# Evidence-Centered Foundations of Stealth Assessment

*Valerie J. Shute* (Florida State University)

*Seyedahmad Rahimi* (University of Florida)

## Abstract

Stealth assessment has emerged as a powerful approach for measuring and supporting learning within digital environments without interrupting engagement. Its conceptual coherence and psychometric rigor, however, rest squarely on the foundations of Robert Mislevy's evidence-centered design (ECD). ECD provides the inferential architecture that enables valid claims about learners' competencies to be drawn from complex streams of behavioral data. In this paper, we argue that ECD is the *sine qua non* of stealth assessment: without it, stealth assessment would be reduced to opportunistic data mining rather than psychometrically sound, principled measurement. We trace the historical development of stealth assessment, describe how ECD structures its core models, and illustrate their integration through an extended example from *Physics Playground*. We conclude by outlining future directions that further deepen, rather than dilute, the evidence-centered foundations of stealth assessment.

## Background

Robert Mislevy didn't just change how we measure learning; he provided the syntax for a language we are still learning to speak when referring to educational assessment. At the heart of this transformation lies Evidence-Centered Design (ECD; Mislevy et al., 2003; Mislevy et al., 2013; Mislevy, 2018)—a framework that shifted assessment from the craft of test-making to the construction of principled evidentiary arguments. ECD reframed assessment as an inferential enterprise: one in which claims about what learners know and can do must be explicitly warranted by evidence.

For our work on stealth assessment, ECD is not merely a reference point but its *sine qua non*. By separating the logic of inference from the mechanics of administration, Mislevy offered a blueprint for moving assessment beyond intrusive "stop-and-test" moments to embedding measurement directly within the activity of learning itself.

This reconceptualization matters because learning is not episodic; rather, it unfolds continuously and is easily disrupted by frequent interruptions. Learning also does not occur in a vacuum. Moreover, learners differ widely in their prior knowledge, skills, motivations, affective states, goals, backgrounds, and opportunities, and these differences shape how learning trajectories emerge over time. What happens during the learning process—moment by moment—thus plays a decisive role in shaping eventual outcomes. An assessment approach grounded in ECD makes it possible to attend to these processes without halting them, supporting inferences that are both psychometrically sound and educationally useful. Stealth assessment builds directly on this evidentiary logic to make learning processes visible, enabling systems not only to produce reliable and valid inferences, but also to

deliver learning supports that are timely, personalized, and instructionally meaningful (Rahimi & Shute, 2025; Shute & Zapata-Rivera, 2012).

Digital learning environments, including intelligent tutoring systems (ITS), simulations, and games, create the conditions under which this vision becomes actionable. Such environments provide unprecedented access to fine-grained data about learning as it unfolds. Every interaction—what a learner does or says, when they do it, how they respond to constraints, and how their strategies evolve—leaves a digital trace. Stealth assessment (Shute, 2011) emerged from a deceptively simple question inspired by ECD's inferential logic: What if these traces could be systematically interpreted as evidence, transformed into valid, real-time inferences about learners' evolving cognitive and noncognitive states, and then used to support learning while it is still in progress?

From its earliest instantiations in student modeling for ITS (Shute & Psotka, 1996) to its current implementations in complex, game-based learning environments (e.g., Rahimi et al., 2024; Shute et al., 2020), stealth assessment has always depended on principled inference—specifically, on ECD (Mislevy et al., 2003; Mislevy et al., 2013). ECD supplies the connective tissue that links observable actions to latent competencies, ensuring that assessment claims are grounded in evidence rather than convenience. Our work on stealth assessment derives its coherence, credibility, and power from Mislevy's formulation of ECD, which serves as its conceptual and psychometric foundation. This chapter makes explicit how stealth assessment is built on ECD's foundations and argues that its future—particularly its promise to support learning—depends on preserving, refining, and extending those foundations.

**A Brief History of Stealth Assessment**

As discussed fully in Shute (2023), the roots of stealth assessment can be traced to late-1980s research in artificial intelligence (AI), education, and measurement. During this period, work on ITS explored how learners' actions could be used to infer knowledge states, guide instruction, and adapt content. Central to this work were student models, Bayesian inference networks, and the idea that assessment could be continuous rather than episodic.

Despite these conceptual advances, early systems were constrained by limited computing power and data infrastructure. Real-time data collection, analysis, and feedback were difficult to achieve, and many assessments remained intrusive or disruptive. As technologies advanced, however, the possibility of embedding assessment (mainly for formative rather than summative purposes; see Shute & Rahimi, 2017) seamlessly within learning activities became increasingly feasible.

The term stealth assessment was introduced circa 2005 to describe assessment that is unobtrusive, continuous, and deeply embedded within authentic activities. Early proof-of-concept work, supported by the Gates Foundation, involved the development of a custom-designed game—initially *Newton's Playground* (Shute et al., 2013) and later *Physics Playground* (Shute et al., 2019)—built explicitly to house stealth assessment machinery. Importantly, these efforts were not ad hoc. From the outset, they were grounded in ECD, with clearly articulated competency models, evidence models, and task models.

Using *Physics Playground*, we designed and validated stealth assessments of various competencies, such as physics understanding (Shute et al., 2020), creativity (Shute & Rahimi, 2020), persistence (Rahimi et al., 2021; Shute et al., 2015), and collaborative problem-solving skills (e.g., Sun et al., 2019). These studies, among others, demonstrated that stealth assessment could achieve acceptable reliability and

validity while remaining invisible to learners. Encouraged by these findings, other projects extended stealth assessment to commercial games such as *Portal 2* (Shute & Wang, 2015) and *Plants vs. Zombies 2* (Shute et al., 2016). These studies showed that ECD-based stealth assessments could be deployed within existing environments and still produce meaningful inferences about problem solving, spatial skills, and persistence. To achieve its formative assessment potential, we conducted design-based research testing various in-game learning supports that offer just-in-time explanations when students succeed and encouragement and instructional scaffolding when they struggle (e.g., Shute, Ke et al., 2019). Across these efforts, a consistent pattern emerged: whenever stealth assessment worked well, it was because ECD ensured that the stealth assessment could produce valid, reliable, and fair estimates of learners' competencies. In addition to the work that we have done, other researchers have adopted this method and used it in other fields (see Rahimi et al., 2023 for a systematic review on this topic; and see Rahimi, Shute, & Almond, 2026 for a summary of a special issue on stealth assessment).

## How Does Stealth Assessment Work?

Stealth assessment weaves ongoing formative assessments deeply and invisibly within digital learning environments. As learners interact with tasks, they naturally generate rich sequences of actions that are automatically captured in log files. These data are identified and scored using in-game rubrics/rules, aggregated in real time using statistical models (e.g., Bayesian networks), and used to update probabilistic estimates of learners' competencies. Stealth assessment is also intended to support learning and maintain flow—a state of optimal experience, where a person is so engaged in the activity at hand that self-consciousness disappears, sense of time is lost, and the person engages in complex, goal-directed activities not for external rewards, but simply for the exhilaration of doing (Csikszentmihalyi, 1990). The goal is to blur the distinction between assessment and learning.

Stealth assessment differs from simple learning analytics not just because it uses data, but because it is built on an inferential argument. Learners' observable actions only matter if they are intentionally designed to provide evidence about specific competencies. In a typical stealth assessment process, learners engage in tasks that elicit evidence; their actions are captured and scored using evidence rules; this evidence is combined using statistical models, often Bayesian networks; and learner models are updated to estimate current competencies. These Bayesian networks allow the system to reason *as if* a student has certain competencies which offers a useful, context-sensitive approximation of their complex socio-cognitive resources—rather than claiming to fully reveal their underlying cognitive structure. These evolving estimates can then support adaptive instruction, personalized learning supports, and meaningful feedback for both learners and instructors.

Without ECD, this pipeline would lack coherence. With ECD, it becomes a defensible assessment system grounded in explicit claims, designed evidence, and principled inference. Another way to understand stealth assessment is basically as the bridge between *emic* (i.e., the learner's internal cognitive and affective resources) and *etic* (i.e., the external linguistic, cultural, and substantive patterns and formal models of the domain). Stealth assessment then uses a person's observable behavior (interaction data) to make inferences about how their emic resource system is aligning—or not—with etic patterns. This is where just-in-time supports become especially powerful. They're triggered based on etic interpretations (e.g., evidence that a student is not applying Newton's second law force correctly), but they act directly on the learner's emic system and help them reorganize and refine their internal resources. Over time, these feedback loops help attune emic understanding to etic structure.

## ECD: The Backbone of Stealth Assessment

ECD was formulated to address a fundamental problem in assessment: how to reason systematically from observations to claims. ECD frames assessment as an evidentiary argument composed of interlocking models that specify what is to be measured, what evidence will support claims, and what tasks will elicit that evidence.

At its core, ECD consists of three key models that work together dynamically. The *competency model* (CM) specifies the knowledge, skills, and other personal attributes to be assessed. It defines the latent variables of interest and their relationships, providing the structure for student models. In stealth assessment, the competency model corresponds to the student model that is continuously updated as learners interact with the environment (see Rahimi, Smith et al., 2024 for CM development).

The *evidence model* (EM) defines how observable behaviors are identified, scored, and linked to competency variables. It includes two essential processes: evidence identification, which specifies what counts as observable evidence; and evidence accumulation, which specifies how that evidence updates beliefs about competency. Bayesian networks are frequently used in stealth assessment because they naturally support probabilistic inference under uncertainty and allow evidence to be combined across time and tasks.

The *task model* (TM) specifies the features of tasks that can elicit the desired evidence. In game-based environments, tasks often correspond to levels or challenges designed to evoke targeted behaviors (e.g., specific game mechanics, interactions, affordances). The TM ensures that activities are not merely engaging, but diagnostically meaningful.

The EM serves as the bridge between the TM and the CM, ensuring that tasks generate evidence that can be interpreted in terms of claims about learners. Collectively, these models form the inferential engine of stealth assessment. These models can be mapped directly to Toulmin's Argument Model (Toulmin, 1958) where the CM structures claims, the TM generates data, and the EM acts as probabilistic warrants. The goal in Toulmin's Argument Model is to provide strong evidence that justifies the connection between data and claims. The more accurate (i.e., valid) the warrant—representing the inferences drawn from the EM—the less plausible alternative explanations (i.e., other reasons for what the learner knows or can do) become. Stealth assessment seeks to strengthen these warrants and reduce alternative explanations using ecologically valid tasks. In doing so, it supports a *conditional sense of fairness* (Mislevy et al., 2013), whereby interpretations are continuously refined to minimize construct-irrelevant factors and ensure that observed performance reflects intended competencies.

## Why ECD Is the *Sine Qua Non* of Stealth Assessment

Stealth assessment often operates in environments that are rich with data but poor in structure. Digital games, in particular, can generate vast quantities of interaction data, but data alone do not constitute evidence. ECD provides the structure that transforms raw data into meaningful evidence.

Moreover, ECD ensures that claims about learning are explicitly defined, that evidence is designed rather than discovered *post hoc*, and that inferences are transparent and justifiable. Without ECD, stealth assessment risks devolving into opportunistic data mining, where patterns are detected but

their meaning is unclear and their validity questionable. With ECD, stealth assessment becomes principled measurement.

This distinction is not merely theoretical. In practice, ECD constrains design decisions, guides the selection of indicators, and anchors validation efforts. It ensures that stealth assessment systems remain accountable to the constructs they claim to measure, even as they operate invisibly within complex environments.

In an ECD-based stealth assessment system, actions (e.g., gameplay) and assessment are inseparable. Tasks generated from the TM elicit behaviors that are scored according to the EM and used to update the CM in real time. As learners continue to interact with the learning environment (e.g., game), the system refines its estimates, which can be used to enable adaptive task selection and personalized feedback.

Crucially, this adaptivity is driven not by surface performance alone, such as success or failure, but by inferences about underlying competencies at play for learners' observable actions. A learner who fails a task may do so for different reasons—misconceptions, lack of persistence, or inefficient strategies—and ECD-based models allow these differences to be distinguished. This capacity to diagnose learning processes rather than merely outcomes is central to the value of stealth assessment. Next, we illustrate stealth assessment related to understanding Newtonian physics.

**An Extended Example: Physics Understanding in *Physics Playground***

*Physics Playground* (Shute et al., 2019) provides a concrete illustration of how ECD structures stealth assessment from conception through validation. The game challenges learners to solve physics-based puzzles by drawing and manipulating objects and parameters to achieve a simple goal: hitting a target balloon using a green ball with principles of force and motion.

The CM for physics understanding was developed in collaboration with subject-matter experts and organized hierarchically. At the top level sat overall physics understanding, supported by mid-level constructs such as force and motion, energy, linear momentum, and torque. These, in turn, were linked to specific laws and principles, including Newton's three laws of force and motion.

The TM included two primary task/level types: *sketching* levels, in which learners drew objects on the screen of their devices to create simple machines, and *manipulation* levels, in which learners adjusted parameters such as mass, gravity, and air resistance, to solve problems. A Q-matrix was used to map tasks to competencies and ensure adequate coverage across competencies and difficulty levels. This explicit alignment exemplifies the discipline imposed by ECD.

The EM specified in-game indicators associated with targeted physics understanding competencies. Learners' actions were scored and accumulated using Bayesian networks, which then produced probabilistic estimates of mastery at multiple grain sizes. These estimates were updated continuously as learners progressed through the game. Note that this moment-by-moment capture and scoring of in-game actions maps onto Mislevy's concept of "evidence-bearing opportunities," discussed in Chapter 16, this volume (VERIFY CHAPTER/VOLUME).

Validation focused on face, content, and convergent validity. The face and content validity were conducted before collecting data and during the design process. This approach is also informed by ECD practices. That is, learning science and physics experts continuously validated the content, the game mechanics, and the assessment design. Convergent validity, done after data collection, examined correlations between stealth assessment estimates and relevant external physics tests (e.g.,

Force Concept Inventory; Hestenes, Wells, & Swackhamer 1992). Significant correlations at both overall and sub-construct levels demonstrated that the inferences supported by the ECD-based models were meaningful and defensible. Stealth assessment estimates were then used to drive adaptivity and learning supports. When competency estimates were low, targeted feedback or instructional resources were delivered; when they were high, learners progressed to more challenging tasks. Student-facing dashboards further increased transparency and supported self-regulated learning as students could focus on game levels with low estimates.

## Looking Ahead: The Future of ECD-Based Stealth Assessment

The future of stealth assessment lies not in abandoning ECD, but in extending it. As technology-rich learning environments become more widespread, the importance of assessment approaches that are theory-driven, psychometrically sound, and unobtrusive will continue to grow. Stealth assessment is particularly well-positioned for large-scale adoption because it enables valid inference while preserving learner engagement and minimizing test-like interruptions. Accordingly, we expect stealth assessment to be increasingly implemented at scale across digital learning platforms.

A central direction for this evolution is the integration of theory-driven ECD models with data-driven methods, including machine learning and generative AI. These methods can strengthen internal processes of stealth assessment—particularly evidence identification and scoring—by supporting the interpretation of complex learner behaviors and open-ended products. For example, rubric-based prompt engineering offers a structured way to extract and evaluate evidence in alignment with performance expectations (i.e., competencies and the relevant claims), which is especially useful when learners generate multimodal artifacts (e.g., written explanations, designs, diagrams, audio, or interactive creations). In addition, generative AI can support the design of tasks intended to elicit targeted evidence, accelerating the development of new stealth assessment activities while maintaining alignment with ECD principles.

At the same time, because ECD-based assessment development is often resource intensive, another promising opportunity involves reusing and partially automating ECD models to reduce development time and cost. Early efforts toward automation (e.g., Min et al., 2015) suggest that key components of the ECD pipeline (e.g., evidence specification, scoring rules, and model refinement) can increasingly be supported by computational methods.

Crucially, in each of these directions—whether expanding evidence sources to include multimodal data, incorporating machine learning, or scaling to new competencies and contexts—ECD remains essential. It provides the validity-oriented structure needed to ensure that new data streams and analytic techniques contribute to meaningful inference rather than measurement noise, and that assessment remains interpretable, defensible, and aligned with intended learning constructs.

## Conclusion

Stealth assessment represents a shift from episodic testing to continuous, deeply embedded assessment in service of learning. Its promise lies in supporting learning while simultaneously assessing it—and that promise is best understood by grounding personalized learning supports in learners' evolving resources and their alignment with linguistic, cultural, and substantive patterns in a domain. Within this framing, environments like *Physics Playground* use just-in-time explanations as feedback loops that help students refine and attune their cognitive resources to the disciplinary patterns of Newtonian physics. See Chapters 17 and 2 (VERIFY THESE CHAPTERS) for more on these disciplinary patterns and actionable feedback. This interplay between learner resources and

disciplinary patterns is made visible through stealth assessment because it rests on ECD. ECD provides the inferential backbone that transforms interaction data into defensible claims about how learners build and coordinate resources in relation to these domain-specific patterns. It disciplines design, guides validation, and anchors adaptivity in theory. As educational technologies evolve, preserving this evidence-centered foundation is critical: without ECD, stealth assessment would be stealth in name only; with it, a principled approach to assessment for learning.

We, like many scholars in the field, are deeply indebted to Bob Mislevy for his foundational scholarship and enduring leadership in educational measurement.

## References


Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience.* Harper Perennial.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158. https://doi.org/10.1119/1.2343497

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., & Lester, J. C. (2015). DeepStealth: Leveraging deep learning models for stealth assessment in game-based learning environments. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial intelligence in education* (pp. 277–286). Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_28

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., et al. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation, 19*(2–3), 121–140.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. https://doi.org/10.4324/9781315871691

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.

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., 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. J., Khodabandelou, R., Kuba, R., Babaee, M., & Esmaeiligoujar, S. (2023). Stealth assessment: A systematic review of the literature. In *Proceedings of the 17th International Conference of the Learning Sciences (ICLS 2023)* (pp. 1977–1978). International Society of the Learning Sciences.

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

Rahimi, S., Shute, V. J., & 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. https://doi.org/10.46328/ijtes.118

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 Publishing.

Shute, V. J. (2023). History of stealth assessment and a peek into its future. In M. P. McCreery & S. K. Krach (Eds.), *Games as stealth assessments* (pp. 1–20). IGI Global.

Shute, V. J., Almond, R., & Rahimi, S. (2019). *Physics Playground* (Version 1.3) [Computer software]. https://pluto.coe.fsu.edu/ppteam/pp-links/

Shute, V. J., Ke, F., Almond, R. G., Rahimi, S., Smith, G., & Lu, X. (2019). How to increase learning while not decreasing the fun in educational games. In R. Feldman (Ed.), *Learning science: Theory, research, and practice* (pp. 327-357). McGraw-Hill.

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., ... Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education, 86*, 224–235.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 570–600). Macmillan.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12172

Shute, V. J., & Rahimi, S. (2020). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior, 116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in educational games. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12473

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., & Wang, L. (2015). Measuring problem solving skills in Portal 2. In *E-learning systems, environments and approaches: Theory and implementation* (pp. 11–24). Springer International Publishing.

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

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach (Ed.), *Adaptive technologies for training and education* (pp. 7-27). Cambridge University Press.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2019). Towards a generalized competency model of collaborative problem solving. *Computers & Education, 143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.