermediate pattern was EI: aggregation; EA: regression modeling ($n = 10$). Other intermediate pipelines used factor analysis, hierarchical or multilevel modeling, or multi-step EA sequences such as aggregation followed by regression, factor analysis followed by regression, or aggregation followed by hierarchical

modeling. Most intermediate pipelines were one-step structures ($n = 12$), although some used two-step ($n = 5$) or three-step ($n = 1$) EA processes.

For example, Simons et al. (2021) operationalized gameplay evidence using aggregated performance indicators, such as mean points per turn, and then applied mixed-effects regression models to examine relationships with externally measured managerial skills. This represents an intermediate pipeline in which simple feature aggregation in EI is paired with statistical modeling in EA. Similarly, DiCerbo (2014) implemented stealth assessment within *Poptropica*, a commercial virtual world game in which players complete multi-step quests by navigating environments, collecting and using items, and solving sequential challenges. EI involved extracting and aggregating gameplay-log indicators of persistence, such as time spent on quest-related activities and number of quest events completed. EA then used confirmatory factor analysis to model persistence as a latent construct and generate factor-based estimates. Rahimi et al. (2021), in work on persistence in *Physics Playground*, also exemplify an intermediate pipeline. EI relied on aggregated log-based indicators, such as time on task and in-game trophy counts, while EA used hierarchical linear modeling to account for nested gameplay data, such as problems within students, and to capture both within- and between-student variability.

A typical intermediate pipeline therefore follows: EI: gameplay logs $\rightarrow$ aggregated or engineered indicators; EA: indicators $\rightarrow$ statistical model. These pipelines extend simple aggregation by modeling relationships among indicators, external measures, or latent constructs. They can strengthen construct interpretation and support more nuanced inference, but they are often applied analytically or post hoc rather than embedded as real-time stealth assessment engines.

Advanced Modeling. Advanced pipelines used probabilistic, temporal, network-based, or psychometric models that explicitly represented latent constructs, uncertainty, or patterned relationships among indicators ($n = 38$; 23.2%). In these studies, EI was almost always aggregation-based, but this did not necessarily mean simple counting. Rather, EI often involved theoretically guided transformations of learner-interaction data into meaningful indicators, such as action sequences, tool use, solution strategies, task success, time on task, or scored behaviors aligned with target competencies. The quality and granularity of these EI indicators were central because they formed the evidence base for more complex EA models.

EA in advanced pipelines included Bayesian networks, dynamic Bayesian or temporal state models, hidden Markov models, psychometric latent trait models such as Rasch, item response theory, or diagnostic classification models, and epistemic network analysis. The most common advanced pattern was EI: aggregation; EA: Bayesian network ($n = 18$), followed by multi-step pipelines in which aggregation fed into Bayesian networks ($n = 6$). Other advanced pipelines used aggregation followed by dynamic Bayesian or temporal models, psychometric latent trait models, or epistemic network analysis. Overall, advanced modeling studies most often used Bayesian-network-based accumulation ($n = 26$), followed by psychometric latent trait models ($n = 7$), dynamic Bayesian or temporal state models ($n = 6$), and epistemic network analysis ($n = 2$).

For example, Shute et al. (2021) used rule-based scoring of gameplay behaviors in *Physics Playground*, such as solution success, strategy use, and time on task, as EI indicators. EA then used Bayesian networks to infer students' mastery of Newtonian physics competencies. Other studies used dynamic Bayesian or Bayesian knowledge tracing approaches to update learner models over time and capture temporal dependencies in learning. Liu and Fulwider

(2026), for instance, used gameplay action categories from *Penguin Go* as EI indicators and then applied a hidden Markov model in EA to infer latent problem-solving states such as Explore, Develop, and Debug. Other studies (e.g., Cui et al., 2019; Verma et al., 2022) employed DBN-based approaches (e.g., Bayesian Knowledge Tracing) to capture temporal dependencies in learning and update learner models over time. Psychometric latent trait approaches, such as those using Rasch, item response theory, or diagnostic models, similarly used observed gameplay indicators to estimate latent abilities with psychometric scaling properties.

A typical advanced pipeline therefore follows: EI: gameplay logs → structured competency-aligned indicators; EA: indicators → probabilistic, temporal, network-based, or psychometric model. These pipelines were more likely than simple or intermediate pipelines to explicitly model uncertainty, latent states, or relationships among indicators. Most advanced pipelines were one-step structures ($n = 24$), but a substantial subset used two-step ($n = 13$) or three-step ($n = 1$) modeling sequences, such as aggregation followed by Bayesian modeling or psychometric modeling followed by another inferential layer. This layered structure reflects a deliberate separation between evidence extraction and competency inference, consistent with ECD principles, but it also requires stronger assumptions about how observed indicators relate to latent constructs and often demands substantial upfront design and validation work.

Highly Sophisticated Modeling. At the highest level of modeling sophistication, pipelines incorporated machine-learning or deep-learning approaches in either EI, EA, or both ($n = 58$; 35.4%). Despite this computational complexity, most studies in this category still preserved a conceptual separation between evidence identification and evidence accumulation ($n = 49$; 84.5%). A smaller subset used integrated pipelines ($n = 7$; 12.1%), in which feature

extraction and inference were collapsed into a single modeling process, and only a few used hybrid pipelines ($n = 2$; 3.4%), where some components were partially shared across the two stages. These patterns suggest that even highly computational approaches often retained the broader stealth assessment logic of transforming learner data into evidence before using that evidence for competency inference.

In EI, highly sophisticated pipelines often combined aggregation with computational methods. Aggregation still appeared in most studies in this category ($n = 47$), usually as feature engineering from gameplay logs, such as action counts, task performance, time-on-task, sequences of actions, or scored behaviors. However, these pipelines also incorporated machine learning ($n = 18$) and deep learning ($n = 14$) to identify richer forms of evidence from complex learner data. These approaches included clustering, dimensionality reduction, detector models, automated speech scoring, natural language processing, embeddings, and multimodal representations. Rather than relying only on predefined counts or rubrics, these methods enabled researchers to extract latent behavioral patterns, represent open-ended responses, or model complex temporal and multimodal traces.

Evidence accumulation in this category was especially diverse. Machine learning was the most common accumulation approach ($n = 35$), followed by deep learning ($n = 18$), aggregation ($n = 16$), Bayesian networks ($n = 7$), regression-based models ($n = 4$), and a small number of temporal or network-based approaches. Traditional machine learning models, such as random forests, support vector machines, k-nearest neighbors, gradient boosting, and reinforcement learning, were typically used to classify students, predict posttest performance, or estimate learning gains. Deep learning models, including recurrent neural networks, long short-term memory networks, gated recurrent units, convolutional neural networks, transformer-based

models, and large language models, were used when the assessment required modeling sequential, textual, audio, image-based, or multimodal data. In these cases, machine learning and deep learning served not only predictive functions but also representational ones, making it possible to model nonlinear patterns, temporal dependencies, and complex relationships among evidence sources.

Several studies illustrate variations of this modeling paradigm. In the *ENGAGE* game-based learning environment, where middle school students solve computational thinking tasks involving binary representations and programming-like problem solving, Akram et al. (2018) implemented a two-stage pipeline in which EI combined aggregation with machine learning by extracting n-gram features from sequences of player actions, such as tile manipulations and device interactions, and applying unsupervised clustering to identify latent problem-solving strategies. These sequence-derived representations were then passed to EA, where a long short-term memory network modeled temporal dependencies across actions and predicted posttest learning outcomes. In the same *ENGAGE* environment, Min et al. (2017) adopted a more end-to-end deep learning approach, where EI relied on raw, time-ordered action sequences encoded through one-hot vectors and combined with pretest features, while EA was performed by a long short-term memory-based stealth assessor that learned latent representations and predicted students' competency levels.

Other studies extended this logic to multimodal and multi-layered pipelines. In *Crystal Island*, a narrative-centered microbiology game in which students investigate a disease outbreak by collecting evidence, conducting tests, and writing reflections, Gupta et al. (2023a) implemented a multimodal deep-learning pipeline. EI combined structured gameplay features, such as event sequences and goal-completion vectors, with textual data from student reflections,

represented using ELMo embeddings and reduced through principal component analysis. EA then used gated recurrent unit-based neural networks within a multi-task learning framework to predict posttest performance and reflection quality, while also incorporating fairness evaluation and bias mitigation techniques. Similarly, Chen et al. (2020), using the *Raging Skies* game in which students act as storm chasers identifying weather phenomena, implemented a multi-layered pipeline. EI involved extracting behavioral features from log data, including time on task, click frequencies, retry attempts, and task-level performance, followed by feature selection using support vector machines with recursive feature elimination. EA then combined support vector machine classifiers for mastery prediction with long short-term memory networks to model temporal sequences of student behavior across multiple learning opportunities.

Recent studies show both continuity and expansion in this modeling paradigm. Rowe et al. (2026), for example, used machine-learning detectors in EI to identify computational thinking practices from gameplay, while EA used aggregation and Bayesian networks to estimate students' competencies. Like earlier *ENGAGE* studies, this pipeline used learned models to interpret gameplay traces, but it retained a clearer EI–EA separation by using detectors for evidence identification and Bayesian networks for evidence accumulation. Gupta et al. (2026) further extended the multimodal direction by combining engineered gameplay features, written-reflection embeddings, and deep-learning-based goal recognition in EI, while EA used recurrent neural models to predict posttest and concept-level outcomes. Other recent examples illustrate newer forms of highly sophisticated modeling. For example, Cao et al. (2026) used a large language model to interpret negotiation transcripts and generate ratings and feedback, while Rjiba et al. (2025) used reinforcement learning for adaptive difficulty adjustment in a collaborative game-based learning framework. Compared with earlier studies that focused

mainly on action sequences and posttest prediction, these recent studies increasingly incorporate multimodal data, language-based evidence, adaptive decision-making, and generative or reinforcement-learning models.

Highly sophisticated pipelines were often multi-step and computationally intensive. Within this category, 22 studies (37.9%) used one-step pipelines, 31 studies (53.4%) used two-step pipelines, and 5 studies (8.6%) used three-step pipelines. These designs are well suited for capturing nonlinear, temporal, multimodal, and high-dimensional patterns in large-scale learner data. However, their complexity can also make interpretation, validation, and psychometric grounding more challenging. Although these approaches expand the scalability and analytic power of stealth assessment, many still require stronger evidence of validity, reliability, fairness, and alignment with explicit competency models.

Psychometric Quality and Fidelity Across Modeling Levels

To examine psychometric quality, each study was coded for whether it reported evidence related to validity, reliability, and/or fairness. The psychometric quality score ranged from 0 to 3, where 0 indicated that none of the three qualities was reported and 3 indicated that all three were reported. Fidelity was coded as low, medium, or high based on the extent to which studies aligned with stealth assessment and ECD. Studies were coded as low fidelity when they mentioned stealth assessment or ECD but did not follow the core steps, or when they did not explicitly mention stealth assessment but cited related stealth assessment work and used some ECD-related steps. Studies were coded as medium fidelity when they mentioned stealth assessment or ECD, described ECD models, and followed the relevant steps. Studies were coded as high fidelity when they mentioned stealth assessment or ECD, followed the relevant core design steps (i.e., specified ECD models), and reported validation evidence.

Across the 164 studies, most reported at least one form of psychometric quality evidence. Specifically, 141 studies (86.0%) reported validity evidence, 57 studies (34.8%) reported reliability evidence, and only 12 studies (7.3%) reported fairness evidence. In terms of combined psychometric quality, 23 studies (14.0%) reported none of the three qualities, 84 studies (51.2%) reported only one quality, 45 studies (27.4%) reported two qualities, and 12 studies (7.3%) reported all three. Fidelity scores showed that 85 studies (51.8%) were coded as high fidelity, 49 studies (29.9%) as medium fidelity, and 30 studies (18.3%) as low fidelity. Overall, studies tended toward the medium-to-high fidelity range, but reporting remained uneven: validity evidence was common, reliability was less frequently reported, and fairness was rare.

To examine the relationship between psychometric quality reporting and fidelity, we conducted a Spearman's rho correlation because both variables were ordinal. The analysis showed a significant positive association between fidelity and reporting of psychometric quality, $\rho = .33, p < .001$, indicating that studies with higher fidelity to stealth assessment/ECD principles tended to report a greater number of psychometric qualities which was expected and desired.

We also examined psychometric quality and fidelity across the four modeling sophistication categories. As shown in Table 4, mean psychometric quality scores were similar across modeling levels, ranging from 1.24 to 1.34. A Kruskal–Wallis test showed no statistically significant difference in psychometric quality across modeling categories, $H(3) = 0.88, p = .83$. Fidelity also did not differ significantly by modeling category, $\chi^2(6) = 4.78, p = .57$, Cramér's $V = .12$. Finally, Spearman rank-order correlations showed no significant ordered trend between modeling sophistication and psychometric quality, $\rho = .02, p = .80$, or between modeling sophistication and fidelity, $\rho = .15, p = .056$. This pattern suggests that greater modeling sophistication did not correspond to stronger psychometric quality reporting or closer alignment

with stealth assessment/ECD design principles. In other words, computationally complex models were not necessarily more rigorous from a psychometric or design-fidelity standpoint, and simpler modeling approaches could still demonstrate strong assessment quality when they were theoretically grounded, transparently implemented, and supported by validity evidence. Because validity was the most frequently reported psychometric property, we next examined the types of validity evidence, analytic approaches, and reported metrics used across studies.

Table 4

Psychometric quality and fidelity by modeling sophistication category

Modeling sophistication	<i>n</i>	PQ score <i>M</i> (SD)	Validity (%)	Reliability (%)	Fairness (%)	High fidelity <i>n</i> (%)	Fidelity <i>M</i> (SD)
Simple	50	1.24 (0.94)	78	34	12	21 (42.0)	2.18 (0.80)
Intermediate	18	1.33 (0.59)	100	27.8	5.6	9 (50.0)	2.28 (0.83)
Advanced	38	1.34 (0.81)	84.2	44.7	5.3	20 (52.6)	2.39 (0.72)
Highly sophisticated	58	1.26 (0.71)	89.7	31	5.2	35 (60.3)	2.45 (0.75)
Total	164	1.28 (0.80)	86	34.8	7.3	85 (51.8)	2.34 (0.77)

Note. PQ = psychometric quality. Percent signs are omitted from table cells but indicated in column headers.

Validity Evidence: Types, Analytic Approaches, and Reported Metrics

We first examined the types of validity evidence reported across the reviewed studies. Among the 164 studies, validity evidence was most commonly reported as convergent validity ($n = 62$, 37.8%), followed by content validity ($n = 57$, 34.8%) and predictive validity ($n = 45$, 27.4%). Less frequently reported forms included concurrent validity ($n = 23$, 14.0%) and face validity ($n = 19$, 11.6%). In contrast, ecological validity ($n = 5$, 3.0%) and discriminant validity ($n = 3$, 1.8%) were rarely reported. These results show that validity evidence was concentrated around demonstrating that stealth assessment indicators aligned with intended constructs, correlated with

related external measures, or predicted learning and performance outcomes. Next, we examined the validity analysis approaches and reported metrics used across the reviewed studies (Table 5).

As shown in Table 5, the most common validity analysis approach was correlation with external measures ($n = 68$, 41.5%), followed by statistical, latent, or probabilistic modeling ($n = 40$, 24.4%) and predictive performance metrics ($n = 30$, 18.3%). These findings indicate that validity evidence was most often established through associations with external measures (e.g., convergent validity) or through modeling relationships among indicators, constructs, and outcomes (e.g., criterion or predictive validity).

Table 5
Validity Analysis Approaches and Reported Metrics Across Reviewed Studies

Validity analysis approach	<i>n</i>	% of studies	Most common reported metric
Correlation with external measures	68	41.5	Correlation coefficients; significance tests; accuracy
Statistical / latent / probabilistic modeling	40	24.4	Correlation coefficients; significance tests; model-fit or other details
Predictive performance metrics	30	18.3	Accuracy; correlation coefficients; AUC; F1-score
Machine-learning/deep-learning or cluster-based modeling	20	12.2	Accuracy; error metrics; correlation coefficients; AUC; F1-score
Other analytic approaches	26	15.9	Correlation coefficients; significance tests; kappa/agreement; other details
Not reported	25	15.2	

Note. Categories are not mutually exclusive; studies could be coded with more than one validity analysis approach. Percentages are based on the full sample of 164 studies.

The reported metrics varied by analytic approach. Correlation coefficients were most common overall ($n = 78$, 47.6%) and were especially common in studies using external-measure

validation and statistical modeling. Significance tests or statistical inference were also frequently reported ($n = 34$, 20.7%), followed by accuracy ($n = 32$, 19.5%) and other model-specific metrics or details ($n = 28$, 17.1%). Prediction-oriented and machine-learning/deep-learning studies more often reported performance metrics such as accuracy, AUC, F1-score, precision/recall, and error metrics. Overall, validity evidence was primarily established through external associations and statistical or probabilistic modeling, while prediction-based and machine-learning/deep-learning approaches were less common.

Discussion

Using PRISMA-guided procedures, we synthesized 161 publications representing 164 study entries on stealth assessment in technology-rich learning environments. Our coding captured study contexts, target competencies, learning environments, evidence identification and evidence accumulation pipelines, modeling sophistication, psychometric quality, fidelity to stealth assessment/ECD principles, and validity evidence. This discussion is organized around the major patterns emerging from the review. We first discuss the evolution of stealth assessment as an interdisciplinary methodology, then examine its role in assessing learning and hard-to-measure competencies across contexts. We then discuss the relationship between assessment design, modeling sophistication, psychometric quality, and fidelity, followed by implications for validity, reliability, fairness, reporting standards, and future research.

Stealth Assessment as an Expanding Interdisciplinary Methodology

This review shows that stealth assessment has evolved from a game-based assessment innovation into a broader methodology for integrating assessment within technology-rich learning and performance environments. Although early applications were strongly associated with digital games, the reviewed studies now include games, simulations, intelligent tutoring systems, virtual

reality, interactive texts, and other digital environments. This expansion is consistent with evidence-centered design, which was developed as a general framework for constructing assessments around principled evidentiary arguments rather than as a method tied to any single platform or domain (Mislevy et al., 2003). Stealth assessment extends this logic by embedding evidence capture, evidence identification, and evidence accumulation into authentic learner activity (Almond et al., 2002; Rahimi & Shute, 2024; Shute, 2011; Shute & Ventura, 2013).

The field has also expanded beyond a single disciplinary home. Researchers from learning sciences, educational technology, computer science, artificial intelligence in education, psychometrics, and applied professional domains (e.g., health, bioengineering, business) have taken up stealth assessment as a way to model learners' knowledge, skills, and abilities during meaningful activity. This expansion positions stealth assessment not merely as a technique for educational games but as a general assessment methodology for digital learning and performance systems (Rahimi & Shute, 2024; Shute & Rahimi, 2017). At the same time, the literature remains concentrated in feasibility, design, model development, and validation-stage work. Fewer studies describe full-cycle systems in which assessment outputs are used continuously to drive feedback, adaptation, scaffolding, or personalized learning at scale.

This pattern should be interpreted carefully. Because this review focused on scholarly publications, it may underrepresent large-scale implementations in commercial platforms, workplace learning, military training, or other applied settings that are not described in peer-reviewed outlets. Nevertheless, within the published literature, stealth assessment remains more developed as a design and validation methodology than as a widely operationalized assessment infrastructure. The next phase of research should therefore move from proof-of-concept and validation studies toward full-cycle systems that use evidence to support learners in real time.

Stealth Assessment Across the Lifespan and for Hard-to-Measure Competencies

The reviewed studies also demonstrate the relevance of stealth assessment for lifelong learning. Although many studies remain concentrated in pre-K–12 settings, the literature includes higher education, adult learning, professional training, healthcare, workplace simulations, and other applied domains. This broader use is especially important as education systems move toward lifelong learning and “60-year curriculum” models, in which individuals repeatedly reskill, upskill, and demonstrate new competencies across changing educational and professional contexts (Dede & Richards, 2020). Stealth assessment could provide part of the assessment infrastructure for such models by enabling learning systems to estimate developing competencies continuously, identify learner needs, and provide timely support.

The competencies assessed across the reviewed studies also illustrate both the promise and current boundaries of the field. Stealth assessment has been used to assess disciplinary knowledge, such as mathematics, science, reading, and computational thinking, as well as hard-to-measure competencies such as problem solving, creativity, persistence, collaboration, self-regulation, and socio-emotional or affective states. This trend aligns with the original promise of stealth assessment as a method for assessing competencies that are difficult to measure through traditional tests (Shute, 2011; Shute & Ventura, 2013; Shute & Wang, 2016). However, the literature remains more developed for cognitive, analytical, and disciplinary competencies than for interpersonal, intrapersonal, leadership, and broader workforce-oriented competencies. Future work should therefore expand more deliberately into these domains while ensuring that constructs are modeled with sufficient theoretical and psychometric rigor.

At the same time, newer studies show that stealth assessment is not used only to measure hard-to-measure competencies. It is also increasingly used to optimize adaptive learning systems,

support feedback, guide scaffolding, inform intelligent tutoring, and regulate task difficulty. This broadened purpose is promising, but it also raises the stakes for validity, reliability, and fairness. When learner models influence what support students receive, what tasks they encounter, or how teachers interpret their progress, the quality of the assessment argument becomes consequential.

Design, Modeling Sophistication, and Measurement Quality

One of the most important findings of this review is that studies varied substantially in how they identified and accumulated evidence. Most studies explicitly separated evidence identification and evidence accumulation, consistent with the architecture of stealth assessment and ECD. Evidence identification most often transformed learner-interaction data into indicators, while evidence accumulation combined those indicators into competency estimates or learner states. Although aggregation was dominant in evidence identification, evidence accumulation showed greater methodological diversity, including aggregation, regression, Bayesian networks, psychometric models, machine learning, deep learning, and other computational approaches. This pattern suggests that many studies still begin with relatively simple transformations of log data into indicators, while employing more varied approaches to combine those indicators into competency estimates and eventually student models. This distinction between identifying evidence and accumulating it is consistent with ECD-based assessment architecture, in which evidence capture, evidence identification, evidence accumulation, and activity selection play distinct roles in supporting assessment delivery (Almond et al., 2002; Mislevy et al., 2003).

The distribution of modeling approaches revealed a broad spectrum of sophistication. Simple pipelines relied primarily on aggregation. Intermediate pipelines introduced statistical models such as regression, factor analysis, or hierarchical modeling. Advanced pipelines used probabilistic, psychometric, temporal, or network-based approaches. Highly sophisticated

pipelines incorporated machine learning, deep learning, large language models, reinforcement learning, automated detectors, multimodal modeling, and other computational techniques. However, higher modeling sophistication was not statistically associated with stronger psychometric quality reporting or higher fidelity to stealth assessment/ECD principles. This finding is important because it challenges the assumption that more complex models necessarily produce better assessments. Assessment quality depends on the defensibility of the interpretation and use of scores, not simply on the complexity of the model (AERA, APA, & NCME, 2014; Kane, 2006; Messick, 1994).

Stealth assessment is, and should remain, driven by assessment design rather than by modeling sophistication alone (Rahimi & Shute, 2024; Rahimi et al., 2026). The term “stealth” directs attention to design: assessment should be integrated so seamlessly within learning activity that learners remain focused on meaningful performance rather than shifting into a conventional test-taking mode (Shute, 2011; Shute & Ventura, 2013). This design principle does not mean that assessment is hidden from learners; rather, it means that assessment is experientially unobtrusive while remaining transparent in purpose and use. A simple aggregation-based pipeline can still support rigorous assessment if the competency model is theoretically grounded, the learning environment is well designed, the tasks elicit meaningful evidence, and the evidence rules are aligned with the intended construct (Rahimi et al., 2024). Conversely, a highly sophisticated model may generate accurate predictions but still produce weak assessment inferences if the task design is poor, the evidence model is unclear, or the relationship between learner behavior and construct interpretation is underdeveloped. In stealth assessment, the quality of evidence depends first on the quality of the learning environment and task design. If the environment does not elicit

meaningful behaviors related to the target competency, no computational model can fully compensate for weak evidence.

This point is consistent with critiques of “big data” approaches in education, which caution that data-driven models require theory to determine what counts as meaningful evidence and how patterns should be interpreted (Wise & Shaffer, 2015). Stealth assessment is strongest when computational modeling is guided by an assessment argument rather than when assessment claims are reverse-engineered from available data (i.e., fully bottom-up approaches). Thus, modeling sophistication should be understood as a tool for improving inference, not as a substitute for assessment design.

This does not mean sophisticated modeling is unnecessary. Rather, modeling choices should be selected in service of the assessment argument. Simple aggregation may be appropriate when competencies are well defined, tasks are tightly aligned with those competencies, and evidence rules are transparent. Intermediate statistical models can help examine relationships among indicators, external measures, and outcomes. Advanced probabilistic or psychometric models can support inference under uncertainty and dynamic updating of learner states (Almond et al., 2015; Mislevy et al., 2003). Highly sophisticated models are especially useful for sequential, multimodal, nonlinear, and high-dimensional data. Ideally, future stealth assessment systems will combine rigorous design with appropriate modeling complexity, using sophisticated models only when they add inferential value.

The Promise and Risk of Highly Sophisticated Modeling

The highly sophisticated modeling category illustrates both the cutting edge and the methodological risks of the field. Several patterns emerged. First, there was a shift toward data-driven evidence identification, with machine learning and deep learning used not only for

prediction but also for detecting patterns in complex learner data. Second, temporal modeling became central in many studies, especially through recurrent architectures such as long short-term memory networks and gated recurrent units. Third, many pipelines incorporated multimodal evidence, including gameplay logs, text, audio, images, artifacts, dialogue, and external measures. Fourth, newer studies extended this paradigm through large language models, reinforcement learning, automated speech scoring, multimodal scoring, detector-based evidence identification, and AI-agent or adaptive scaffolding systems.

These approaches offer advantages in scalability, prediction, temporal modeling, adaptive decision-making, and analysis of complex data streams that are difficult to model manually. They can also reduce the labor of evidence identification and support analysis of open-ended or multimodal performances. However, stealth assessment research must ensure that computational sophistication does not outpace psychometric rigor. The risks are significant. Because many machine learning and deep learning models are data-driven rather than theory-driven, they may reduce interpretability and weaken alignment with psychometric frameworks such as Evidence-Centered Design (ECD). End-to-end models can obscure how learner behaviors support claims about competencies, while large language models raise additional concerns about consistency, hallucination, bias, transparency, and overreliance. These issues align with broader work on explainable AI (XAI), which emphasizes transparency, accountability, fairness, and stakeholder trust in educational systems (Khosravi et al., 2022).

Across the four modeling levels (i.e., ranging from simple to highly sophisticated pipelines), a clear trade-off emerges between transparency, theoretical grounding, and computational power. Simple pipelines offer interpretability but limited analytical capacity.

Intermediate pipelines support statistical validation but are often applied post hoc. Advanced pipelines align strongly with ECD and support inference under uncertainty. Highly sophisticated pipelines offer predictive power, scalability, and multimodal modeling, but can lack interpretability and psychometric transparency. These findings indicate that increasing modeling sophistication does not consistently correspond to stronger measurement quality and highlight a critical gap between computational modeling and measurement theory in stealth assessment research.

Validity Evidence Is Common but Narrowly Operationalized

The psychometric findings show that validity evidence was common across the reviewed studies, but unevenly distributed. Validity reporting was concentrated mainly in convergent, content, and predictive types. Far fewer studies examined whether stealth assessment scores were distinct from unrelated constructs (i.e., discriminant validity), whether the assessment appeared credible or meaningful to users and experts (i.e., face validity), or whether assessment performance reflected authentic real-world functioning (i.e., ecological validity).

This pattern suggests that validity reporting has primarily emphasized construct alignment, convergence with external measures, and prediction. These are important forms of evidence, however they do not constitute a complete validity argument. Contemporary validity theory emphasizes that validity is not established by a single statistic or isolated correlation; rather, it is built through a coherent argument linking evidence and theory to intended interpretations and uses of assessment results (AERA, APA, & NCME, 2014; Kane, 2006; Messick, 1994). For stealth assessment, this means that validity evidence should connect the competency model, evidence model, task model, evidence identification and accumulation processes, and intended feedback or adaptation decisions.

The field, therefore, needs to move from isolated validity checks toward more comprehensive validity arguments. Rather than reporting only that stealth assessment scores correlate with an external measure or predict an outcome, researchers should explain why the selected tasks elicit the target competencies, why the observables are meaningful indicators, how raw interaction data are transformed into evidence, how evidence is accumulated into competency estimates, and how those estimates are used to support learners. Providing these pieces of evidence for validity are especially important when studies use machine learning, deep learning, large language models, or other AI-based approaches. Predictive accuracy alone does not establish construct validity. A model may predict an outcome well while relying on construct-irrelevant features or opaque patterns that are difficult to interpret. Strong validity arguments are needed to show that predictions are not only accurate but also educationally meaningful and appropriate for the intended use. These arguments should begin during the design phase and continue iteratively across the life cycle of the learning environment, rather than being treated as one-time post hoc validation checks.

Reliability and Fairness as Ethical and Psychometric Requirements

Reliability and fairness were reported much less frequently than validity. This is a major concern because stealth assessment systems increasingly operate in adaptive, data-rich, and AI-mediated environments where assessment outputs may shape feedback, task selection, scaffolding, or learner opportunities. Reliability is necessary for stable inference, particularly when process data are noisy, contextual, and longitudinal. If competency estimates fluctuate because of task variation, interface familiarity, random behavior, or unstable model outputs, then the assessment cannot support valid interpretation or trustworthy adaptation (AERA, APA, & NCME, 2014).

Fairness is equally central. In technology-rich environments, fairness concerns extend beyond traditional item bias to include interface design, prior gaming experience, access to technology, language background, disability status, demographic characteristics, and differential familiarity with digital environments. Traditional measurement work has long treated fairness as central to assessment quality, particularly when assessment results are used to make or inform decisions about learners (Dorans & Cook, 2016; Holland & Wainer, 1993). These concerns are amplified in AI-mediated systems because algorithms may reproduce or intensify inequities if biased evidence, construct-irrelevant features, or uneven training data shape predictions. Recent work on AI in education similarly emphasizes fairness, accountability, transparency, and ethics in educational AI systems (Khosravi et al., 2022; Rahimi et al., 2025).

Reliability and fairness should therefore not be treated as optional reporting categories or technical checkboxes. They are ethical and methodological requirements for assessment systems that make claims about learners (Rahimi et al., 2025). This is particularly important because stealth assessment systems can directly influence learners' educational experiences. If a learner model determines what feedback learners receive, which tasks they encounter next, whether they are identified as needing support, or how their competencies are represented to teachers, then unreliable or unfair inferences can have real consequences. Future research should incorporate fairness from the design stage through participatory design, subgroup analyses, bias evaluation, documentation of mitigation strategies, and clear reporting of how assessment outputs are used. ECD supports such a design mindset, therefore, fidelity to evidence-centered design principles matter which was supported by the findings of our review.

Fidelity to ECD and the “Glass Box” Promise of Stealth Assessment

Studies with higher fidelity tended to report more psychometric qualities, suggesting that fidelity to the methodology may support stronger measurement practice. Although this association partly reflects the coding structure, it reinforces a broader conceptual point: stealth assessment is strongest when it is treated as a design methodology rather than as a label for embedded analytics.

A major advantage of ECD-grounded stealth assessment is that it can function as a “glass box” assessment system. Unlike purely data-driven models that may produce accurate predictions without a clear account of why those predictions are educationally meaningful (i.e., “black box”), stealth assessment can make the full assessment chain visible: from competency claims, to task design, to evidence rules, to evidence identification, to evidence accumulation, to feedback or adaptation. This top-down and bottom-up transparency distinguishes stealth assessment from some forms of learning analytics and educational data mining that prioritize prediction without sufficient validity evidence or construct interpretation. In a well-designed stealth assessment, explainability is not added after modeling; it is built into the architecture of the assessment system.

This glass-box quality is also what distinguishes stealth assessment from “just learning analytics.” Learning analytics and educational data mining have contributed powerful tools for discovering patterns in learning data (Baker & Inventado, 2014; Long & Siemens, 2011). However, stealth assessment begins with an assessment argument. It asks what competency is being assessed, what tasks can elicit evidence of that competency, what behaviors count as evidence, how that evidence should be accumulated, and how the resulting inference will be used. This design-first orientation protects stealth assessment from becoming an opaque analytics

pipeline. If the field loses this orientation, the term stealth assessment risks being applied to any embedded or automated scoring system, regardless of whether it has defensible links among competency, task, evidence, and inference.

Interdisciplinary Innovation and the Future of Stealth Assessment

The innovations identified in this review reinforce the need for interdisciplinary collaboration in the next phase of stealth assessment research. Across the literature, stealth assessment has moved beyond early rule-based implementations toward more automated, multimodal, adaptive, and scalable systems. Machine learning and deep learning models have been used to infer competencies from temporal patterns, behavioral sequences, and feature representations in gameplay and interaction logs (Akram et al., 2018; Chen et al., 2020; Henderson et al., 2020; Min et al., 2017, 2020; Rajendran & Prasanna, 2022). Other studies have expanded the evidentiary base of stealth assessment by integrating multimodal data sources, including written reflections, eye tracking, facial expressions, biometric signals, speech, and behavioral traces, to capture cognitive, affective, and metacognitive states that are difficult to observe through log files alone (Gupta et al., 2023b; Parra Vargas et al., 2022; Plass et al., 2013; Tadayon et al., 2018). A further line of work has connected stealth assessment to real-time feedback, scaffolding, adaptive difficulty, and intervention, using ongoing learner-model inferences to personalize support during learning (Avi Shena et al., 2019; Rahimi et al., 2023; Sabourin et al., 2013; Saleh et al., 2022; Shute et al., 2021).

Recent studies extend these innovations in several directions. Some incorporate machine-learning detectors into evidence identification while retaining more transparent evidence accumulation models, such as Bayesian networks, to estimate competencies (Rowe et al., 2026). Others use multimodal deep learning to combine gameplay logs, written reflections, and goal-

recognition signals for posttest or concept-level prediction (Gupta et al., 2026). Large language models are also beginning to enter the stealth assessment pipeline: for example, Cao et al. (2026) examined how a large language model could interpret negotiation transcripts and generate ratings and feedback, while Rahimi et al. (2026) used multimodal large language models to assess creativity in learning games. Other recent work has expanded stealth assessment toward adaptive system optimization, such as reinforcement-learning-based difficulty adjustment in collaborative game-based learning (Rjiba et al., 2025), and AI-agent systems that use evidence, decision, and feedback layers to support adaptive scaffolding (Cohn et al., 2026). These studies show that stealth assessment is becoming more technically powerful, but they also underscore the need to maintain clear links among competency claims, task design, evidence rules, model outputs, and learner support.

This collaboration is especially important with the growth of generative AI, large language models, and other computational techniques. Large language models may support evidence identification by interpreting open-ended text, dialogue, explanations, code, design artifacts, multimodal descriptions, and learner reflections. They may also support evidence accumulation by synthesizing heterogeneous evidence streams, generating formative feedback, or reasoning over learner models. However, automation alone will not guarantee assessment quality. Without strong competency models, task models, evidence models, and validation arguments, more automated systems may become more efficient without becoming more defensible.

Future stealth assessment systems are likely to become hybrid engineered systems in which different parts of evidence identification and evidence accumulation use different levels of modeling sophistication. Some indicators may be best handled through simple rule-based

aggregation; others may require regression, factor analysis, Bayesian networks, psychometric latent trait models, machine learning, deep learning, reinforcement learning, or large-language-model-based interpretation. The goal should not be to make every component maximally sophisticated, but to use the simplest defensible model for each inferential need while preserving transparency and validity. In this way, artificial intelligence can help scale principled stealth assessment design rather than replacing it.

Toward Reporting Standards for Stealth Assessment

The findings of this review suggest a need for clearer reporting standards. Future studies should explicitly report the competency model, evidence model, task model, other ECD models such as assembly models, evidence identification procedures, evidence accumulation procedures, intended use of the assessment, validation evidence, reliability evidence, fairness analyses, and whether the system was used for feedback, adaptation, summative scoring, or research only. Such reporting standards would align stealth assessment research with contemporary expectations in educational measurement, where validity, reliability, fairness, intended interpretation, and intended use are central to assessment quality (AERA, APA, & NCME, 2014; Kane, 2006; Messick, 1994).

These standards would also help readers distinguish between design proposals, simulation studies, validation studies, full-cycle implementations, and summative applications. This distinction matters because study maturity varied considerably across the literature. Some studies validated operational systems with real learners; others proposed frameworks, conducted simulations, or compared methods without validating a deployed assessment. Clear reporting would support cumulative knowledge building by making it easier to compare modeling approaches, psychometric evidence, fidelity, and implementation maturity across studies.

Such standards are especially important as stealth assessment expands into AI-supported and GenAI-enhanced environments. As systems become more complex, the need for transparency increases. Researchers should document not only model performance but also the assessment argument that justifies why model outputs can be interpreted as evidence of learning or competence. This documentation would help ensure that stealth assessment remains a psychometrically grounded methodology rather than becoming another label for automated prediction (Rahimi & Shute, 2024).

Limitations

This review has several limitations. First, it focused on scholarly publications and may therefore underrepresent stealth assessment systems implemented in commercial platforms, workplace learning, military training, or other applied settings where methods and results are not publicly reported. Second, despite efforts to retrieve full texts, some studies may have been missed because they were inaccessible, behind paywalls, or described using terminology not captured by our search strategy. Third, because stealth assessment is an evolving interdisciplinary field, some coding decisions required interpretive judgment, particularly for fidelity, evidence identification and accumulation separation, modeling sophistication, study purpose, and validity subtypes. Fourth, study maturity varied: some entries described validated empirical systems, whereas others were conceptual, simulation-based, or proof-of-concept studies. This variation should be considered when interpreting patterns related to validity, fidelity, and implementation. We mitigated interpretive bias through iterative coding and review, but future reviews may further refine these coding categories as the field develops.

Conclusion

This review synthesized 161 publications representing 164 study entries on stealth assessment from 2004 through 2026. The findings show that stealth assessment has become a broad methodology for embedding continuous and unobtrusive assessment into technology-rich learning environments, including games, simulations, intelligent tutoring systems, virtual reality, and other digital contexts. The field has demonstrated strong potential for assessing both disciplinary knowledge and complex competencies while supporting scalable, adaptive, and responsive learning systems. However, the review also shows that psychometric rigor has not kept pace with computational sophistication. Validity evidence was common, but reliability and especially fairness were reported far less often. Moreover, higher modeling sophistication was not associated with stronger psychometric quality or greater fidelity to stealth assessment and ECD principles. These findings suggest that rigorous stealth assessment depends not on model complexity alone, but on a defensible chain linking competency claims, task design, evidence rules, evidence identification, evidence accumulation, and intended use. The next generation of stealth assessment research must ensure that computational sophistication does not outpace psychometric rigor. When guided by design-first, evidence-centered, and ethically grounded practices, stealth assessment can transform assessment from an interruption of learning into a transparent, valid, reliable, fair, and supportive infrastructure for learning across the lifespan.

References

- * *Included studies in the review (see the supplementary materials for other included studies).*
- * Abbasi, S., & Kazi, H. (2020). Stealth assessment in serious games to improve OO learning outcomes. In *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)* (pp. 1–5). IEEE.
<https://doi.org/10.1109/AECT47998.2020.9194170>
- * Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). Improving stealth assessment in game-based learning with LSTM-based analytics. In *Proceedings of the*

- International Conference on Educational Data Mining*.
<https://par.nsf.gov/biblio/10100664>
- Almond, R. G., Mislevy, R. J., Steinberg, L. S., Williamson, D. M., & Yan, D. (2015). *Bayesian networks in educational assessment*. Springer.
- Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). A four-process architecture for assessment delivery, with connections to evidence-centered design. *Practical Assessment, Research & Evaluation*, 8(1), Article 13.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. *American Educational Research Association*.
- * Avi Shena, B. S., Sitohang, B., & Rukmono, S. A. (2019). Application of dynamic difficulty adjustment on evidence-centered design framework for game-based learning. In *2019 International Conference on Data and Software Engineering (ICoDSE)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICoDSE48700.2019.9092725>
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
- * Bang, H. J., Li, L., & Flynn, K. (2023). Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. *Early Childhood Education Journal*, 51, 717–732. <https://doi.org/10.1007/s10643-022-01332-3>
- * Basu, S., DiSalvo, B., Rutstein, D., Xu, Y., Roschelle, J., & Holbert, N. (2020). The role of evidence-centered design and participatory design in a playful assessment for computational thinking about data. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 985–991). Association for Computing Machinery. <https://doi.org/10.1145/3328778.3366881>
- * Bauer, M. I., Williamson, D. M., Mislevy, R. J., & Behrens, J. T. (2003). Using evidence-centered design to develop advanced simulation-based assessment and training. In A. Rossett (Ed.), *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn 2003)* (pp. 1495–1502). Association for the Advancement of Computing in Education (AACE).
- * Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015, June). Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. In *International conference on artificial intelligence in education* (pp. 545–549). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_58
- 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>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- * Bley, S. (2017). Developing and validating a technology-based diagnostic assessment using the evidence-centered game design approach: An example of intrapreneurship competence. *Empirical Research in Vocational Education and Training*, 9(1), 6. <https://doi.org/10.1186/s40461-017-0049-0>
- Borrella, I., & Ponce-Cueto, E. (2025). Design principles and impact of a learning analytics dashboard: Evidence from a randomized MOOC experiment. *Applied Sciences*, 15(21), Article 11493. <https://doi.org/10.3390/app152111493>

- * Bounajim, D., Rachmatullah, A., Boulden, D., Mott, B., Lester, J., Lord, T., Reichsman, F., Horwitz, P., Dorsey, C., & Wiebe, E. (2020). Utilizing cognitive load theory and evidence-centered design to inform the design of game-based learning environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1), 826–830. <https://doi.org/10.1177/1071181320641192>
- * Caballero Hernández, J. A., Palomo Duarte, M., Dodero Beardo, J. M., & Gašević, D. (2023). *Supporting skill assessment in learning experiences based on serious games through process mining techniques*. <https://doi.org/10.9781/ijimai.2023.05.002>
- * Cao, L., Etemadi, A., Wheeler, M., & Dede, C. (2026). Keeping the “glass box” transparent: Comparing expert and AI-generated ratings and feedback in stealth assessment for judgement-focused negotiation simulations. *Journal of Research on Technology in Education*, 58(1), 151–173. <https://doi.org/10.1080/15391523.2025.2568526>
- * Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, 30, 481–503. <https://doi.org/10.1007/s40593-020-00202-6>
- * Chin, D. B., Blair, K. P., & Schwartz, D. L. (2016). Got game? A choice-based learning assessment of data literacy and visualization skills. *Technology, Knowledge and Learning*, 21, 195–210. <https://doi.org/10.1007/s10758-016-9279-7>
- * Cohn, C., Guo, S., Rayala, S., Wang, H. D., Mohammed, N., Timalsina, U., Jain, S., Eeds, A., Deweese, M., Osborn Popp, P. J., Stanton, R., Walker, S., Ma, M., & Biswas, G. (2026). *Evidence-decision-feedback: Theory-driven adaptive scaffolding for LLM agents*. arXiv. <https://doi.org/10.48550/arXiv.2602.01415>
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper & Row.
- * Cui, Y., Chu, M.-W., & Chen, F. (2019). Analyzing student process data in game-based assessments with Bayesian knowledge tracing and dynamic Bayesian networks. *Journal of Educational Data Mining*, 11(1), 80–100.
- de Vreugd, L., van Leeuwen, A., Jansen, R., & van der Schaaf, M. (2024). Learning analytics dashboard design and evaluation to support student self-regulation of study behaviour. *Journal of Learning Analytics*, 11(3), 249–262. <https://doi.org/10.18608/jla.2024.8529>
- * de-Juan-Ripoll, C., Llanes-Jurado, J., Giglioli, I. A. C., Marín-Morales, J., & Alcañiz, M. (2021). An immersive virtual reality game for predicting risk taking through the use of implicit measures. *Applied Sciences*, 11(2), Article 825. <https://doi.org/10.3390/app11020825>
- Dede, C. J., & Richards, J. (Eds.). (2020). *The 60-year curriculum: New models for lifelong learning in the digital economy*. Routledge. <https://doi.org/10.4324/9781003013617>
- * DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society*, 17(1), 17–28.
- Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. Routledge. <https://doi.org/10.4324/9781315774527>
- Entertainment Software Association. (2019). Essential facts about the computer and video game industry. <https://www.theesa.com/resource/essential-facts-about-the-computer-and-video-game-industry/>
- * Forsyth, C. M., Graesser, A. C., & Millis, K. (2020). Predicting learning in a multi-component serious game. *Technology, Knowledge and Learning*, 25, 251–277. https://doi.org/10.1007/978-1-4939-9888-8_10

- * Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2021). Multimodal multi-task stealth assessment for reflection-enriched game-based learning. In *MAIED@AIED* (pp. 93–102).
- * Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2023a). Enhancing stealth assessment in collaborative game-based learning with multi-task learning. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (Vol. 13916, pp. 243–253). Springer. https://doi.org/10.1007/978-3-031-36272-9_25
- * Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2023b). Detecting and mitigating encoded bias in deep learning-based stealth assessment models for reflection-enriched game-based learning environments. *International Journal of Artificial Intelligence in Education*, 34, 1138–1165. <https://doi.org/10.1007/s40593-023-00379-6>
- * Gupta, A., Min, W., Carpenter, D., Azevedo, R., & Lester, J. (2026). Enhancing stealth assessment in game-based learning through goal recognition. *Journal of Research on Technology in Education*, 58(1), 53–72. <https://doi.org/10.1080/15391523.2025.2555246>
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, 4(3), 365–379. <https://doi.org/10.1080/0969594970040304>
- * Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020). Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. In *Proceedings of the International Conference on Educational Data Mining (EDM 2020)*. <https://eric.ed.gov/?id=ED607823>
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed., pp. 17–64)*. Praeger.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274.
- * Ke, F., & Shute, V. J. (2015). Design of game-based stealth assessment and learning support. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Advances in game-based learning*. Springer. https://doi.org/10.1007/978-3-319-05834-4_13
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, Article 100074.
- * Kiili, K., & Ketamo, H. (2017). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies*.
- * Kiili, K., Devlin, K., Perttula, A., Tuomi, P., & Lindstedt, A. (2015). Using video games to combine learning and assessment in mathematics education. *International Journal of Serious Games*, 2(4). <https://doi.org/10.17083/ijsg.v2i4.98>

- * Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, 87, 340-356. <https://doi.org/10.1016/j.compedu.2015.07.009>
- * Kim, Y. J., Knowles, M. A., Scianna, J., Lin, G., & Ruipérez-Valiente, J. A. (2023). Learning analytics application to examine validity and generalizability of game-based assessment for spatial reasoning. *British Journal of Educational Technology*, 54(1), 355–372.
- Kleimola, R., Hirsto, L., & Ruokamo, H. (2024). Promoting higher education students' self-regulated learning through learning analytics dashboards: A qualitative study. *Education and Information Technologies*, 30, 4959–4986.
- * Kleitman, S., Jackson, S. A., Zhang, L. M., Blanchard, M. D., Rizvandi, N. B., & Aidman, E. (2022). Applying evidence-centered design to measure psychological resilience: The development and preliminary validation of a novel simulation-based assessment methodology. *Frontiers in Psychology*, 12, 717568. <https://doi.org/10.3389/fpsyg.2021.717568>
- * Levy, L., Solomon, R., Johnson, J., Wilson, J., Lambeth, A., Gandy, M., & Liu, R. (2016). Grouches, extraverts, and jellyfish: Assessment validity and game mechanics in a gamified assessment. In *Proceedings of the DiGRA/FDG 2016 Conference*. <https://doi.org/10.26503/dl.v2016i1.779>
- * Liu, Z., & Fulwider, G. C. (2026). Modeling hidden states in learning: A hidden Markov model approach to stealth assessment for problem solving in computational thinking. *Journal of Research on Technology in Education*, 58(1), 32-52.
- Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review (Online)*.
- Masiello, I., Mohseni, Z. A., Palma, F., Nordmark, S., Augustsson, H., & Rundquist, R. (2024). A current overview of the use of learning analytics dashboards. *Education Sciences*, 14(1), 82. <https://doi.org/10.3390/educsci14010082>
- * McCashin, L., McGarvey, L., Carbonaro, M., & Yuen, C. (2019). Assessing spatial geometry through digital gameplay in a Minecraft summer camp. *Ubiquitous Learning*, 12(4), 37. [10.18848/1835-9795/CGP/v12i04/37-57](https://doi.org/10.18848/1835-9795/CGP/v12i04/37-57)
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23. <https://doi.org/10.3102/0013189X023002013>
- * Min, W., Frankosky, M. H., Mott, B. W., Wiebe, E. N., Boyer, K. E., & Lester, J. C. (2017, June). Inducing stealth assessors from game interaction data. In *International Conference on Artificial Intelligence in Education* (pp. 212-223). Cham: Springer International Publishing.
- * Min, W., Mott, B. W., Rowe, J. P., Liu, B., & Lester, J. C. (2020). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, 13(2), 312–325. <https://doi.org/10.1109/TLT.2019.2922356>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus Article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02
- National Research Council, Board on Science Education, & Committee on a Conceptual Framework for New K-12 Science Education Standards. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. national academies press.

- National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press.
<https://doi.org/10.17226/13398>
- * Nowlan, N., Arya, A., Qorbani, H. S., & Abdinejad, M. (2023). Higher-order thinking skills assessment in 3D virtual learning environments using motifs and expert data. *Computers & Education: X Reality*, 2, 100012. <https://doi.org/10.1016/j.cexr.2023.100012>
- * Nowlan, N. S. (2022). *Assessment of Higher Order Thinking Skills in Virtual Learning Environments* (Doctoral dissertation, Carleton University).
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
- * Parra Vargas, E., García Delgado, A., Torres, S. C., Carrasco-Ribelles, L. A., Marín-Morales, J., & Alcañiz, M. (2022). Virtual reality stimulation and organizational neuroscience for the assessment of empathy. *Frontiers in Psychology*, 13, Article 993162.
<https://doi.org/10.3389/fpsyg.2022.993162>
- Partnership for 21st Century Learning. (2019). *Framework for 21st century learning definitions*. Battelle for Kids.
https://static.battelleforkids.org/documents/p21/p21_framework_definitionsbfk.pdf
- * Plass, J. L., Homer, B. D., & Kinzer, C. K. (2013). Metrics in simulations and games for learning. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics* (pp. 649–689). Springer. https://doi.org/10.1007/978-1-4471-4769-5_31
- Pluye, P., Gagnon, M. P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. *International journal of nursing studies*, 46(4), 529–546.
- * Rahimi, S., Li, H., Esmaeiligoujar, S., Ercan, D., & Botelho, A. (2026). A Rubric-Guided Multimodal Approach Using High-Capacity LLMs Provides Psychometrically Sound Creativity Assessment in Learning Games. *Creativity Research Journal*, 1–18.
<https://doi.org/10.1080/10400419.2026.2638384>
- Rahimi, S., & Shute, V. J. (2021). Learning analytics dashboards in educational games. In *Visualizations and dashboards for learning analytics* (pp. 527–546). Springer International Publishing. https://doi.org/10.1007/978-3-030-81222-5_24
- Rahimi, S., Almond, R. G., Shute, V. J., & Sun, C. (2023). Getting the first and second decimals right: Psychometrics of stealth assessment. In *Games as stealth assessments* (pp. 125–153). IGI Global. <https://10.4018/979-8-3693-0568-3.ch006>
- Rahimi, S., Dede, C., Esmaeiligoujar, S., & Babae, M. (2025). Augmenting human creativity with responsible and ethical use of generative AI. *Generative Artificial Intelligence and Creativity: Precautions, perspectives, and possibilities*, 87.
- * Rahimi, S., & Shute, V. J. (2024). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. *Educational Technology Research and Development*, 72, 2417–2441. <https://doi.org/10.1007/s11423-023-10232-1>
- Rahimi, S., & Shute, V. J. (2025). Personalized learning in educational games using stealth assessment. In M. L. Bernacki, C. Walkington, A. Emery, & L. Zhang (Eds.), *Handbook of personalized learning* (Chap. 5). Routledge. <https://doi.org/10.4324/9781032719467-7>

- Rahimi, S., Almond, R. G., Ramírez-Salgado, A., Wusylko, C., Weisberg, L., Song, Y., & Wright, E. (2024). Competency model development: The backbone of successful stealth assessments. *Journal of Computer Assisted Learning*, 40(6), 2772–2789. <https://doi.org/10.1111/jcal.13025>
- Rahimi, S., Shute, V. J., & Almond, R. G. (2026). Stealth assessments in digital learning environments: Current trends, new directions, and ethical considerations. *Journal of Research on Technology in Education*, 58(1), 1–10. <https://doi.org/10.1080/15391523.2025.2587551>
- * Rahimi, S., Shute, V., & Zhang, Q. (2021). The effects of game and student characteristics on persistence in educational games: A hierarchical linear modeling approach. *International Journal of Technology in Education and Science*, 5(2), 141–165.
- * Rajendran, D., & Prasanna, S. (2022). Stealth assessment strategy in distributed systems using optimal deep learning with game-based learning. *The Journal of Supercomputing*, 78, 8285–8301. <https://doi.org/10.1007/s11227-021-04236-y>
- * Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics. *British Journal of Educational Technology*, 46(1), 98–122. <https://doi.org/10.1111/bjet.12128>
- * Rjiba, A., Cheniti-Belcadhi, L., & Kasperuniene, J. (2025). Adaptive Difficulty and Stealth Assessment in Collaborative Game-Based Learning. *Computer Applications in Engineering Education*, 33(6), e70102. <https://doi.org/10.1002/cae.70102>.
- * Rowe, E., Almond, R. G., & Almeda, M. V. (2026). Validating game-based learning assessment of students' computational thinking practices using Bayesian networks and machine-learning based detectors. *Journal of Research on Technology in Education*, 58(1), 14–31. <https://doi.org/10.1080/15391523.2025.2590018>
- * Ruipérez-Valiente, J. A., Kim, Y. J., Baker, R. S., Martínez, P. A., & Lin, G. C. (2022). The affordances of multivariate elo-based learner modeling in game-based assessment. *IEEE transactions on learning technologies*, 16(2), 152–165. 10.1109/TLT.2022.3203912
- * Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, 4(1), 49–110. <https://doi.org/10.5281/zenodo.3554643>
- * Sabourin, J. L. (2013). *Stealth assessment of self-regulated learning in game-based learning environments* (Doctoral dissertation, North Carolina State University).
- * Saleh, A., Phillips, T., Glazewski, K. D., Hmelo-Silver, C. E., Mott, B., & Lester, J. (2022). Designing an adaptive collaborative problem-solving system with expanded evidence-centered design. In A. Weinberger, W. Chen, D. Hernández-Leo, & B. Chen (Eds.), *Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning—CSCL 2022* (pp. 328–331). International Society of the Learning Sciences.
- * Sánchez Castro, S., Pascual Sevillano, M. Á., & Fombona Cadavieco, J. (2024). Learning analytics in serious games as predictors of linguistic competence in students at risk. *Technology, Knowledge and Learning*, 29(3), 1551–1577. <https://doi.org/10.1007/s10758-024-09726-5>
- Sevin, R., & DeCamp, W. (2020). Video game genres and advancing quantitative video game research with the genre diversity score. *The Computer Games Journal*, 9(4), 401–420. DOI: 10.1007/s40869-020-00115-3

- Shute, V. J. (2023). The history of stealth assessment and a peek into its future. In *Games as stealth assessments* (pp. 1-23). IGI Global.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Information Age Publishers.
- Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 43-58). New York, NY: Springer New York.
- Shute, V. J., and Rahimi, S. (2017) Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33: 1–19. doi: 10.1111/jcal.12172.
- * Shute, V. J., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. In *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 535–562). Wiley.
<https://doi.org/10.1002/9781118956588.ch22>
- * Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R. G., Dai, C. P., & Liu, Z. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in educational games. *Journal of Computer Assisted Learning*, 37(1), 127–141. <https://doi.org/10.1111/jcal.12473>
- * Shute, V. J., Smith, G., Kuba, R., Dai, C. P., Rahimi, S., & Sun, C. (2021). The design, development, and testing of learning supports for the Physics Playground game. *International Journal of Artificial Intelligence in Education*, 31, 357–379.
<https://doi.org/10.1007/s40593-020-00196-1>
- * Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton’s Playground. *The Journal of Educational Research*, 106(6), 423–430.
- Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in video games: Stealth assessment*. MIT Press.
- * Simons, A., Wohlgenannt, I., Weinmann, M., & Fleischer, S. (2021). Good gamers, good managers? A proof-of-concept study with Sid Meier’s Civilization. *Review of Managerial Science*, 15, 957–990. <https://doi.org/10.1007/s11846-020-00378-0>
- Singh, J. (2013). Critical appraisal skills programme. *Journal of Pharmacology and Pharmacotherapeutics*, 4(1), 76-77. <https://doi.org/10.4103/0976-500X.107697>
- * Smith, A., Aksit, O., Min, W., Wiebe, E., Mott, B. W., & Lester, J. C. (2016, June). Integrating real-time drawing and writing diagnostic models: an evidence-centered design framework for multimodal science assessment. In *International Conference on Intelligent Tutoring Systems* (pp. 165-175). Cham: Springer International Publishing.
- * Smith, D. (2014). *Earthquake Rebuild: A game for the stealth learning of middle school math* (Master's thesis, The Florida State University).
- * Smith, G., Shute, V., & Muenzenberger, A. (2019). Designing and validating a stealth assessment for calculus competencies. *Journal of Applied Testing Technology*, 52-59.
- * Spain, R., Min, W., Saville, J., Brawner, K., Mott, B., & Lester, J. (2021). Automated assessment of teamwork competencies using evidence-centered design-based natural language processing approach. In *Proceedings of the Ninth Annual GIFT Users Symposium (GIFTSym9)* (p. 140). U.S. Army DEVCOM Soldier Center.
- Suchy, Y., DesRuisseaux, L. A., Gereau Mora, M., Brothers, S. L., & Niermeyer, M. A. (2024). Conceptualization of the term “ecological validity” in neuropsychological research on

- executive function assessment: A systematic review and call to action. *Journal of the International Neuropsychological Society*, 30(5), 499–522.
<https://doi.org/10.1017/S1355617723000735>
- Sweet, S. J., & Rupp, A. A. (2012). Using the ECD framework to support evidentiary reasoning in the context of a simulation study for detecting learner differences in epistemic games. *Journal of Educational Data Mining*, 4(1), 183–223.
<https://doi.org/10.5281/zenodo.3554649>
- * Tadayon, R., Amresh, A., McDaniel, T., & Panchanathan, S. (2018). Real-time stealth intervention for motor learning using player flow-state. In *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1–8). IEEE.
<https://doi.org/10.1109/SeGAH.2018.8401360>
- * Timms, M. J. (2012). *Session N—SimScientists: An example of how technology can support differentiated instruction in the classroom*. ACER Research Conference.
- * Udeozor, C., Abegão, F. R., & Glassey, J. (2024). Measuring learning in digital games: Applying a game-based assessment framework. *British Journal of Educational Technology*, 55(2), 957–991. <https://doi.org/10.1111/bjet.13407>
- * Udeozor, C., Russo, A., & Glassey, J. (2021). Exploring log data for behaviour and solution pattern analysis in a serious game. In *Gamification and social networks in education*. Veritas Health Innovation. (n.d.). *Covidence systematic review software*. Retrieved from <https://www.covidence.org/>
- * Verma, V., Craig, S. D., Levy, R., Bansal, A., & Amresh, A. (2022). Domain Knowledge and Adaptive Serious Games: Exploring the Relationship of Learner Ability and Affect Adaptability. *Journal of Educational Computing Research*, 60(2), 406–432.
<https://doi.org/10.1177/07356331211031287>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advances, and future directions. *Journal of Research on Educational Effectiveness*, 13(2), 235–261. DOI: 10.1080/15391523.2020.1747757
- * Walsh, C., & Bokhove, C. (2021). Targeting data collection in games based assessment. *Computers and Education Open*, 2, 100054. <https://doi.org/10.1016/j.caeo.2021.100054>
- * Westera, W., Prada, R., Mascarenhas, S., Santos, P. A., Dias, J., & Guimarães, R. (2020). Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies*, 25, 351–380. <https://doi.org/10.1007/s10639-019-09968-2>
- Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press
- Wise, A. F., & Shaffer, D. W. (2015). Why theory matters more than ever in the age of big data. *Journal of Learning Analytics*, 2(2), 5–13. <https://doi.org/10.18608/jla.2015.22.2>
- World Economic Forum. (2025). *The future of jobs report 2025*. World Economic Forum.
<https://www.weforum.org/publications/the-future-of-jobs-report-2025/>
- * Xie, C., Schimpf, C., Chao, J., Nourian, S., & Massicotte, J. (2018). Learning and teaching engineering design through modeling and simulation on a CAD platform. *Computer Applications in Engineering Education*, 26(4), 824–840.
<https://doi.org/10.1002/cae.21920>
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Brief Author Biographies

Seyedahmad Rahimi

Seyedahmad Rahimi, Ph.D., is an Assistant Professor of Educational Technology in the School of Teaching and Learning, College of Education, at the University of Florida. His research lies at the intersection of learning sciences, artificial intelligence in education, educational technology, and computational psychometrics.

Salah Esmaeiligoujar

Salah Esmaeiligoujar is a doctoral student in the School of Teaching and Learning, College of Education, at the University of Florida.

Deniz Ercan

Deniz Ercan is a doctoral candidate in the School of Teaching and Learning, College of Education, at the University of Florida.

Valerie J. Shute

Valerie J. Shute, Ph.D., is Professor Emerita at Florida State University in the Anne Spencer Daves College of Education, Health, and Human Sciences. Her research focuses on assessment, cognitive diagnosis, learning from advanced instructional systems, educational games, and stealth assessment.

Acknowledgments

The authors would like to thank Rouhollah Khodabandelou, Renata Kuba, and Maryam Babaei for their help during the initial screening phase.

Funding Statement

The authors received no external funding for this work.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References for Included Studies in the Review

* *Included studies in the review*

- * Abbasi, S., & Kazi, H. (2020). Stealth assessment in serious games to improve OO learning outcomes. In *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)* (pp. 1–5). IEEE.
<https://doi.org/10.1109/AECT47998.2020.9194170>
- * Akram, B. (2019). *Assessment of students' computer science focal knowledge, skills, and abilities in game-based learning environments* (Doctoral dissertation, North Carolina State University).
- * Akram, B., Azizolsoltani, H., Min, W., Wiebe, E. N., Navied, A., Mott, B. W., ... & Lester, J. C. (2020). A Data-Driven Approach to Automatically Assessing Concept-Level CS Competencies Based on Student Programs. In *CSEDM@ EDM*.
- * Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2019). Improving stealth assessment in game-based learning with LSTM-based analytics. In *Proceedings of the International Conference on Educational Data Mining*.
<https://par.nsf.gov/biblio/10100664>
- * Al-Smadi, M., Capuano, N., & Guetl, C. (2018). Decoupling assessment and serious games to support guided exploratory learning in smart education. *Journal of Ambient Intelligence and Humanized Computing*, 9, 497–511. <https://doi.org/10.1007/s12652-016-0436-6>
- * Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2019). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*, 36(3), 350–358.
<https://doi.org/10.1111/jcal.12405>
- * Amorim, A. N., Jeon, L., Abel, Y., Pape, S., Albuquerque, E. X. S., Soares, M., Silva, V. C., Aguiar, D., Oliveira Neto, J. R., Costin, C., Rodrigues, R. L., Leon, M., de Paula, C. A., Lopes, J., Silva, M. S., do Nascimento, M. V., Patricio, G. A., da Silva, V. F., & Florentino, R. (2023). Exploring the use of Escribo Play mobile learning games to foster early mathematics for low-income first-grade children. *Computers & Education*, 199, Article 104759. <https://doi.org/10.1016/j.compedu.2023.104759>
- * Antoniou, P. E., Siountas, A., Zilidou, V. I., & Bamidis, P. D. (2017). Virtual scenarios for stealth assessment of the elderly: Perceptions and acceptance of technology-based health and wellness interventions. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 696–701). IEEE.
<https://doi.org/10.1109/CBMS.2017.115>
- * Avi Shena, B. S., Sitohang, B., & Rukmono, S. A. (2019). Application of dynamic difficulty adjustment on evidence-centered design framework for game-based learning. In *2019 International Conference on Data and Software Engineering (ICoDSE)* (pp. 1–6). IEEE.
<https://doi.org/10.1109/ICoDSE48700.2019.9092725>

- * Bang, H. J., Li, L., & Flynn, K. (2023). Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. *Early Childhood Education Journal*, 51, 717–732. <https://doi.org/10.1007/s10643-022-01332-3>
- * Baron, T., & Amresh, A. (2015). Word towers: Assessing domain knowledge with non-traditional genres. In *Proceedings of the European Conference on Games-Based Learning* (Vol. 638). Academic Conferences International Limited.
- * Basu, S., DiSalvo, B., Rutstein, D., Xu, Y., Roschelle, J., & Holbert, N. (2020). The role of evidence-centered design and participatory design in a playful assessment for computational thinking about data. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 985–991). Association for Computing Machinery. <https://doi.org/10.1145/3328778.3366881>
- * Bauer, M. I., Williamson, D. M., Mislevy, R. J., & Behrens, J. T. (2003). Using evidence-centered design to develop advanced simulation-based assessment and training. In A. Rossett (Ed.), *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn 2003)* (pp. 1495–1502). Association for the Advancement of Computing in Education (AACE). <https://www.learntechlib.org/p/12157/>
- * Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015). Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial intelligence in education* (pp. 545–549). Springer.
- * Bley, S. (2017). Developing and validating a technology-based diagnostic assessment using the evidence-centered game design approach: An example of intrapreneurship competence. *Empirical Research in Vocational Education and Training*, 9(1), 6. [10.1186/s40461-017-0049-0](https://doi.org/10.1186/s40461-017-0049-0)
- * Bounajim, D., Rachmatullah, A., Boulden, D., Mott, B., Lester, J., Lord, T., Reichsman, F., Horwitz, P., Dorsey, C., & Wiebe, E. (2020). Utilizing cognitive load theory and evidence-centered design to inform the design of game-based learning environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1), 826–830. <https://doi.org/10.1177/1071181320641192>
- * Caballero Hernández, J. A., Palomo Duarte, M., Dodero Beardo, J. M., & Gašević, D. (2023). *Supporting skill assessment in learning experiences based on serious games through process mining techniques*. <https://doi.org/10.9781/ijimai.2023.05.002>
- * Calvo-Morata, A., Alonso-Fernández, C., Freire-Morán, M., Martínez-Ortiz, I., & Fernández-Manjón, B. (2019). Game learning analytics: Facilitating the use of serious games in the class. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, 14(4), 168–176. <https://doi.org/10.1109/RITA.2019.2952296>

- * Capuano, N., & King, R. (2015). Knowledge-based assessment in serious games: An experience on emergency training. *Journal of e-Learning and Knowledge Society*, 11(3). Retrieved April 27, 2026 from <https://www.learntechlib.org/p/151927/>
- * Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, 30, 481–503. <https://doi.org/10.1007/s40593-020-00202-6>
- * Chin, D. B., Blair, K. P., & Schwartz, D. L. (2016). Got game? A choice-based learning assessment of data literacy and visualization skills. *Technology, Knowledge and Learning*, 21, 195–210. <https://doi.org/10.1007/s10758-016-9279-7>
- * Christhlf, K. (2023). *Stealth Assessment of Reading Comprehension Skill and Vocabulary Knowledge using Game Performance* (Master's thesis, Arizona State University).
- * Chu, M. W., & Chiang, A. (2018). *Raging skies: Development of a digital game-based science assessment using evidence-centred game design*.
- * Cui, Y., Chu, M.-W., & Chen, F. (2019). Analyzing student process data in game-based assessments with Bayesian knowledge tracing and dynamic Bayesian networks. *Journal of Educational Data Mining*, 11(1), 80–100.
- * de Heer, J., & Porskamp, P. (2019). Predictive analytics for leadership assessment. In J. I. Kantola, S. Nazir, & T. Barath (Eds.), *Advances in human factors, business management and society* (Vol. 783, pp. 516–523). Springer. https://doi.org/10.1007/978-3-319-94709-9_51
- * de-Juan-Ripoll, C., Llanes-Jurado, J., Giglioli, I. A. C., Marín-Morales, J., & Alcañiz, M. (2021). An immersive virtual reality game for predicting risk taking through the use of implicit measures. *Applied Sciences*, 11(2), Article 825. <https://doi.org/10.3390/app11020825>
- * de-Juan-Ripoll, C., Soler-Domínguez, J. L., Chicchi Giglioli, I. A., Contero, M., & Alcañiz, M. (2020). The Spheres & Shield Maze Task: A Virtual Reality Serious Game for the Assessment of Risk Taking in Decision Making. *Cyberpsychology, Behavior, and Social Networking*, 23(11), 773-781. <https://doi.org/10.1089/cyber.2019.0761>
- * DeRosier, M. E., & Thomas, J. M. (2018). Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Developmental Psychology*, 55, 52–61. <https://doi.org/10.1016/j.appdev.2017.03.001>
- * Desai, K., Ashtiani, O., & Prabhakaran, B. (2023). Assessment HTN (A-HTN) for automated task performance assessment in 3D serious games. arXiv. <https://arxiv.org/abs/2302.05795>
- * DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society*, 17(1), 17–28.

- * DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology*, 7-18.
- * DiCerbo, K. E., Xu, Y., Levy, R., Lai, E., & Holland, L. (2017). Modeling student cognition in digital and nondigital assessment environments. *Educational Assessment*, 22(4), 275–297. <https://doi.org/10.1080/10627197.2017.1382343>
- * Drey, T., Jansen, P., Fischbach, F., Frommel, J., & Rukzio, E. (2020). Towards progress assessment for adaptive hints in educational virtual reality games. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–9). Association for Computing Machinery. <https://doi.org/10.1145/3334480.3382789>
- * Falakmasir, M. H., González-Brenes, J. P., Gordon, G. J., & DiCerbo, K. E. (2016). A data-driven approach for inferring student proficiency from game activity logs. In *Proceedings of the Third ACM Conference on Learning @ Scale*(pp. 341–349). ACM.
- * Fang, Y., Allen, L. K., Roscoe, R. D., & McNamara, D. S. (2023). Stealth literacy assessment: Leveraging games and NLP in iSTART. In V. Yaneva & A. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 183–199). Routledge.
- * Fang, Y., Li, T., Roscoe, R. D., & McNamara, D. S. (2021). Predicting literacy skills via stealth assessment in a simple vocabulary game. In R. A. Sottolare & J. Schwarz (Eds.), *Adaptive instructional systems: Adaptation strategies and methods* (Vol. 12793, pp. 34–45). Springer. https://doi.org/10.1007/978-3-030-77873-6_3
- * Feng, X., & Yamada, M. (2019). Effects of game-based learning on informal historical learning: A learning analytics approach. In *Proceedings of the 27th International Conference on Computers in Education (ICCE 2019)* (pp. 505–514). Asia-Pacific Society for Computers in Education. <https://doi.org/10.58459/icce.2019.496>
- * Feng, X., & Yamada, M. (2021). An analytical approach for detecting and explaining the learning path patterns of an informal learning game. *Educational Technology & Society*, 24(1), 176-190.
- * Foroughi, C. K., Serraino, C., Parasuraman, R., & Boehm-Davis, D. A. (2016). Can we create a measure of fluid intelligence using Puzzle Creator within Portal 2?. *Intelligence*, 56, 58-64. <https://doi.org/10.1016/j.intell.2016.02.011>
- * Forsyth, C. M., Graesser, A. C., & Millis, K. (2020). Predicting learning in a multi-component serious game. *Technology, Knowledge and Learning*, 25, 251–277. https://doi.org/10.1007/978-3-030-34350-7_21
- * Georgiadis, K., Faber, T., & Westera, W. (2019). Bolstering stealth assessment in serious games. In A. Liapis, G. Yannakakis, M. Gentile, & M. Ninaus (Eds.), *Games and Learning Alliance (GALA 2019) (Lecture Notes in Computer Science, Vol. 11899, pp. 211–220)*. Springer. https://doi.org/10.1007/978-3-030-34350-7_21

- * Georgiadis, K., van Lankveld, G., Bahreini, K., & Westera, W. (2019). Reinforcing stealth assessment in serious games. In *International Conference on Games and Learning Alliance* (pp. 512–521). Springer. https://doi.org/10.1007/978-3-030-34350-7_49
- * Georgiadis, K., van Lankveld, G., Bahreini, K., & Westera, W. (2021). On the robustness of stealth assessment. *IEEE Transactions on Games*, 13(2), 180–192. <https://doi.org/10.1109/TG.2020.3020015>
- * Giglioli, I. A. C., Carrasco-Ribelles, L. A., Parra, E., Marín-Morales, J., & Alcañiz, M. (2021). An immersive serious game for the behavioral assessment of psychological needs. *Applied Sciences*, 11(4), Article 1971. <https://doi.org/10.3390/app11041971>
- * Guo, Q. (2018). *LSTM cluster: An integrated approach to cluster students' problem-solving sequences in log files* (Doctoral dissertation, University of Alberta).
- * Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2021). Multimodal multi-task stealth assessment for reflection-enriched game-based learning. In *MAIED@AIED* (pp. 93–102).
- * Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2023a). Enhancing stealth assessment in collaborative game-based learning with multi-task learning. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (Vol. 13916, pp. 243–253). Springer. https://doi.org/10.1007/978-3-031-36272-9_25
- * Gupta, A., Carpenter, D., Min, W., Rowe, J. P., Azevedo, R., & Lester, J. C. (2023b). Detecting and mitigating encoded bias in deep learning-based stealth assessment models for reflection-enriched game-based learning environments. *International Journal of Artificial Intelligence in Education*, 34, 1138–1165. <https://doi.org/10.1007/s40593-023-00379-6>
- * Halverson, R., & Owen, V. E. (2014). Game-based assessment: An integrated model for capturing evidence of learning in play. *International Journal of Learning Technology*, 9(2), 111–138. <https://doi.org/10.1504/IJLT.2014.064489>
- * Harpstead, E., Myers, B. A., & Alevan, V. (2013). In search of learning: Facilitating data analysis in educational games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 79–88). Association for Computing Machinery. <https://doi.org/10.1145/2470654.2470667>
- * Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2022). Enhancing stealth assessment in game-based learning environments with generative zero-shot learning. In *Proceedings of the International Conference on Educational Data Mining (EDM 2022)*. <https://eric.ed.gov/?id=ED624031>
- * Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2020). Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. In *Proceedings of the*

International Conference on Educational Data Mining (EDM 2020).
<https://eric.ed.gov/?id=ED607823>

- * Hutchins, N. M., Biswas, G., Maróti, M., Lédeczi, Á., Grover, S., Wolf, R., Blair, K. P., Chin, D., Conlin, L., Basu, S., & McElhaney, K. (2020). C2STEM: A system for synergistic learning of physics and computational thinking. *Journal of Science Education and Technology*, 29, 83–100. <https://doi.org/10.1007/s10956-019-09804-9>
- * Jaffal, Y., & Wloka, D. (2015). Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content. *Journal of e-Learning and Knowledge Society*, 11(3). <https://www.learntechlib.org/p/151924/>
- * Jemmali, C., Bunian, S., Mambretti, A., & El-Nasr, M. S. (2018, August). Educational game design: an empirical study of the effects of narrative. In *Proceedings of the 13th international conference on the foundations of digital games*(pp. 1-10). 10.1145/3235765.3235783
- * Ke, F., & Shute, V. J. (2015). Design of game-based stealth assessment and learning support. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Advances in game-based learning*. Springer. https://doi.org/10.1007/978-3-319-05834-4_13
- * Kiili, K., & Ketamo, H. (2017). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies*. <https://doi.org/10.1109/TLT.2017.2687458>
- * Kiili, K., Devlin, K., Perttula, A., Tuomi, P., & Lindstedt, A. (2015). Using video games to combine learning and assessment in mathematics education. *International Journal of Serious Games*, 2(4). <https://doi.org/10.17083/ijsg.v2i4.98>
- * Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, 87, 340-356. <https://doi.org/10.1016/j.compedu.2015.07.009>
- * Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in Physics Playground. *International Journal of Testing*, 16(2), 142–163. <https://doi.org/10.1080/15305058.2015.1108322>
- * Kim, Y. J., Knowles, M. A., Scianna, J., Lin, G., & Ruipérez-Valiente, J. A. (2023). Learning analytics application to examine validity and generalizability of game-based assessment for spatial reasoning. *British Journal of Educational Technology*, 54(1), 355–372. <https://doi.org/10.1111/bjet.13286>
- * Klaassen, M. (2021). “My log has something to tell you”: The value of log data in game-based learning (Master’s thesis, Utrecht University).
- * Kleitman, S., Jackson, S. A., Zhang, L. M., Blanchard, M. D., Rizvandi, N. B., & Aidman, E. (2022). Applying evidence-centered design to measure psychological resilience: The development and preliminary validation of a novel simulation-based assessment

methodology. *Frontiers in Psychology*, 12, 717568.
<https://doi.org/10.3389/fpsyg.2021.717568>

- * Levy, L., Solomon, R., Johnson, J., Wilson, J., Lambeth, A., Gandy, M., & Liu, R. (2016). Grouches, extraverts, and jellyfish: Assessment validity and game mechanics in a gamified assessment. In *Proceedings of the DiGRA/FDG 2016 Conference*.
<https://doi.org/10.26503/dl.v2016i1.779>
- * Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, 54(6), 771–794.
<https://doi.org/10.1080/00273171.2019.1590794>
- * Li, J., Bai, J., Zhu, S., & Yang, H. H. (2024). Game-based assessment of students' digital literacy using evidence-centered game design. *Electronics*, 13(2), Article 385.
<https://doi.org/10.3390/electronics13020385>
- * Liu, Z., & Moon, J. (2023). A framework for applying sequential data analytics to design personalized digital game-based learning for computing education. *Educational Technology & Society*, 26(2), 181–197. <https://doi.org/10.301>
- * Masduki, I. (2015). *Assessing job negotiation competencies of college students using evidence-centered design and branching simulations*.
- * Mayer, I. (2018). Assessment of teams in a digital game environment. *Simulation & Gaming*, 49(6), 602–619. <https://doi.org/10.1177/1046878118770831>
- * McCashin, L., McGarvey, L., Carbonaro, M., & Yuen, C. (2019). Assessing spatial geometry through digital gameplay in a Minecraft summer camp. *Ubiquitous Learning*, 12(4), 37. 10.18848/1835-9795/CGP/v12i04/37-57
- * McCreery, M. P., & Krach, S. K. (2019). Can video games be used as a stealth assessment of aggression? A criterion-related validity study. *International Journal of Gaming and Computer-Mediated Simulations*, 11(2), 40–49.
<https://doi.org/10.4018/IJGCMS.2019040103>
- * Min, W., Frankosky, M. H., Mott, B. W., Wiebe, E. N., Boyer, K. E., & Lester, J. C. (2017, June). Inducing stealth assessors from game interaction data. In *International Conference on Artificial Intelligence in Education* (pp. 212–223). Cham: Springer International Publishing.
- * Min, W., Mott, B. W., Rowe, J. P., Liu, B., & Lester, J. C. (2020). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, 13(2), 312–325. <https://doi.org/10.1109/TLT.2019.2922356>
- * Nowlan, N. S. (2022). *Assessment of Higher Order Thinking Skills in Virtual Learning Environments* (Doctoral dissertation, Carleton University).
- * Nowlan, N., Arya, A., Qorbani, H. S., & Abdinejad, M. (2023). Higher-order thinking skills assessment in 3D virtual learning environments using motifs and expert data. *Computers & Education: X Reality*, 2, Article 100012. <https://doi.org/10.1016/j.cexr.2023.100012>

- * Orozco-García, L., González, C., Montaña, J., Mondragón, C., & Tobar-Muñoz, H. (2019). A formative assessment tool to support computational thinking in the classroom. In *2019 International Conference on Virtual Reality and Visualization (ICVRV)* (pp. 185–188). IEEE. <https://doi.org/10.1109/ICVRV47840.2019.00043>
- * Parra Vargas, E., García Delgado, A., Torres, S. C., Carrasco-Ribelles, L. A., Marín-Morales, J., & Alcañiz, M. (2022). Virtual reality stimulation and organizational neuroscience for the assessment of empathy. *Frontiers in Psychology*, *13*, Article 993162. <https://doi.org/10.3389/fpsyg.2022.993162>
- * Parra, E., Chicchi Giglioli, I. A., Philip, J., Carrasco-Ribelles, L. A., Marín-Morales, J., & Alcaniz Raya, M. (2021). Combining virtual reality and organizational neuroscience for leadership assessment. *Applied Sciences*, *11*(13), 5956. <https://doi.org/10.3390/app11135956>
- * Pedersen, M. K., Díaz, C. M. C., Wang, Q. J., Alba-Marrugo, M. A., Amidi, A., Basaiawmoit, R. V., ... & Sherson, J. F. (2023). Measuring cognitive abilities in the wild: Validating a population-scale game-based cognitive assessment. *Cognitive Science*, *47*(6), e13308. [10.1111/cogs.13308](https://doi.org/10.1111/cogs.13308)
- * Plass, J. L., Homer, B. D., & Kinzer, C. K. (2013). Metrics in simulations and games for learning. In M. Seif El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics* (pp. 649–689). Springer. https://doi.org/10.1007/978-1-4471-4769-5_31
- * Polyak, S. T., von Davier, A. A., & Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Frontiers in Psychology*, *8*, Article 2029. <https://doi.org/10.3389/fpsyg.2017.02029>
- * Pool, J. (2019). *Stealth assessment and computer game learning: Could this benefit the children of tomorrow?* (Master's thesis, Anglia Ruskin University)
- * Poole, F. J. (2020). *Developing and validating stealth assessments for an educational game to assess young dual language immersion learners' reading comprehension* (Doctoral dissertation, Utah State University).
- * Presser, A. L., Vahey, P., & Zanchi, C. (2013). Designing early childhood math games: A research-driven approach. *Proceedings of the 12th International Conference on Interaction Design and Children*.
- * Rafner, J., Wang, Q. J., Gadjac, M., Badts, T., Baker, B., Bergenholtz, C., & Sherson, J. (2023). Towards game-based assessment of creative thinking. *Creativity Research Journal*, *35*(4), 763–782. <https://doi.org/10.1080/10400419.2023.2198845>
- * Rahimi, S., & Shute, V. J. (2021). First inspire, then instruct to improve students' creativity. *Computers & Education*, *174*, Article 104312. <https://doi.org/10.1016/j.compedu.2021.104312>
- * Rahimi, S., & Shute, V. J. (2024). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in

- technology-rich environments. *Educational Technology Research and Development*, 72, 2417–2441. <https://doi.org/10.1007/s11423-023-10232-1>
- * Rahimi, S., Shute, V., & Zhang, Q. (2021). The effects of game and student characteristics on persistence in educational games: A hierarchical linear modeling approach. *International Journal of Technology in Education and Science*, 5(2), 141-165.
 - * Rahimi, S., Smith, J. B., Truesdell, E. J. K., Vinay, A., Boyer, K. E., Magerko, B., Freeman, J., & McKlin, T. (2024). An automated, unobtrusive, formative assessment of creativity in a computer science and music remixing learning environment. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. <https://doi.org/10.1037/aca0000683>
 - * Rajendran, D., & Prasanna, S. (2022). Stealth assessment strategy in distributed systems using optimal deep learning with game-based learning. *The Journal of Supercomputing*, 78, 8285–8301. <https://doi.org/10.1007/s11227-021-04236-y>
 - * Rambhatla, S., & Gupta, D. (2016, December). Evidence Centred Approach to Measuring Learning Outcomes Amongst Management Students Using Epistemic Games. In *2016 IEEE Eighth International Conference on Technology for Education (T4E)* (pp. 120-123). IEEE.
 - * Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics. *British Journal of Educational Technology*, 46(1), 98–122. <https://doi.org/10.1111/bjet.12128>
 - * Reilly, J. M. (2020). *Dynamic feedback as automated scaffolding to support learners and teachers in guided authentic scientific inquiry settings* (Doctoral dissertation, Harvard University).
 - * Rezapour, M. M., Fatemi, A., & Nematbakhsh, M. A. (2024). Learning experience assessment through players' chat content in multiplayer online games. *Computers in Human Behavior*, 151, Article 108003. <https://doi.org/10.1016/j.chb.2023.108003>
 - * Roberto, C., Shi, Y., Amanmyradova, A., & Meghani, S. Predicting student's post-test performance using game data. Tech. Rep. 10.13140/RG.2.2.21373.20963
 - * Rosen, Y., Jaeger, G., Newstadt, M., Bakken, S., Rushkin, I., Dawood, M., & Purifoy, C. (2023). A multidimensional approach for enhancing and measuring creative thinking and cognitive skills. *International Journal of Information and Learning Technology*, 40(4), 334–352. <https://doi.org/10.1108/IJILT-12-2022-0227>
 - * Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in Zoombinis puzzle gameplay. *Computers in Human Behavior*, 120, Article 106707. <https://doi.org/10.1016/j.chb.2021.106707>
 - * Ruipérez-Valiente, J. A., Kim, Y. J., Baker, R. S., Martínez, P. A., & Lin, G. C. (2022). The affordances of multivariate elo-based learner modeling in game-based assessment. *IEEE transactions on learning technologies*, 16(2), 152-165. 10.1109/TLT.2022.3203912

- * Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, 4(1), 49–110. <https://doi.org/10.5281/zenodo.3554643>
- * Sabourin, J. L. (2013). *Stealth assessment of self-regulated learning in game-based learning environments* (Doctoral dissertation, North Carolina State University).
- * Saleh, A., Phillips, T., Glazewski, K. D., Hmelo-Silver, C. E., Mott, B., & Lester, J. (2022). Designing an adaptive collaborative problem-solving system with expanded evidence-centered design. In A. Weinberger, W. Chen, D. Hernández-Leo, & B. Chen (Eds.), *Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning—CSCL 2022* (pp. 328–331). International Society of the Learning Sciences.
- * Sánchez Castro, S., Pascual Sevillano, M. Á., & Fombona Cadavieco, J. (2024). Learning analytics in serious games as predictors of linguistic competence in students at risk. *Technology, Knowledge and Learning*, 29(3), 1551-1577. <https://doi.org/10.1007/s10758-024-09726-5>
- * Scianna, J., Gagnon, D., & Knowles, M. A. (2021). Counting the game: Visualizing changes in play by incorporating game events. In A. R. Ruis & S. B. Lee (Eds.), *Advances in quantitative ethnography (Communications in Computer and Information Science, Vol. 1312, pp. 231–246)*. Springer. https://doi.org/10.1007/978-3-030-67788-6_15
- * Shute, V. J., & Kim, Y. J. (2012). Does Playing World of Goo Facilitate Learning?. In *Design research on learning and thinking in educational settings* (pp. 243-267). Routledge.
- * Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 31–51). Emerald Publishing Limited. <https://doi.org/10.1108/978-1-68123>
- * Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, 116, Article 106647. <https://doi.org/10.1016/j.chb.2020.106647>
- * Shute, V. J., & Wang, L. (2015). Measuring problem solving skills in Portal 2. In *E-learning systems, environments and approaches: Theory and implementation* (pp. 11-24). Cham: Springer International Publishing.
- * Shute, V. J., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. In *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 535–562). Wiley. <https://doi.org/10.1002/9781118956588.ch22>
- * Shute, V. J., Ke, F., & Wang, L. (2016). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Eds.), *Techniques to improve the effectiveness of serious games* (pp. xxx–xxx). Springer. https://doi.org/10.1007/978-3-319-39298-1_4

- * Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R. G., Dai, C. P., & Liu, Z. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in educational games. *Journal of Computer Assisted Learning*, 37(1), 127–141. <https://doi.org/10.1111/jcal.12473>
- * Shute, V. J., Smith, G., Kuba, R., Dai, C. P., Rahimi, S., & Sun, C. (2021). The design, development, and testing of learning supports for the Physics Playground game. *International Journal of Artificial Intelligence in Education*, 31, 357–379. <https://doi.org/10.1007/s40593-020-00196-1>
- * Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton’s Playground. *The Journal of Educational Research*, 106(6), 423–430. <https://doi.org/10.1080/00220671.2013.832970>
- * Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In *Serious games* (pp. 295–321). Routledge.
- * Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem-solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, 63, 106–117. <https://doi.org/10.1016/j.chb.2016.05.047>
- * Simons, A., Wohlgenannt, I., Weinmann, M., & Fleischer, S. (2021). Good gamers, good managers? A proof-of-concept study with Sid Meier’s Civilization. *Review of Managerial Science*, 15, 957–990. <https://doi.org/10.1007/s11846-020-00378-0>
- * Simons, A., Wohlgenannt, I., Zelt, S., & Fleischer, S. (2023). Intelligence at play: Game-based assessment using a virtual-reality application. *Virtual Reality*, 27, 1827–1843. <https://doi.org/10.1007/s10055-023-00752-9>
- * Smith, D. (2014). *Earthquake Rebuild: A game for the stealth learning of middle school math* (Master’s thesis, The Florida State University).
- * Smith, A., Aksit, O., Min, W., Wiebe, E., Mott, B. W., & Lester, J. C. (2016). Integrating real-time drawing and writing diagnostic models: An evidence-centered design framework for multimodal science assessment. In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 9684, pp. 149–159). Springer. https://doi.org/10.1007/978-3-319-39583-8_16
- * Smith, G., Shute, V., & Muenzenberger, A. (2019). Designing and validating a stealth assessment for calculus competencies. *Journal of Applied Testing Technology*, 52-59.
- * Snow, E. L. (2014). *Stealth assessment of self-regulative behaviors within a game-based environment* (Master’s thesis, Arizona State University).
- * Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter? Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education*, 82, 378–392. <https://doi.org/10.1016/j.compedu.2014.12.011>

- * Spain, R., Min, W., Saville, J., Brawner, K., Mott, B., & Lester, J. (2021). Automated assessment of teamwork competencies using evidence-centered design-based natural language processing approach. In *Proceedings of the Ninth Annual GIFT Users Symposium (GIFTSym9)* (p. 140). U.S. Army DEVCOM Soldier Center.
- * Steinrücke, J., Veldkamp, B. P., & de Jong, T. (2019). Determining the effect of stress on analytical skills performance in digital decision games towards an unobtrusive measure of experienced stress in gameplay scenarios. *Computers in Human Behavior*, 99, 144–155. <https://doi.org/10.1016/j.chb.2019.05.014>
- * Steinrücke, J., Veldkamp, B. P., & de Jong, T. (2020). Information literacy skills assessment in digital crisis management training for the safety domain: Developing an unobtrusive method. *Frontiers in Education*, 5, Article 140. <https://doi.org/10.3389/feduc.2020.00140>
- * Suleiman, I., Arslan, M., Alhajj, R., & Rokne, J. (2016). The power of online genetic algorithm in stealth assessment for school readiness. *Journal of Computers in Education*, 3, 209–246. <https://doi.org/10.1007/s40692-016-0054-5>
- * Tadayon, R., Amresh, A., McDaniel, T., & Panchanathan, S. (2018). Real-time stealth intervention for motor learning using player flow-state. In *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1–8). IEEE. <https://doi.org/10.1109/SeGAH.2018.8401360>
- * Tanner, O. D., & Lou, Y. (2017). Design of a digital science inquiry game-based assessment. *Paper presented at the American Educational Research Association Annual Meeting*.
- * Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. *Computers in Human Behavior*, 76, 641–655. <https://doi.org/10.1016/j.chb.2017.01.038>
- * Tenorio, M., Arango, P. S., & Aparicio, A. (2023). BENDI: Improving cognitive assessments in toddlers and children with Down syndrome using stealth assessment. *Children*, 10(12), Article 1923. <https://doi.org/10.3390/children10121923>
- * Timms, M. J. (2012). *Session N—SimScientists: An example of how technology can support differentiated instruction in the classroom*. ACER Research Conference.
- * Toutkoushian, E., & Ryoo, K. (2018). Developing a method to use log files to understand NGSS-aligned science learning. In *Proceedings of the International Conference on Educational Data Mining (EDM 2018)*.
- * Udeozor, C., Abegão, F. R., & Glassey, J. (2024). Measuring learning in digital games: Applying a game-based assessment framework. *British Journal of Educational Technology*, 55(2), 957–991. <https://doi.org/10.1111/bjet.13407>
- * Udeozor, C., Dominguez Alfaro, J. L., & Glassey, J. (2023). Assessment framework for immersive learning: Application and evaluation. In M. L. Bourguet, J. M. Krüger, D. Pedrosa, A. Dengel, A. Peña-Rios, & J. Richter (Eds.), *Immersive Learning Research*

Network (iLRN 2023) (Communications in Computer and Information Science, Vol. 1904, pp. 221–236). Springer. https://doi.org/10.1007/978-3-031-47328-9_15

- * Udeozor, C., Russo, A., & Glassey, J. (2021). Exploring log data for behaviour and solution pattern analysis in a serious game. In *Gamification and social networks in education*.
- Veritas Health Innovation. (n.d.). *Covidence systematic review software*. Retrieved from <https://www.covidence.org/>
- * Verma, V. (2021). *Content agnostic game-based stealth assessment* (Doctoral dissertation, Arizona State University).
- * Verma, V., Craig, S. D., Levy, R., Bansal, A., & Amresh, A. (2022). Domain knowledge and adaptive serious games: Exploring the relationship of learner ability and affect adaptability. *Journal of Educational Computing Research*, 60(2), 406–432. <https://doi.org/10.1177/07356331211031287>
- * Walsh, C., & Bokhove, C. (2021). Targeting data collection in games based assessment. *Computers and Education Open*, 2, 100054. <https://doi.org/10.1016/j.caeo.2021.100054>
- * Westera, W., Prada, R., Mascarenhas, S., Santos, P. A., Dias, J., & Guimarães, R. (2020). Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies*, 25, 351–380. <https://doi.org/10.1007/s10639-019-09968-2>
- * Xie, C., Schimpf, C., Chao, J., Nourian, S., & Massicotte, J. (2018). Learning and teaching engineering design through modeling and simulation on a CAD platform. *Computer Applications in Engineering Education*, 26(3), 824–840. <https://doi.org/10.1002/cae.21920>
- * Yang, D., Zargar, E., Adams, A. M., Day, S. L., & Connor, C. M. (2021). Using interactive e-book user log variables to track reading processes and predict digital learning outcomes. *Assessment for Effective Intervention*, 46(4), 292–303. <https://doi.org/10.1177/1534508420941935>
- * Yu, J., Ma, W., Moon, J., & Denham, A. (2022). Developing a stealth assessment system using a continuous conjunctive model. *Journal of Learning Analytics*, 9(3), 11–31. <https://doi.org/10.18608/jla.2022.7639>
- * Zhang, J., Ober, T. M., Jiang, Y., Plass, J. L., & Homer, B. D. (2021). Predicting executive functions in a learning game: Accuracy and reaction time. In *Proceedings of the International Conference on Educational Data Mining (EDM 2021)*.

Appendix A

Competency Categorization Using Different Frameworks

Code	Category	What Counts	Decision Rule	Example Constructs
NRC-C	Cognitive	Domain knowledge and reasoning skills (problem solving, executive functions, scientific reasoning, academic skills)	Select when the construct involves thinking, reasoning, or knowledge application	Math skills, physics understanding, reading comprehension, creativity, systems thinking, spatial reasoning
NRC-I	Intrapersonal	Internal regulation, motivation, persistence, SRL, personality, resilience	Select when the construct reflects self-regulation, traits, or motivation	Persistence, resilience, agency, risk-taking, SRL behaviors, Big Five traits
NRC-P	Interpersonal	Social interaction and teamwork skills	Select when the construct requires interaction or collaboration	Collaboration, teamwork, leadership, empathy, negotiation
P21-4C	Learning & Innovation (4Cs)	Creativity, critical thinking, communication, collaboration	Use when explicitly assessing one of the 4Cs	Creative thinking, collaborative problem solving, teamwork
P21-ICT	Information & Technology Skills	Digital literacy, ICT fluency, programming, data literacy	Use when the construct involves technology or data use	Programming, computational thinking, digital literacy
P21-LC	Life & Career Skills	Self-direction, flexibility, leadership, productivity	Use for work habits and adaptive behaviors	Leadership, resilience, agency, management skills

P21-CS	Core Subjects / Academic Competencies	Discipline-based academic outcomes and foundational cognitive skills tied to academic achievement	Use when targeting core academic domains or academic cognitive abilities	Mathematics knowledge, science concepts, reading comprehension, spatial reasoning, executive functions
WEF-AN	Analytical Thinking	Logical reasoning, quantitative analysis, rule-based problem solving	Use for reasoning-intensive cognitive constructs	Scientific reasoning, math modeling, information analysis
WEF-CR	Creative Thinking	Originality, idea generation, innovation	Use for creativity-focused outcomes	Divergent thinking, innovation
WEF-RES	Resilience & Adaptability	Persistence, coping, flexibility	Use when focus is on adaptability or stress tolerance	Persistence, resilience
WEF-LEAD	Leadership & Social Influence	Leading, influencing, coordinating others	Use for leadership or social influence outcomes	Leadership styles, negotiation
WEF-TECH	Technology Literacy	Digital, AI, or computational competence	Use for digital/technology skill constructs	Programming, networking, digital literacy
WEF-LLL	Lifelong Learning	Metacognition, learning-to-learn, SRL	Use for metacognitive or continuous learning constructs	SRL behaviors, metacognition, self-explanation

Appendix B

Search Strings without any restrictions

Databases	Search Strings	Result
SCOPUS	("stealth assessment" OR "unobtrusive assessment" OR "embedded-assessment" OR "embedded assessment" OR "game-based assessment") AND ("evidence-centered design" OR "evidence centered design" OR "computer-based assessment for learning" OR "CA4L") AND ("game*" OR "online-learning" OR "virtual reality" OR "technology*" OR "computer supported" OR "simulation" OR digital)	535
Google Scholar	("stealth assessment" OR "embedded assessment" OR "game-based assessment") AND ("evidence centered design" OR ECD) AND ("game*" OR "online-learning" OR "virtual reality" OR "technology*" OR "computer supported" OR "simulation" OR digital)	1,246
Web of Science	ALL=(("stealth assessment" OR "unobtrusive assessment" OR "embedded-assessment" OR "embedded assessment" OR "game-based assessment" OR "evidence-centered design" OR "evidence centered design" OR "computer-based assessment for learning" OR "CA4L") AND ("game*" OR "online-learning" OR "virtual reality" OR "technology*" OR "computer supported" OR "simulation" OR digital))	208
Science Direct	("stealth assessment" OR "embedded-assessment" OR "embedded assessment" OR "game-based assessment" OR "evidence-centered design") AND (game OR "technology" OR "computer supported" OR digital)	335
ERIC Proquest	("stealth assessment" OR "embedded-assessment" OR "embedded assessment" OR "game-based assessment" OR "evidence-centered design" OR "evidence centered design") AND ("game*" OR "technology*" OR ("computer support" OR "computer supported") OR "digital*" OR "video*")	182
Total		2,506

Appendix C

Studies included in this review

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Bauer (2003)	87.5	Ed Tech / LS	NetPASS	Computer Networking	Adult & Cont. Ed.	NA	Full cycle	Advanced
Shute (2009)	87.5	Ed Tech / LS	Elder Scrolls IV: Oblivion	Creative PS	Not Specified	NA	Design	Advanced
Timms (2012)	87.5	Ed Tech / LS	SimScientists	Scientific Inquiry, Science Content Knowledge	pre-K–12	5867	Full cycle	Advanced
Rupp (2012)	100	Ed Tech / LS	Packet Tracer	Computer Networking	Higher Ed.	2,269	Full cycle	Simple
Shute (2012)	100	Ed Tech / LS	World of Goo	Physics, PS	pre-K–12	5	Full cycle	Advanced
Sweet (2012)	81.25	Ed Tech / LS	Land Science	PS	Not Specified	NA	Design	Advanced
Shute (2013)	81.25	Ed Tech / LS	Physics Playground	Physics	pre-K–12	154	Validation	Simple
Plass (2013) – study 1	100	Ed Tech / LS	Kinetic Molecular Theory	Metacognitive Processes/Control	pre-K–12	457	Validation	Advanced
Plass (2013) – study 2	100	Ed Tech / LS	FactorReactor	Mathematics & Cognitive Flexibility	pre-K–12	64	Validation	Advanced
Plass (2013) – study 3	100	Ed Tech / LS	Noobs vs. Leets	Geometry / Spatial Reasoning	pre-K–12	89	Full cycle	Advanced

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Sabourin (2013)	100	CS / AI in Ed	Crystal Island	SLR	pre-K–12	260	Full cycle	HS
Harpstead (2013)	81.25	CS / AI in Ed	RumbleBlocks	Structural Stability in Physics	pre-K–12	174	Validation	HS
Presser (2013)	100	Ed Tech / LS	Next Generation Preschool Math" project	Early Numeracy, Rational Numbers	pre-K–12	50	Validation	Simple
Shute (2013)	87.5	Ed Tech / LS	Portal 2	PS	pre-K–12	220	Validation	Simple
DiCerbo (2014)	93.75	Psychometric	Poptropica	Persistence	pre-K–12	892	Design	Intermediate
Halverson (2014)	100	Ed Tech / LS	Progenitor X	Stem-Cell Science	pre-K–12	110	Validation	Simple
Snow (2014)	100	Ed Tech / LS	iSTART-ME	SLR	pre-K–12	40	Validation	Simple
Smith (2014)	87.5	CS / AI in Ed	Earthquake Rebuild	Math	pre-K–12	4	Design	Simple
Kiili (2015)	68.75	Ed Tech / LS	Semideus & Wuzzit Trouble	Math	pre-K–12	66	Validation	Simple
Capuano (2015)	75	AD	A 3D Model of a School	Safety Skills	pre-K–12	45	Full cycle	Simple
Snow (2015)	100	Ed Tech / LS	iSTART-2	Agency	Higher Ed.	70	Validation	Simple
Ke (2015) – study 1	75	Ed Tech / LS	Portal 2	Spatial Reasoning	Higher Ed.	77	Design	Simple

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Ke (2015) – study 2	75	Ed Tech / LS	Earthquake Rebuild	Ratio And Proportional Reasoning	pre-K–12	NA	Design	Simple
Baron (2015)	100	CS / AI in Ed	Word Towers	Spelling Skills	pre-K–12	17	Design	Simple
Bertling (2015)	87.5	Psychometric	Mars Generation One	CT	pre-K–12	950	Validation	Intermediate
Jaffal (2015)	100	CS / AI in Ed	Bicycle World	Traffic Education	pre-K–12	109	Validation	Simple
Reese (2015)	100	Ed Tech / LS	Selene: A Lunar Construction GaME	Geology	pre-K–12	357	Validation	Simple
Kim (2015)	100	Ed Tech / LS	Physics Playground	Physics	Higher Ed.	102	Validation	Intermediate
Masduki (2015)	100	Ed Tech / LS	Branching Simulation Tool	Job Negotiation Skills	Higher Ed.	137	Full cycle	Intermediate
Kiili (2016)	100	Ed Tech / LS	Semideus	Math, Fraction Knowledge	pre-K–12	51	Full cycle	Simple
Shute (2016)	75	Ed Tech / LS	Physics Playground	Creativity	pre-K–12	NA	Validation	Advanced
Shute (2016)	87.5	Ed Tech / LS	Use Your Brainz (PvZ2)	PS	pre-K–12	47	Validation	Advanced
Chin (2016)	100	Ed Tech / LS	Stroylet	Data Literacy	pre-K–12	93	Validation	Simple
Shute (2016)	75	Ed Tech / LS	Use Your Brainz	PS	pre-K–12	47	Validation	Advanced

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Suleiman (2016)	87.5	CS / AI in Ed	NA	School Readiness	Not Specified	NA	Design	Advanced
Falakmasir (2016)	50	Psychometric	Alice in Arealand	Geometry	pre-K–12	129	Validation	HS
Smith (2016)	75	CS / AI in Ed	LEONARDO Digital Science Notebook	Scientific Inquiry	pre-K–12	98	Validation	HS
Kim (2016)	100	Ed Tech / LS	Physics Playground	Physics, Creativity	pre-K–12	167	Full cycle	Advanced
Levy (2016)	100	Ed Tech / LS	Bubble Trip	Personality Traits	Adult & Cont. Ed.	212	Validation	Simple
Foroughi (2016)	75	Ed Tech / LS	Portal 2 Test Battery	Fluid Intelligence	Higher Ed.	135	Validation	Intermediate
DiCerbo (2017)	87.5	Psychometric	Alice in Arealand	Math Geometry	pre-K–12	131	Validation	Advanced
Shute (2017)	100	Ed Tech / LS	Physics Playground	Creativity	pre-K–12	77	Validation	Advanced
Antoniou (2017)	100	AD	Serious talk	Elderly Tech Acceptance	Adult & Cont. Ed.	21	Validation	Simple
Taub (2017)	87.5	CS / AI in Ed	Crystal Island	SLR	Higher Ed.	50	Validation	Intermediate
Polyak (2017)	100	CS / AI in Ed	Circuit Runner	PS	pre-K–12	159	Validation	HS
Tanner (2017)	100	Ed Tech / LS	Chicken Collision	Science Content Knowledge	pre-K–12	82	Design	Simple
Bley (2017)	100	Ed Tech / LS	ALUSIM Ltd	Intrapreneurship Competence	Adult & Cont. Ed.	919	Full cycle	Advanced

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Rambhatla (2017)	87.5	AD	STRAT UP	Management Skills	Higher Ed.	12	Validation	Advanced
Min (2017)	68.75	CS / AI in Ed	ENGAGE	CT	pre-K–12	191	Validation	HS
DiCerbo (2017)	81.25	Psychometric	Insight Learning System	Geometry	pre-K–12	131	Validation	Simple
Akram (2018)	75	CS / AI in Ed	ENGAGE	CT	pre-K–12	244	Validation	HS
de Heer (2018)	75	CS / AI in Ed	DILEMMA	Leadership	Not Specified	NA	Design	HS
Tadayon (2018)	100	CS / AI in Ed	Fruit Island	Motor Skills	Not Specified	1	Full cycle	HS
DeRosier (2018)	93.75	AD	Zoo U	Social Emotional Skills	pre-K–12	270	Validation	Simple
Mayer (2018)	100	AD	TEAMUP	Collaboration	Adult & Cont. Ed.	424	Validation	Simple
Guo (2018)	87.5	Ed Tech / LS	TRE Science Laboratory	PS, Data Analysis	pre-K–12	193	Validation	HS
Toutkoushian (2018)	87.5	Ed Tech / LS	Properties of Matter (POM) & Chemical Reactions (CR).	NGSS Science Education	pre-K–12	158	Validation	Advanced
Xie (2018)	62.5	Ed Tech / LS	Energy3D	Renewable Energy System Design	pre-K–12	60	Full cycle	HS

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Al-Smadi (2018)	100	CS / AI in Ed	Fire Evacuation Game	Fire Evacuation Skills	pre-K–12	45	Validation	Simple
Jemmali (2018)	87.5	Ed Tech / LS	May's Journey	OOP Skills	Adult & Cont. Ed.	73	Design	Simple
Chu (2018)	81.25	Ed Tech / LS	Raging Skies	Science Content Knowledge	pre-K–12	NA	Design	Simple
Georgiadis (2019)	87.5	CS / AI in Ed	The poisoned lake	Big Five Personality Traits	NA	80	Validation	HS
Georgiadis (2019)	68.75	Ed Tech / LS	abcdeSIM	Emergency Care	Adult & Cont. Ed.	267	Validation	HS
Levy (2019)	93.75	CS / AI in Ed	Save Patch	Math	pre-K–12	851	Validation	Advanced
Smith (2019)	87.5	Ed Tech / LS	Variant: LimitsTM	Calculus	Higher Ed.	148	Validation	Simple
McCreery (2019)	62.5	Ed Tech / LS	The Deed	Aggression	Higher Ed.	157	Validation	Simple
Pool (2019)	100	AD	Mangahigh	Mathematics	pre-K–12	10	Design	Advanced
Akram (2019)	87.5	CS / AI in Ed	ENGAGE	Problem Solving Strategies	pre-K–12	300	Full cycle	HS
Alonso- Fernández (2019)	100	CS / AI in Ed	First-Aid Game	Basic Life-Support Manoeuvres	pre-K–12	227	Validation	HS
Calvo-Morata (2019)	100	CS / AI in Ed	Conectado and First Aid Game	Bullying Awareness	Not Specified	NA	Validation	HS

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Feng (2019)	87.5	Ed Tech / LS	Hist Maker	Informal Historical Learning	pre-K–12	185	Validation	HS
Avi Shena (2019)	100	Ed Tech / LS	Digital Quiz Game	PS	pre-K–12	39	Validation	Simple
Orozco-Garcia (2019)	87.5	CS / AI in Ed	HERA	CT	pre-K–12	57	Design	Simple
Cui (2019)	100	Ed Tech / LS	Raging Skies	Weather Phenomena	pre-K–12	460	Validation	Advanced
Steinrücke (2019)	100	Ed Tech / LS	The Mayor Game	Experienced Stress	Higher Ed.	80	Validation	Advanced
Roberto (2019)	75	CS / AI in Ed	Ratio Rancher	Ratios And Proportions	pre-K–12	245	Validation	HS
McCashin (2019)	87.5	Ed Tech / LS	Minecraft	Spatial Geometry	pre-K–12	77	Validation	HS
Abbasi (2020)	75	CS / AI in Ed	OOg	OOP Skills	NA	NA	Design	Advanced
de-Juan-Ripoll (2020)	75	AD	Spheres & Shield Maze Task	Risk Taking	Adult & Cont. Ed.	41	Validation	Simple
Bounajim (2020)	75	CS / AI in Ed	Geniventure	Genetics	pre-K–12	394	Design	Simple
Shute (2020)	100	Ed Tech / LS	Physics Playground	Physics	pre-K–12	263	Full cycle	Advanced
Yang (2020)	87.5	Ed Tech / LS	WKe-Book	Word Knowledge & Strategic Reading Skills	pre-K–12	573	Validation	Simple

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Chen (2020)	100	Ed Tech / LS	Raging Skie	Whether Phenomena	pre-K–12	460	Validation	HS
Henderson (2020)	100	CS / AI in Ed	Geniventure	Genetics	pre-K–12	462	Validation	HS
Reilly (2020)	87.5	Ed Tech / LS	EcoXPT	Eco System Knowledge	pre-K–12	587	Validation	Advanced
Frederick (2020)	87.5	Ed Tech / LS	Legend of the Dragon	Chinese Reading	pre-K–12	32	Full cycle	Advanced
Min (2020)	100	CS / AI in Ed	ENGAGE	CT	pre-K–12	182	Full cycle	HS
Westera (2020)	100	CS / AI in Ed	VIBOA	NA	Not Specified	NA	Design	HS
Steinrücke et al. (2020)	100	Ed Tech / LS	dilemma game	Information Literacy	Adult & Cont. Ed.	40	Validation	HS
Forsyth (2020)	87.5	Ed Tech / LS	Operation ARA	Research Methodology, Scientific Reasoning	Higher Ed.	192	Validation	Intermediate
Hutchins (2020)	87.5	Ed Tech / LS	C2STEM	Physics, CT	pre-K–12	174	Validation	Simple
Basu (2020)	100	CS / AI in Ed	Beats Empire	Data Literacy	pre-K–12	28	Validation	Simple
Akram (2020)	87.5	CS / AI in Ed	ENGAGE	CT	pre-K–12	69	Validation	HS
Shute (2021)	100	Ed Tech / LS	Physics Playground	Creativity	pre-K–12	167	Validation	Advanced
Georgiadis (2021)	100	CS / AI in Ed	NA	NA	Not Specified	NA	Validation	HS
Verma (2021)	100	Ed Tech / LS	Chemo-o-crypt	Chemistry	Higher Ed.	107	Validation	Advanced

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Gupta (2021)	93.75	CS / AI in Ed	Crystal Island	Microbiology	pre-K–12	119	Validation	HS
Verma (2021)	100	Psychometric	Chem-o-crypt	Chemistry	Higher Ed.	107	Full cycle	Advanced
de-Juan-Ripoll (2021)	100	AD	AEMIN	Risk Taking	Adult & Cont. Ed.	88	Full cycle	HS
Giglioli (2021)	100	AD	ATHENEA	Core Psychological Needs	Adult & Cont. Ed.	61	Validation	HS
Zhang et al. (2021)	100	Ed Tech / LS	All You Can E.T. (AYCET)	Executive Functions	pre-K–12	137	Validation	Intermediate
Simons (2021)	100	Ed Tech / LS	Sid Meier's Civilization	Managerial Skills	Higher Ed.	40	Summative Assessment	Intermediate
Fang (2021)	87.5	Ed Tech / LS	iSTART	Reading	Higher Ed.	211	Validation	Intermediate
Klaassen (2021)	50	Ed Tech / LS	Zeldenrust	Mathematics	pre-K–12	355	Validation	Intermediate
Rowe (2021)	87.5	Ed Tech / LS	Zoombinis	CT	pre-K–12	194	Validation	HS
Shute (2021)	87.5	Ed Tech / LS	Physics Playground	Physics	pre-K–12	263	Full cycle	Advanced
Spain (2021)	87.5	CS / AI in Ed	GIFT	Teamwork Competencies	Adult & Cont. Ed.	NA	Design	Simple
Udeozor (2021)	87.5	AD	CosmiClean	PS	Higher Ed.	58	Validation	HS
Scianna (2021)	100	Ed Tech / LS	Lakeland	Systems Thinking	pre-K–12	NA	Summative Assessment	HS
Rahimi (2021)	100	Ed Tech / LS	Physics Playground	Creativity	Higher Ed.	114	Full cycle	Simple

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Walsh (2021)	87.5	Ed Tech / LS	Bidmas Blaster, ©Blue Duck Education.	Numeracy, PS	pre-K–12	200	Validation	Advanced
Parra (2021)	87.5	AD	NA	Leadership Styles	Adult & Cont. Ed.	51	Validation	HS
Rahimi (2021)	87.5	Ed Tech / LS	Physics Playground	Physics & Persistence	pre-K–12	137	Full cycle	Intermediate
Feng (2021)	87.5	Ed Tech / LS	Hist Maker	Chinese History Knowledge	Mixed	548	Validation	HS
Rajendran (2022)	100	CS / AI in Ed	NA	PS	Not Specified	NA	Validation	HS
Henderson (2022)	87.5	CS / AI in Ed	Geniventure	Genetics	pre-K–12	462	Validation	HS
Drey (2022)	100	Ed Tech / LS	Social Engineer	PS	Adult & Cont. Ed.	NA	Design	Simple
Yu (2022)	50	Ed Tech / LS	The Nomads	Mathematics	pre-K–12	84	Full cycle	Advanced
Parra Vargas (2022)	75	AD	VROE	Cognitive And Affective Empathy	Adult & Cont. Ed.	82	Validation	HS
Saleh (2022)	75	Ed Tech / LS	CRYSTAL ISLAND	PS	pre-K–12	NA	Design	Advanced
Kleitman (2022)	93.75	AD	Driving Simulation	Psychological Resilience	Higher Ed.	90	Validation	Simple
Nowlan (2022)	100	Ed Tech / LS	Carleton Virtual 3DVLE	Higher-Order Thinking	Higher Ed.	44	Validation	HS

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Desai (2023)	93.75	CS / AI in Ed	two 3D VR serious games	Chemistry	Higher Ed.	45	Validation	Simple
Rosen (2023)	87.5	Ed Tech / LS	BrainPOP	4C and Cognitive Skills	pre-K–12	211	Validation	Advanced
Amorim (2023)	75	Ed Tech / LS	Escribo Play	Early Mathematics	pre-K–12	2980	Validation	Simple
Nowlan (2023)	100	Ed Tech / LS	3DVLE Virtual Chemistry Lab	Higher-Order Thinking	Higher Ed.	36	Validation	HS
Rafner (2023)	87.5	Ed Tech / LS	CREA Suite	Creative Thinking	Higher Ed.	408	Validation	Simple
Kim (2023)	87.5	Ed Tech / LS	Shadowspect	Geometry	pre-K–12	44	Full cycle	HS
Simons (2023)	93.75	AD	Job Simulator	Creativity	Higher Ed.	103	Validation	Simple
Rezapour (2023)	87.5	CS / AI in Ed	Dota 2	PS	Adult & Cont. Ed.	35517	Validation	HS
Bang (2023)	87.5	Ed Tech / LS	My Math Academy	Math Skills	pre-K–12	886	Validation	Intermediate
Liu (2023)	87.5	Ed Tech / LS	Penguin Go	CT	pre-K–12	79	Validation	Simple
Fang (2023)	87.5	CS / AI in Ed	iSTART	Reading Comprehension	pre-K–12	121	Validation	Intermediate
Rahimi (2023)	100	Ed Tech / LS	Physics Playground	Creativity	pre-K–12	430	Full cycle	Advanced
Tenorio (2023)	87.5	Ed Tech / LS	BENDI	Core Cognitive Functions	pre-K–12	68	Full cycle	Simple
Christhlf (2023)	87.5	Ed Tech / LS	iSTART games	Reading Comprehension	Higher Ed.	405	Validation	Intermediate

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Pedersen (2023)	87.5	AD	Rat Catch	Core Cognitive Functions	Adult & Cont. Ed.	10,725	Validation	Intermediate
Caballero- Hernández (2023)	75	CS / AI in Ed	Custom Serious Game for E/R Diagram Design	PS	Higher Ed.	110	Design	HS
Ruiperez- Valiente (2023)	87.5	Ed Tech / LS	Shadowspect	Geometric Spatial Reasoning	pre-K–12	322	Validation	Intermediate
Gupta (2023a)	87.5	CS / AI in Ed	Crystal Island: EcoJourneys	Ecosystem Science & Collaboration Skills	pre-K–12	72	Validation	HS
Gupta (2023b)	100	CS / AI in Ed	Crystal Island	Decision-Making Processes	pre-K–12	119	Validation	HS
Udeozor (2024)	100	Ed Tech / LS	MAR (Mobile augmented reality (MAR) titration game)	Chemistry, Lab Safety	Adult & Cont. Ed.	3	Validation	Simple
Udeozor (2024)	100	Ed Tech / LS	CosmiClean	Chemical Separation Processes	Higher Ed.	146	Design	Simple
Li (2024)	93.75	Ed Tech / LS	Guogan's Journey to A Mysterious Planet	Digital Literacy	pre-K–12	188	Validation	Advanced
Rahimi (2024)	100	Ed Tech / LS	EarSketch	Creativity	pre-K–12	53	Full cycle	Simple

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Sánchez Castro (2024)	75	Ed Tech / LS	Walinwa and Leobien	Linguistic Competence	pre-K–12	75	Validation	Intermediate
Luo (2025)	81.3	CS / AI in Ed	GCTPt	Computational Thinking	K-12	74	Validation	Simple
Fiorentini (2025)	93.8	Ed Tech / LS	WWTW	Comprehension	Higher Education	87	Validation	Simple
Poole (2025)	87.5	Ed Tech / LS	Legend of the Dragon	Chinese Reading Comprehension	K-12	34	Validation	Advanced
Lu (2025)	81.3	CS / AI in Ed	Mission HydroSci	Science content knowledge	K-12	632	Validation	HS
Lu (2026)	75.0	Ed Tech / LS	Mission HydroSci	Oral Reading Fluency	K-12	632	Validation	HS
Rjiba (2025\)	81.3	Ed Tech / LS	GRADES Framework	Collaboration	Not Specified	100	Validation	HS
Rowe (2026)	93.8	Psychometric	Zoombinis	Computational Thinking	Mixed	716	Validation	HS
Rahimi (2026)	87.5	Ed Tech / LS	Physics Playground	Physics	Higher Education	114	Validation	HS
Cohn (2026)	75.0	CS / AI in Ed	Cupa	Physics	K-12	33	Full cycle	HS
Acosta (2026)	81.3	CS / AI in Ed	ecojourneys	Problem Solving	K-12	76	Full cycle	HS
Gupta (2026)	81.3	Ed Tech / LS	Crystal Island	Microbiology knowledge	K-12	119	Validation	HS
Adair (2026)	93.8	CS / AI in Ed	Inq-ITS platform	Mathematical modeling	K-12	121	Validation	HS

First Author (Year)	Quality Assessment	Field of Study	Game/ Environment	Competency	Target Audience	N	Status	Modeling Sophistication
Hadyaoui (2026)	93.8	CS / AI in Ed	GenSolve	Self-Regulated Learning	Higher Education	250	Validation	HS
Cao (2026)	100	Ed Tech / LS	Wonda Platform	Negotiation	Not Specified	7	Summative Assessment	HS
Liu (2026)	87.5	Ed Tech / LS	Online Community Harmony Keeper	digital etiquette	K-12	82	Full cycle	HS
Beigman Klebanov (2026)	81.3	Psychometric	Relay Reader	Reading comprehension	K-12	431	Validation	HS
Wang (2026)	87.5	Ed Tech / LS	Relay Reader	Oral Reading Fluency	K-12	80	Validation	HS
Liu (2026)	87.5	Ed Tech / LS	Penguin Go	Computational Thinking	K-12	79	Validation	Advanced

Note: Ed Tech = Educational Technology, LS = Learning Sciences, CS = Computer Science, AI in Ed = Artificial Intelligence in Education, BN = Bayesian Networks, DBN = Dynamic Bayesian Networks, ENA = Epistemic Network Analysis, EI = Evidence Identification, EA = Evidence Accumulation, QA = Quality Assessment, HS = Highly Sophisticated, Adult & Cont. Ed. = Adult & Continuing Education, AD = Applied Domains, CT = Computational Thinking, OOP Skills = Object Oriented Programming Skills, PS = Problem Solving, SLR = Self-Regulation Learning, Higher Ed. = Higher Education