# Non-Question Proficiency Evaluation Framework

Methodology Comparison and Ranking Report

Research Analysis Team

January 8, 2026

**Abstract**

This report provides a systematic evaluation of thirteen methodologies for estimating student proficiency based solely on non-question engagement data. Drawing from heuristic, probabilistic, and machine learning domains, we analyze feasibility, predictive accuracy, and implementation requirements. The analysis confirms that while proficiency estimation from engagement is an "estimation problem under uncertainty," advanced models like Deep Knowledge Tracing (DKT) and Stealth Assessment (SPRING) offer moderate to strong correlations ($r = 0.40 - 0.65$) with actual learning outcomes.

# Contents

# 1 Executive Summary

## 1.1 Overview

The Non-Question Proficiency Evaluation Framework aims to determine the feasibility and best methods for estimating learner proficiency without direct assessment questions. Based on an analysis of current research and platform practices, estimating proficiency from non-question activities (reading, watching videos, simulations) is feasible but carries inherent uncertainty. It relies on interpreting "rich engagement data" through probabilistic models rather than binary right/wrong signals.

## 1.2 Key Findings

Our analysis of thirteen distinct methods reveals three primary tiers of estimation capability:

- **High Precision (ML-Based):** Methods like **Deep Knowledge Tracing (DKT)** and **Stealth Assessment** provide the highest predictive power ($r \approx 0.55$), acting like a "satellite" to analyze subtle terrain changes.
- **Structural Estimation (Model-Based):** Approaches like **Half-Life Regression (HLR)** and **Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)** offer grounded estimates based on cognitive theories of forgetting and learning curves.
- **Baseline Tracking (Heuristic):** Systems like **XP Models** and **Time-on-Task** serve well for gamification but are considered "lower-confidence evidence" for actual proficiency.

## 1.3 Recommendations

For the NEXS-399 project, we recommend a hybrid implementation strategy tailored to content type:

- **Video Content:** **Cognitive Load Proxy Model (CLPM)** and **Video Engagement Analytics**, leveraging data on pauses and rewinds which correlate with performance.
- **Text/PDF Content:** **Half-Life Regression (HLR)**, utilizing spacing effects to model memory decay over time.
- **Interactive Content:** **Stealth Assessment (SPRING)**, which has demonstrated the ability to predict test outcomes from game logs.

# 2 Introduction

## 2.1 Problem Statement

Traditional education relies heavily on direct questioning (quizzes, exams) to measure proficiency. However, a significant portion of learning occurs during passive or semi-active engagement—reading, watching, and exploring—where no direct questions are asked. There is a need to quantify the proficiency gained during these non-question activities to provide a continuous, holistic view of learner progress.

## 2.2 Feasibility Foundation

According to the source material, estimating proficiency from non-question activities is feasible. Modern frameworks suggest that every digital interaction signals a learning state. However, because passive study is less predictive than active practice, these estimates must be treated as "lower-confidence evidence" until validated.

Empirical evidence supports this feasibility:

- **Engagement Correlations:** Higher engagement (time, completion, re-engagement) consistently correlates with better test scores.
- **Predictive Accuracy:** Engagement-based estimation achieves moderate to strong correlations with assessment outcomes, typically ranging from $r = 0.40$ to 0.65.
- **Behavioral Indicators:** Fine-grained actions, such as pausing and rewinding videos, are positively correlated with higher exam performance.

## 2.3 Scope

This report compares thirteen specific methods identified in the source text, categorized into:

1. **Heuristic-Based Methods:** Rules and point systems.
2. **Model-Based Methods:** Probabilistic and statistical theories.
3. **Machine Learning (ML)-Based Methods:** Data-driven deep learning and classification.

# 3 Comprehensive Methodology Review

This section details the thirteen methods identified for estimating proficiency from engagement data.

## 3.1 Category 1: Heuristic-Based Methods

These methods rely on pre-defined rules and expert judgment. They are often used for immediate feedback and gamification rather than deep statistical inference.

### 3.1.1 1. Heuristic Point Systems (XP Models)

Assigns experience points or progress percentages for completing content. Major platforms like Duolingo (XP) and Khan Academy (energy points) use this for motivation.

- **Status:** Used primarily for gamification; typically not a direct measure of mastery.

### 3.1.2 2. Time-on-Task & Completion Metrics

Uses normalized time spent and completion rates (e.g., percentage of a video watched) as direct predictors of success.

- **Basis:** Khan Academy research notes that every additional minute spent is associated with gains on standardized tests.

### 3.1.3 3. Mastery-Based Engagement Scoring (MBES)

A threshold-based system where proficiency levels are assigned based on triggers.

- **Levels:** Attempted, Familiar, Proficient, Mastered.

### 3.1.4 4. Time-Weighted Completion Model (TWCM)

A simple model that weights content completion by the quality of time spent relative to expected duration, attempting to filter out passive or unengaged time.

## 3.2 Category 2: Model-Based (Probabilistic) Methods

These methods use established psychological or cognitive theories to model how knowledge is acquired or forgotten.

### 3.2.1 5. Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)

An extension of standard BKT. It treats content interactions as "learning opportunities," using signals like completion and interaction density to modify the probability that a student has transitioned from an unlearned to a learned state.

### 3.2.2 6. Half-Life Regression (HLR)

Combines the Ebbinghaus forgetting curve with engagement data. It estimates the "strength" of a learner's memory and predicts the probability of recall over time.

- **Evidence:** Duolingo uses this to model forgetting curves, improving daily user retention by 12%.

#### 3.2.3 7. Performance Factors Analysis (PFA) with Engagement Covariates

A logistic regression model that incorporates both a student's prior performance and current engagement factors to predict proficiency.

#### 3.2.4 8. Item Response Theory (IRT) Analogy

Adapts traditional IRT by treating content pieces as "items" with specific difficulty levels. Future success on related questions is used to validate the proficiency gained from that content.

#### 3.2.5 9. Cognitive Load Proxy Model (CLPM)

Estimates the "germane load" (productive learning effort) versus "extraneous load" by analyzing engagement patterns.

- **Key Indicators:** Video pauses and rewinds.

### 3.3 Category 3: Machine Learning (ML)-Based Methods

Data-driven approaches that learn complex mappings between behavioral logs and outcomes.

#### 3.3.1 10. Deep Knowledge Tracing (DKT) with Engagement Features

Uses Recurrent Neural Networks (RNNs) or Transformers to process sequences of student interactions (including non-question data) to predict future performance.

- **Recent Advances:** Integrating DKT with cognitive load estimation has led to more efficient personalized learning paths (Tong & Ren, 2025).

#### 3.3.2 11. Stealth Assessment (SPRING)

Pearson's SPRING (Student PRoficiency INferrer from Game data) uses Evidence-Centered Design (ECD) to infer proficiency from action sequences and game logs.

- **Accuracy:** Demonstrated a correlation of approximately 0.55 with test outcomes.

#### 3.3.3 12. Multi-Modal Attention Models (MMAE)

Combines disparate signals—such as scroll depth, video playback speed changes, and session frequency—to infer the quality of attention and subsequent learning.

#### 3.3.4 13. Regression/Classification Predictive Models

Direct models (Random Forests, Logistic Regression) trained to predict final exam scores or mastery states based on "clickstream" features.

- **Utility:** Research in MOOCs shows utilization rates in the first week can predict passing probabilities.

# 4 Systematic Ranking and Evaluation

The following ranking is derived from the evidence provided in the source text regarding predictive power, empirical support, and complexity.

## 4.1 Ranking Criteria

- **Evidence (Evid):** Strength of empirical backing in the source text.
- **Accuracy (Acc):** Reported correlation or predictive capability.
- **Depth (Dept):** Ability to model complex learning states (satellite vs. steps).
- **Utility (Util):** Practical application for personalization or retention.

## 4.2 Methodology Ranking Matrix

| Method | Cat | Acc | Evid | Util | Overall Tier |
|---|---|---|---|---|---|
| **1. Stealth Assessment (SPRING)** *Rationale: Proven correlation ($r \approx 0.55$), validated by Pearson.* | ML | High | High | High | **Tier 1** |
| **2. Deep Knowledge Tracing (DKT)** *Rationale: State-of-the-art sequencing, capable of analyzing terrain changes.* | ML | High | Med | High | **Tier 1** |
| **3. Half-Life Regression (HLR)** *Rationale: Proven 12% retention boost at Duolingo; strong theoretical basis.* | Model | Med | High | High | **Tier 1** |
| **4. Cog. Load Proxy Model (CLPM)** *Rationale: Links behavior (pausing) to learning processing (germane load).* | Model | Med | Med | Med | **Tier 2** |
| **5. EW-BKT** *Rationale: Extension of the widely cited BKT (Corbett & Anderson).* | Model | Med | High | Med | **Tier 2** |
| **6. Reg/Class Predictive Models** *Rationale: Effective for early intervention/at-risk flagging in MOOCs.* | ML | Med | High | Med | **Tier 2** |

| | | | | | |
|---|---|---|---|---|---|
| **7. Multi-Modal Attention (MMAE)**<br>*Rationale: Rich data usage (scroll/speed), but higher complexity.* | ML | Med | Low | Med | **Tier 2** |
| **8. Time-on-Task**<br>*Rationale: Correlates with gains (Khan Academy) but lacks granularity.* | Heur | Low | High | Low | **Tier 3** |
| **9. PFA with Engagement**<br>*Rationale: Good statistical basis, but less emphasized for pure engagement.* | Model | Med | Med | Low | **Tier 3** |
| **10. IRT Analogy**<br>*Rationale: Theoretical adaptation, less direct evidence provided.* | Model | Low | Low | Low | **Tier 3** |
| **11. Time-Weighted Comp. (TWCM)**<br>*Rationale: Refinement of time metrics, but still a heuristic.* | Heur | Low | Low | Low | **Tier 4** |
| **12. Mastery-Based Score (MBES)**<br>*Rationale: Threshold-based, lacks probabilistic nuance.* | Heur | Low | Low | Low | **Tier 4** |
| **13. Heuristic Point Systems**<br>*Rationale: Gamification tool; not a reliable proof of proficiency.* | Heur | Low | High | Low | **Tier 4** |

### 4.3 Ranking Summary

- **Tier 1 (Best in Class):** Stealth Assessment, DKT, and HLR. These methods move beyond simple tracking to actual prediction of memory strength and test outcomes.
- **Tier 2 (Strong Contenders):** CLPM and EW-BKT. These connect behavioral signals (pausing, rewinding) to cognitive states.
- **Tier 3 & 4 (Baseline):** Heuristics and Point Systems. Useful for motivation ("counting steps") but insufficient for accurate proficiency estimation without validation.

# 5 Comparative Analysis

## 5.1 Strengths and Weaknesses by Category

| Category | Strengths | Weaknesses |
|---|---|---|
| **Heuristic** | - Simple to implement<br>- Immediate feedback for users<br>- High motivational value (Gamification) | - Low predictive accuracy<br>- "Lower-confidence evidence"<br>- Watching videos doesn't ensure mastery |
| **Model-Based** | - Grounded in cognitive theory (Forgetting curves)<br>- Can model memory strength over time<br>- Balances load estimation | - Requires calibration of parameters<br>- Assumes valid theoretical mapping (e.g., meaningful interactions) |
| **ML-Based** | - Highest predictive accuracy ($r = 0.65$)<br>- Captures non-linear patterns<br>- Validated success (SPRING) | - High complexity ("Satellite" approach)<br>- Requires large datasets (Big Data)<br>- "Black box" nature of Deep Learning |

Table 2: Comparative Analysis of Methodology Categories

## 5.2 The Analogy of Estimation

The source text provides a critical analogy for comparing these methods:

> "Estimating proficiency from content engagement is like tracking a hiker's progress... **Heuristics** are like counting the steps taken; **Model-based** approaches are like using a map and known walking speeds to estimate location; and **ML-based** approaches are like using a satellite to analyze every subtle movement and terrain change to predict exactly when the traveler will arrive."

# 6 Content-Specific Recommendations

Different content types generate different data signals. Based on the source material, we recommend specific methods for Video, Text, and Interactive content.

## 6.1 Video Content

**Recommended Method:** **Cognitive Load Proxy Model (CLPM)** combined with **Video Engagement Analytics**.

### 6.1.1 Rationale

The source highlights that fine-grained behaviors are critical for video:

- **Pausing and Rewinding:** These actions are "positively correlated with higher exam performance."
- **Fast-Forwarding:** Associated with lower performance.
- **Utilization Rate:** Early course utilization predicts passing.

CLPM is specifically suited to interpret these behaviors as "germane load" (effortful learning) rather than just passive consumption.

## 6.2 Text/PDF Content

**Recommended Method:** **Half-Life Regression (HLR)**.

### 6.2.1 Rationale

Text consumption is passive and harder to track than video. However, HLR (used by Duolingo) models the **forgetting curve** and memory strength over time.

- Since reading does not generate "rich engagement data" like simulations, estimating the decay of knowledge (forgetting) is more reliable than estimating immediate mastery.
- HLR helps predict when a user needs to review the text to maintain proficiency.

## 6.3 Interactive Content (Simulations/Games)

**Recommended Method:** **Stealth Assessment (SPRING)**.

### 6.3.1 Rationale

Interactive content provides the richest data stream.

- **Evidence:** Pearson's SPRING research showed a correlation of 0.55 by inferring proficiency from game logs.
- **Mechanism:** It uses Evidence-Centered Design to map action sequences (decisions made in the simulation) to proficiency states without asking explicit questions.
- **Validation:** This approach validates that "action sequences can predict learning."

# 7 Implementation Roadmap

## 7.1 Phase 1: Foundation (Heuristic Layer)

**Objective:** Establish data collection and baseline tracking.

- Implement **Time-on-Task** and **Completion Metrics**.
- **Goal:** Collect raw engagement logs (time, clicks, completion %).
- **Note:** Treat this as "unconfirmed" evidence.

## 7.2 Phase 2: Modeling (Probabilistic Layer)

**Objective:** contextualize data with cognitive theory.

- Implement **Half-Life Regression (HLR)** for retention tracking.
- Apply **Cognitive Load Proxy Models** to video player data (tracking pause/rewind events).
- **Goal:** Differentiate between passive watching and active study.

## 7.3 Phase 3: Advanced Prediction (ML Layer)

**Objective:** High-fidelity proficiency estimation.

- Deploy **Deep Knowledge Tracing (DKT)** or **Stealth Assessment** algorithms.
- Train models on historical data to map action sequences to assessment outcomes.
- **Goal:** Achieve correlation levels of $r = 0.40 - 0.65$ with standardized norms.

# 8 Limitations and Future Research

## 8.1 Uncertainty

Proficiency estimation from engagement remains an "estimation problem under uncertainty." Platforms like Khan Academy and Coursera still rely on assessments for "mastery," viewing engagement primarily as a support tool.

## 8.2 Data Quality

Passive study is less predictive than active practice. Estimates derived solely from non-question data should be treated as "lower-confidence evidence" until validated by a later assessment event.

## 8.3 Validation

Success is measured by predictive accuracy. Future implementation must validate "stealth" estimates against external standardized tests (like MAP Growth) to ensure that students identified as engaged are indeed showing growth norms.

# 9 Conclusion

The research confirms that while direct assessment remains the gold standard for mastery verification, **Non-Question Proficiency Evaluation** is a feasible and valuable component of modern learning systems. By moving from simple heuristic tracking to advanced ML-based methods like **Stealth Assessment** and **Deep Knowledge Tracing**, educational platforms can achieve significant predictive power ($r \approx 0.65$).

**Final Recommendation:** Adopt a multi-modal approach. Use **CLPM** for video to capture cognitive load, **HLR** for text to manage retention, and **Stealth Assessment** for interactive components to infer skill application. This composite strategy balances implementation complexity with the need for accurate, actionable learner insights.

# 10 References


1. Corbett, A. T., & Anderson, J. R. (1994). *Knowledge tracing: Modeling the acquisition of procedural knowledge.*
2. Settles, B., & Meeder, B. (2016). *A trainable spaced repetition model.*
3. Gonzalez-Brenes et al. (2016). *A Data-Driven Approach for Inferring Student Proficiency from Game Activity Logs.*
4. Piech, C., et al. (2015). *Deep knowledge tracing.*
5. Guo, P. J., Kim, J., & Rubin, R. (2014). *How video production affects student engagement.*
6. Yürüm et al. (2022). *The use of video clickstream data to predict university students' test performance.*
7. Tong & Ren (2025). *Deep knowledge tracing and cognitive load estimation for personalized learning path.*
8. Chi, M. T. H., & Wylie, R. (2014). *The ICAP framework: Linking cognitive engagement to active learning outcomes.*