Stealth Assessment Implementation Using Evidence-Centered Design Frameworks

Stealth assessment represents a paradigm shift in educational measurement—embedding rigorous assessment directly into gameplay so students demonstrate competencies without experiencing test anxiety or interruption. The Evidence-Centered Design (ECD) framework, (Fsu) developed by Robert Mislevy, Russell Almond, and colleagues at the Educational Testing Service beginning in the late 1990s, (Circlcenter) provides the theoretical and technical architecture for implementing these invisible assessments. Research from Valerie Shute's Physics Playground project demonstrates that well-implemented stealth assessments achieve **significant correlations with external measures** (r = .52-.66) while maintaining student engagement, with validation studies showing learning gains of **d = 0.61** compared to control groups.

The core insight of ECD is that all assessments require an evidentiary argument linking observable behaviors to claims about competency. (ed) In game-based contexts, this argument must be constructed before development, with every game mechanic, logged action, and feedback loop designed to generate interpretable evidence. This report provides technical implementation details across five interconnected domains: the ECD framework architecture, Bayesian network construction for competency modeling, validated behavioral indicators from Physics Playground and the ENGAGE platform, log data feature engineering, and real-time competency estimation algorithms.

The five-model ECD architecture structures assessment design

The Conceptual Assessment Framework (CAF) comprises five principal design objects that together specify what is measured, how evidence is gathered, and how inferences are drawn. Mislevy, Steinberg, and Almond's foundational 2003 paper "On the Structure of Educational Assessments" established this architecture, (Circlcenter) which has since become the standard for complex assessment design.

The **Student Model** (also termed Proficiency Model) defines the latent competencies being assessed. (PubMed Central) These variables are unobservable—we express what we know about them as probability distributions updated by evidence. (fsu) (ed) In Physics Playground, the student model comprises nine physics competencies ordered by conceptual difficulty: Newton's First Law (easiest), energy transfer, energy dissipation, properties of momentum, conservation of momentum, properties of torque, equilibrium, Newton's Second Law, and Newton's Third Law (most difficult). Variables can be continuous (IRT-style proficiency θ) or discrete categorical levels (low/medium/high), and multiple variables can form networks showing associations among aspects of knowledge. (NCBI)

The **Evidence Model** provides instructions for interpreting student work products and contains two essential components. Evidence Rules specify how observable variables summarize performance— (ResearchGate) (ERIC) in games, this means classifying actions like "student draws a lever to solve the level" as evidence for torque/equilibrium competency. The Statistical Model then defines the probabilistic relationship between observables and competencies (ERIC) using structures like Bayesian networks. (fsu) As Mislevy noted, "the

Evidence Model is the heart of ECD, because it provides the credible argument for how students' behaviors constitute evidence about targeted aspects of proficiency."

The **Task Model** describes how to structure situations that elicit needed evidence. Crucially, a task model represents a family of potential tasks, not a single task. (ed) In Physics Playground, task models define game levels as either sketching levels (where players draw simple machines) or manipulation levels (where players adjust physics parameters). (Springer) Task Model Variables include difficulty indices for game mechanics (GM: 1-5) and physics understanding (PU: 1-5), plus specifications for which competencies each level addresses via a Q-matrix structure.

The **Assembly Model** specifies how Student, Evidence, and Task Models work together, determining targets for measurement accuracy, constraints for content coverage, and the number and mix of tasks needed for reliability. (ed) In game-based assessment, this model faces unique challenges—the desire to gain specific evidence must be balanced with the internal logic of gameplay without disrupting player flow. (fsu)

The **Presentation Model** describes how tasks are rendered in physical or electronic environments. For games, this includes the full rules and mechanics of the simulation. (fsu) Almond and colleagues emphasize that "there needs to be a compelling need for the player to produce the evidence needed for the assessment that does not interrupt the flow of the game." (fsu)

Bayesian networks enable probabilistic competency inference

Educational Bayesian networks employ a hierarchical three-layer architecture connecting latent competencies to observable actions. **Competency nodes** represent unobservable knowledge, skills, and abilities—typically discrete with ordered states like {low, medium, high}. **Evidence/intermediate nodes** bridge competencies and observables, including context variables that absorb domain-irrelevant covariance. **Observable action nodes** are directly measurable indicators from gameplay: object manipulations, response patterns, tool usage, and solution strategies.

The directed acyclic graph (DAG) structure enables factorization of the joint probability distribution. For a network with competency variables A, B, C and observable variables X_1, X_2, the joint distribution factors as: $\mathbf{P(A,B,C,X_1,X_2) = P(A)P(B|A)P(C|A)P(X_1|B)P(X_2|C)}$. This factorization exploits conditional independence to make inference computationally tractable.

Conditional probability table construction methods

The DiBello-Samejima method (also called θ-projection) combines structural expert judgments with Item Response Theory to populate CPTs. (mlr) The core formula uses Samejima's Graded Response Model: (fsu)

$$\mathbf{P(Xs,m \geq xk \mid \theta s) = logit^{-1}(am(\theta s + bm,k))}$$

where θs represents latent response propensity, am is the discrimination parameter (typically 0.3-3.0 for game-based assessment), and bm,k are difficulty parameters for K ordered categories. (mlr) For multivariate competency footprints, a projection function combines parent node effects:

$\tilde{\theta}_{s,m} = \Sigma_j (c_{m,j} \cdot S_{m,j} + d_{m,j})$ (compensatory structure)

$\tilde{\theta}_{s,m} = \min(\tilde{\theta}_{s,m,1}, \tilde{\theta}_{s,m,2}, ..., \tilde{\theta}_{s,m,J})$ (conjunctive structure)

$\tilde{\theta}_{s,m} = \max(\tilde{\theta}_{s,m,1}, \tilde{\theta}_{s,m,2}, ..., \tilde{\theta}_{s,m,J})$ (disjunctive structure) (mlr)

Expert elicitation focuses on four key parameters: the structure relationship (compensatory vs. conjunctive vs. disjunctive), task difficulty rating (mapped to d = +1/0/-1), conditional dependence among observables, and relative importance of skills. (mlr) The **Noisy-OR gate** models disjunctive relationships: $P(Y=0|X_1=x_1,...,X_n=x_n) = p_{0,0} \cdot \prod_i p_{i,0}^{x_i}$, where $p_{i,0}$ is the inhibition probability for cause i. The **Noisy-AND gate** models conjunctive relationships where all skills are necessary.

Dynamic Bayesian networks for temporal learning

DBNs extend static networks with time-slice representations to model learning progression:

$P(X_{1:T}, Z_{1:T}) = P(Z_1)P(X_1|Z_1) \cdot \prod_{t=2}^{T} P(Z_t|Z_{t-1})P(X_t|Z_t)$

The transition model $P(Z_t|Z_{t-1})$ captures learning dynamics. The classic Bayesian Knowledge Tracing model assumes $P(Z_t = \text{learned} \mid Z_{t-1} = \text{learned}) = 1$ (no forgetting), though relaxed models allow $P(Z_t = \text{unlearned} \mid Z_{t-1} = \text{learned}) = p_{forget}$. (PubMed Central)

For computational efficiency, the **rollup procedure** computes the posterior $P(S_t|X_{1:t})$ after each time slice, uses this as the prior $P(S_{t+1})$ for the next slice, then discards old evidence nodes. This maintains constant memory while preserving inferential validity.

Physics Playground validates behavioral indicators for physics competencies

Physics Playground (formerly Newton's Playground), developed by Valerie Shute at Florida State University, represents the most thoroughly validated implementation of game-based stealth assessment. (fsu) (Fsu) Players guide a green ball to hit a red balloon by drawing simple machines—ramps, levers, pendulums, and springboards—or manipulating physics parameters like gravity, mass, and air resistance. (ceur-ws) (fsu) The game automatically identifies drawn agents with **>95% accuracy** compared to human raters. (ceur-ws)

Validated behavioral indicators and evidence rules

Each simple machine provides evidence for specific competencies through validated mappings:

Ramps (inclined planes) provide evidence for energy transfer (potential → kinetic) and Newton's First Law. Drawing a ramp that correctly guides the ball demonstrates understanding of how objects in motion continue unless acted upon by external forces.

Levers provide evidence for torque and equilibrium competencies. Correct lever placement demonstrates understanding of rotational mechanics and force multiplication around pivot points.

Pendulums provide evidence for momentum properties. Using pendulums to direct impulse tangent to the direction of motion shows understanding of force application and energy transfer.

Springboards provide evidence for elastic potential energy and energy transfer. Storing energy from a falling weight and releasing it to move the ball vertically demonstrates understanding of energy conservation.

The game logs level entrance/exit times, time spent interacting, number of objects created, number of restarts, types of agents used, whether the level was solved, and badge type earned (gold for "under par" solutions, silver otherwise). (fsu) (Fsu) Gold badges indicate mastery of the relevant physics concept.

Gaming-the-system detection

Three problematic behaviors are automatically detected: **stacking** (drawing consecutive short lines beneath the ball to lift it), **breaking the system** (drawing random lines until the physics engine crashes), and **cutting corners** (drawing a quick line beneath a moving ball spanning to the balloon). These behaviors exploit the system without demonstrating physics understanding and are filtered from evidence accumulation.

Empirical validation results

Shute and colleagues conducted multiple validation studies. In the primary 2013 study with **167 eighth- and ninth-grade students**, significant pretest-posttest physics gains emerged alongside significant correlations between in-game indicators and learning. (Taylor & Francis Online) A 2019 study with **263 high school students** showed significant pretest-posttest improvements ($F(1, 198) = 9.53$, $p < .01$) while the control group showed no gains ($F(1, 63) = 0.002$, $p = .97$). (Springer) Test reliability reached $\alpha = .77$ (pretest) and $\alpha = .82$ (posttest). (ResearchGate)

Persistence validation compared stealth assessment measures (average time on unsolved problems + revisits to unsolved problems) against external measures. Self-report persistence showed no correlation with stealth measures, but the Performance-Based Measure of Persistence correlated significantly: **$r = 0.51$ ($p < .01$)** for low performers and **$r = 0.22$ ($p < .05$)** for high performers, demonstrating that behavioral indicators outperform self-report for measuring persistence. (fsu)

Log data feature engineering extracts predictive signals from gameplay

Game telemetry generates high-volume, high-velocity data requiring systematic feature engineering to extract meaningful competency indicators. The Mission HydroSci project developed a multi-layer pipeline achieving **82-94% classification accuracy** across different learning units. (educationaldatamining)

Taxonomy of telemetry features

Frequency features capture raw counts and rates: action frequencies (number of hint requests), rate metrics (clicks per minute), proportions (percentage of time in exploration vs. problem-solving), and event ratios (correct/incorrect action ratio).

Temporal features capture time dynamics: time-on-task (duration on specific activities), latencies (time between stimulus and response), pacing metrics (inter-event intervals), and moving averages (exponentially-weighted moving averages for trend detection).

Sequence features capture action patterns: n-grams (bigram/trigram action subsequence frequencies), transition matrices (probability of state transitions via Markov chain modeling), edit distance (Levenshtein distance to expert solution paths), and navigation trajectories through game space.

Derived features capture higher-order constructs: efficiency ratios (correct actions / total actions), persistence metrics (attempts after initial failure), exploration indices (unique locations / total locations), and help-seeking patterns (hints used before correct answer).

Multi-layer dimension reduction pipeline

The Mission HydroSci approach applies three layers of unsupervised learning. **Layer 1** applies dimension reduction (PCA, SVD, ICA, NMF, t-SNE, UMAP, or autoencoders) within behavior types to extract principal components. **Layer 2** performs factor analysis across feature types within behavior categories to capture relationships between frequency, speed, and share features. **Layer 3** integrates across all behavior types to generate holistic behavioral constructs.

Validated predictive features

For scientific argumentation assessment, the highest-contributing feature categories were argumentation system interactions (**28% contribution**), dialogue reading behaviors (**25%**), and task completion patterns (**22%**). For content knowledge prediction, dialogue-related behaviors contributed **27-48%** across units, item interaction behaviors **18-29%**, and task completion behaviors **13%**.

Handling sparse and noisy data requires SMOTE (Synthetic Minority Oversampling) for class imbalance, Tomek Links to remove overlapping samples, temporal smoothing for irregular data, and outlier exclusion for anomalous sessions. Random Forest consistently performs well across studies, with ensemble methods improving robustness and cross-game generalizability.

Real-time algorithms update competency beliefs during gameplay

Stealth assessment requires sub-second belief updates to maintain real-time competency estimates without disrupting gameplay. The fundamental Bayesian update follows $\mathbf{P(\theta|X) \propto P(X|\theta)P(\theta)}$, where evidence X updates prior beliefs $P(\theta)$ to posterior $P(\theta|X)$.

Bayesian Knowledge Tracing update equations

The classic BKT model uses four parameters: $P(L_0)$ for initial knowledge probability, $P(T)$ for learning/transition probability, $P(G)$ for guess probability, and $P(S)$ for slip probability. The update equations are:

$$\mathbf{P(L_n|correct) = [P(L_{n-1})(1-P(S))] / [P(L_{n-1})(1-P(S)) + (1-P(L_{n-1}))P(G)]}$$

$$\mathbf{P(L_n|incorrect) = [P(L_{n-1})P(S)] / [P(L_{n-1})P(S) + (1-P(L_{n-1}))(1-P(G))]}$$

$$\mathbf{P(L_n) = P(L_n|obs) + (1-P(L_n|obs))P(T)}$$

BKT achieves sub-millisecond updates with O(n) complexity for n observations per skill.

Deep Knowledge Tracing (DKT) uses LSTM architecture: $\mathbf{h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h)}$ and $\mathbf{y_t = \sigma(W_{\gamma h}h_t + b_\gamma)}$, where x_t encodes student actions and y_t outputs probability of correctness for each skill. DKT requires no expert annotation and captures complex skill dependencies, with inference latency of 1-10ms per prediction.

Computational optimization strategies

Pre-computation dramatically reduces runtime: conditional probability tables are cached, Q-matrices define task-competency mappings in advance, and KL distances between cognitive pattern pairs are pre-calculated before assessment. **Model simplification** removes weak dependencies, applies arc removal for sparse approximations, and uses state-space abstraction to reduce variable cardinalities. **Approximation methods** include mini-bucket partitioning with adjustable accuracy, variational inference to delink nodes, and importance sampling (AIS-BN) for stochastic estimation.

Anytime algorithms produce valid results even when interrupted: bounded conditioning processes high-probability configurations first, incremental probabilistic inference computes larger terms first, and stochastic sampling improves precision with additional samples.

Mastery threshold setting

Common threshold values include **0.70** for minimal competency, **0.80** for standard mastery (typical in CAT systems), **0.90** for high-confidence decisions, and **0.95** for critical applications. The confidence interval approach computes 95% CI after each item: $[\theta - 1.96\times SE(\theta), \theta + 1.96\times SE(\theta)]$. If the interval does not contain the cut-score, classification proceeds; otherwise testing continues.

For cognitive diagnostic testing using Expected A Posteriori estimation, threshold criteria vary by required precision: $P(\text{mastery}|\text{responses}) > 0.65$ yields ±0.23 precision, > 0.75 yields ±0.19, > 0.85 yields ±0.13, and > 0.95 yields ±0.05 precision.

Adaptive feedback integration

The GRADES framework (Game-based Reinforcement Learning Adaptive Difficulty and Evaluation with Stealth assessment) implements three layers: a data collection layer capturing continuous performance metrics, a stealth assessment integration layer for real-time cognitive engagement assessment, and an adaptive decision layer using reinforcement learning for difficulty adjustments.

The RL formulation uses state s (current player metrics), action a (difficulty adjustment), and reward r (engagement/learning outcomes): $\mathbf{Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]}$. SARSA implementations adapt to each player during gameplay, typically targeting ~50% win/loss balance to maintain flow state.

Conclusion

Stealth assessment using Evidence-Centered Design represents a mature methodology with validated implementations across multiple educational domains. The key technical requirements include careful upfront

design linking game mechanics to competency evidence, Bayesian networks calibrated through either expert elicitation or data-driven parameter estimation, comprehensive log data infrastructure capturing actions, timing, and sequences, and computationally efficient real-time inference algorithms.

Physics Playground demonstrates that well-designed stealth assessments achieve meaningful correlations with external measures while maintaining engagement and producing learning gains superior to control conditions. (ERIC) The critical insight is that discrimination parameters must be set conservatively (around 0.3) due to the many confounds in game-based assessment—each piece of evidence is weaker than in traditional testing, but the volume of evidence accumulated during gameplay compensates. (fsu)

Future implementations should leverage the multi-layer feature engineering pipelines validated in Mission HydroSci, apply ensemble machine learning methods alongside traditional Bayesian networks, and implement adaptive feedback loops that adjust difficulty in real-time based on competency estimates. The 10-step development process established by Shute's team—from competency model development through Bayesian network calibration to external validation—provides a proven roadmap for new stealth assessment initiatives.