# Non-Question Proficiency Evaluation Framework

Methodology Comparison and Ranking Report

*Which Method is Best for Non-Question Proficiency Evaluation?*

**Author:** Research Analysis Team

**Version:** 1.0
**Date:** January 8, 2026
**Status:** Final Report
**Project:** NEXS-399

## Abstract

This report provides a comprehensive evaluation and ranking of methodologies for estimating learner proficiency from non-question learning activities. Based on systematic analysis of thirteen identified methods across three categories—Heuristic-Based, Model-Based, and Machine Learning-Based—this document delivers evidence-based recommendations for implementation in the Non-Question Proficiency Evaluation Framework. The analysis draws exclusively from established research evidence documenting correlations between engagement data and learning outcomes, with statistical accuracy typically ranging from $r = 0.40$ to $0.65$.

# Contents

# 1 Executive Summary

## 1.1 Overview

This report addresses the critical question: **Which method is best for the Non-Question Proficiency Evaluation Framework?** The analysis evaluates thirteen distinct methodologies for estimating proficiency gain from engagement data, organized into three categories: Heuristic-Based Methods (4 methods), Model-Based Methods (5 methods), and Machine Learning-Based Methods (4 methods).

The fundamental premise—that estimating proficiency from non-question activities is feasible—is strongly supported by research evidence. Studies consistently demonstrate correlations between engagement metrics and learning outcomes, with behavioral indicators such as video pausing and rewinding positively correlated with exam performance. Engagement-based proficiency estimation achieves moderate to strong correlation with assessment outcomes, typically ranging from $r = 0.40$ to 0.65.

## 1.2 Key Findings

### 1.2.1 Top-Ranked Methods by Category

**Tier 1 (Highest Recommendation):**

1. **Half-Life Regression (HLR)** — Combines the Ebbinghaus forgetting curve with engagement data; demonstrated 12% improvement in daily user retention at Duolingo
2. **Deep Knowledge Tracing (DKT) with Engagement Features** — Neural network approach processing sequences of interactions; supported by peer-reviewed research (Piech et al., 2015)
3. **Stealth Assessment (SPRING)** — Data-driven pipeline achieving correlation of approximately 0.55 with test outcomes from game logs

**Tier 2 (Strong Recommendation):**

4. **Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)** — Extends classical BKT with engagement signals
5. **Performance Factors Analysis (PFA) with Engagement Covariates** — Logistic regression incorporating engagement factors
6. **Multi-Modal Attention Models (MMAE)** — Combines multiple signals for attention quality inference

### 1.2.2 Best Methods by Content Type

| Content Type | Primary Recommendation | Secondary Recommendation |
|---|---|---|
| Video Content | Cognitive Load Proxy Model (CLPM) | DKT with Engagement Features |
| Text/PDF Content | Half-Life Regression (HLR) | Time-Weighted Completion Model |
| Interactive Content | Stealth Assessment (SPRING) | DKT with Engagement Features |

Table 1: Content-Specific Method Recommendations

## 1.3 Implementation Priorities

Based on the analysis, the recommended implementation sequence is:

1. **Phase 1:** Implement Heuristic methods (Time-on-Task, TWCM) for baseline functionality
2. **Phase 2:** Deploy Model-Based methods (HLR, EW-BKT) for improved accuracy
3. **Phase 3:** Integrate ML-Based methods (DKT, SPRING) for advanced capabilities

# 2 Introduction

## 2.1 Problem Statement

Traditional assessment of learner proficiency relies heavily on direct questioning—quizzes, tests, and examinations that provide explicit "right or wrong" signals. However, modern digital learning environments generate vast amounts of engagement data from non-question activities: watching videos, reading articles, interacting with simulations, and navigating educational content. The challenge addressed by the Non-Question Proficiency Evaluation Framework is to **estimate proficiency gain from these passive and semi-active learning interactions without requiring explicit assessment questions**.

This approach offers several potential benefits: reduced test anxiety for learners, continuous rather than discrete proficiency tracking, more engaging learning experiences, and the ability to identify learning progress without interrupting the learning flow. However, it presents significant methodological challenges, as passive study is generally less predictive of retention than active practice, requiring these estimates to be treated as "lower-confidence evidence" until validated by subsequent assessment.

## 2.2 Research Objectives

This report aims to:

1. Provide a **comprehensive review** of all methodologies identified for estimating proficiency from engagement data
2. Develop a **systematic ranking** based on multiple evaluation criteria
3. Deliver a **comparative analysis** of strengths and weaknesses for each approach
4. Offer **content-specific recommendations** for video, text/PDF, and interactive content
5. Propose an **implementation roadmap** for the NEXS-399 project

## 2.3 Scope and Methodology

This analysis is based exclusively on the source document "Feasibility and Methods.txt," which synthesizes research evidence and industry practices regarding proficiency estimation from engagement data. The evaluation considers:

- Empirical evidence supporting each method's validity
- Reported accuracy and predictive power metrics
- Theoretical foundations from cognitive science and learning theory
- Practical implementation considerations
- Industry adoption patterns from major learning platforms

The source material documents practices from Khan Academy, Duolingo, Coursera, edX, Brilliant, and Codecademy, as well as peer-reviewed research from venues including conferences on knowledge tracing, educational data mining, and learning analytics.

### 2.4 Evidence Base for Feasibility

The source material establishes that estimating proficiency from content engagement is **feasible as an estimation problem under uncertainty**. Key supporting evidence includes:

- **Engagement Correlations:** Every additional minute on Khan Academy was associated with gains on standardized tests
- **Behavioral Indicators:** Pausing and rewinding videos positively correlates with higher exam performance; frequent fast-forwarding correlates with lower performance
- **Stealth Assessment Validation:** Pearson's SPRING research demonstrated correlation of approximately 0.55 between game log predictions and test outcomes
- **Early Prediction:** MOOC research found that first-week attendance and utilization rates highly predict eventual course completion
- **Statistical Accuracy Range:** Engagement-based estimation typically achieves $r = 0.40$ to 0.65 correlation with assessment outcomes

# 3 Comprehensive Methodology Review

This section provides detailed descriptions of all thirteen methods identified in the source material, organized by their methodological category. Each method is characterized according to its theoretical basis, operational mechanism, and documented applications.

## 3.1 Category 1: Heuristic-Based Methods

Heuristic-based methods rely on pre-defined rules and expert judgment rather than statistical inference. They are typically employed for immediate feedback and gamification purposes.

### 3.1.1 Method 1: Heuristic Point Systems (XP Models)

**Description:** Assigns experience points or progress percentages for completing content. This approach is exemplified by Duolingo's XP system and Khan Academy's energy points.

**Mechanism:** Points are awarded based on activity completion, with potential weighting by activity type or difficulty. The accumulated points serve as a proxy for progress and engagement.

**Characteristics:**

- Simple to implement and understand
- Provides immediate feedback to learners
- Primarily serves gamification rather than precise proficiency estimation
- Does not directly measure learning or retention

**Industry Example:** Duolingo uses XP for gamification purposes, though their core proficiency model (HLR) relies on different mechanisms.

### 3.1.2 Method 2: Time-on-Task & Completion Metrics

**Description:** Uses normalized time spent and completion rates as direct predictors of success. Metrics include percentage of video watched, time spent on reading materials, and completion status of activities.

**Mechanism:** Engagement is quantified through temporal measures (duration, frequency) and completion indicators, which are then correlated with expected learning outcomes.

**Characteristics:**

- Based on well-documented correlation between time-on-task and learning
- Easy to collect and process at scale
- Susceptible to gaming (leaving content open without engagement)
- Does not distinguish quality of engagement

**Research Support:** Studies show higher engagement—measured by time spent, completion rates, and re-engagement—correlates with better test scores.

### 3.1.3 Method 3: Mastery-Based Engagement Scoring (MBES)

**Description:** A threshold-based system where proficiency levels are assigned based on engagement triggers. Levels typically include: Attempted, Familiar, Proficient, and Mastered.

**Mechanism:** Predefined thresholds determine level transitions. For example, completing 80% of content might advance a learner from "Attempted" to "Familiar."

**Characteristics:**

- Provides discrete, interpretable proficiency levels
- Aligns with mastery learning pedagogical approaches
- Threshold selection requires expert judgment or calibration
- May not capture continuous proficiency development

### 3.1.4 Method 4: Time-Weighted Completion Model (TWCM)

**Description:** A simple model that weights content completion by the quality of time spent relative to expected duration.

**Mechanism:** Compares actual time spent against expected/normative time to assess engagement quality. Completion weighted by this ratio provides an adjusted proficiency estimate.

**Characteristics:**

- Accounts for both completion and engagement quality
- Requires calibration of expected durations
- More nuanced than simple completion metrics
- Still relies on time as primary signal

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

Model-based methods employ established psychological or cognitive theories to model how knowledge is acquired or forgotten over time. They provide probabilistic estimates grounded in learning science.

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

**Description:** An extension of standard Bayesian Knowledge Tracing that treats content interactions as "learning opportunities." Uses engagement signals to modify the probability of knowledge state transitions.

**Mechanism:** Standard BKT models the probability that a student has learned a skill based on response sequences. EW-BKT extends this by incorporating engagement signals (completion, interaction density) to adjust the transition probability from unlearned to learned states.

**Theoretical Foundation:** Based on Corbett & Anderson's (1994) foundational work on knowledge tracing for modeling procedural knowledge acquisition.

**Characteristics:**

- Grounded in well-established knowledge tracing theory
- Provides probabilistic uncertainty estimates

- Adapts classical assessment-based model to engagement data
- Requires parameter estimation from data

**Reference:** Corbett, A. T., & Anderson, J. R. (1994). *Knowledge tracing: Modeling the acquisition of procedural knowledge.*

#### 3.2.2 Method 6: Half-Life Regression (HLR)

**Description:** Combines the Ebbinghaus forgetting curve with engagement data to estimate memory strength and predict recall probability over time.

**Mechanism:** Models the "half-life" of memory—the time until recall probability drops to 50%—as a function of engagement history, practice frequency, and item characteristics. Longer half-lives indicate stronger memory/proficiency.

**Theoretical Foundation:** Based on Settles & Meeder's (2016) trainable spaced repetition model, which integrates forgetting curve theory with machine learning.

**Characteristics:**

- Explicitly models memory decay and retention
- Supports spaced repetition scheduling
- Trainable parameters adapt to individual learners
- Demonstrated success at Duolingo (12% retention improvement)

**Reference:** Settles, B., & Meeder, B. (2016). *A trainable spaced repetition model.*

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

**Description:** A logistic regression model incorporating both prior performance history and current engagement factors to predict proficiency.

**Mechanism:** Extends PFA by including engagement metrics (time spent, interaction patterns) as additional covariates alongside traditional success/failure counts.

**Characteristics:**

- Interpretable logistic regression framework
- Combines assessment and engagement data
- Allows identification of important predictive factors
- Requires sufficient data for reliable parameter estimation

#### 3.2.4 Method 8: Item Response Theory (IRT) Analogy

**Description:** Adapts IRT concepts to treat content pieces as "items" with specific difficulty levels, where future success on related questions validates proficiency gained from that content.

**Mechanism:** Content is characterized by difficulty parameters analogous to IRT item parameters. Engagement with content of known difficulty informs latent proficiency estimates.

**Characteristics:**

- Leverages well-established psychometric theory

- Requires subsequent assessment for validation
- Provides principled difficulty calibration
- Indirect measure requiring assessment confirmation

### 3.2.5 Method 9: Cognitive Load Proxy Model (CLPM)

**Description:** Estimates "germane load" (productive learning effort) versus "extraneous load" by analyzing engagement patterns such as video pauses and rewinds.

**Mechanism:** Interprets behavioral signals (pausing, rewinding, re-reading) as indicators of cognitive processing. Patterns suggesting effortful processing indicate learning engagement.

**Theoretical Foundation:** Based on cognitive load theory and supported by research showing pausing/rewinding correlates with higher exam performance.

**Characteristics:**

- Grounded in cognitive load theory from learning science
- Particularly applicable to video content
- Distinguishes productive struggle from confusion
- Requires behavioral pattern interpretation

**Supporting Evidence:** Guo, Kim, & Rubin (2014) documented how video production affects student engagement; Yürüm et al. (2022) demonstrated video clickstream data predicts test performance.

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

Machine learning approaches learn complex, non-linear mappings between behavioral logs and proficiency outcomes from data, without requiring explicit cognitive models.

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

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

**Mechanism:** Neural networks learn temporal patterns in interaction sequences, modeling how engagement over time relates to knowledge state evolution.

**Theoretical Foundation:** Based on Piech et al.'s (2015) deep knowledge tracing framework, extended with engagement features.

**Characteristics:**

- Captures complex temporal patterns
- Does not require explicit feature engineering
- Requires substantial training data
- Less interpretable than model-based approaches

**Recent Development:** Tong & Ren (2025) integrated DKT with cognitive load estimation for personalized learning paths.

**Reference:** Piech, C., et al. (2015). *Deep knowledge tracing.*

### 3.3.2 Method 11: Stealth Assessment (e.g., Pearson's SPRING)

**Description:** A data-driven pipeline using Evidence-Centered Design (ECD) to infer proficiency from action sequences and game logs without direct questioning.

**Mechanism:** SPRING (Student PRoficiency INferrer from Game data) processes game activity logs through a pipeline that extracts evidence of learning from player actions, validated against external assessments.

**Documented Performance:** Achieved correlation of approximately 0.55 with test outcomes, demonstrating that action sequences can predict learning without quiz questions.

**Characteristics:**

- Designed specifically for game-based and interactive learning
- Evidence-Centered Design provides principled framework
- Validated against external assessments
- Requires structured activity logging

**Reference:** Gonzalez-Brenes et al. (2016). *A Data-Driven Approach for Inferring Student Proficiency from Game Activity Logs.*

### 3.3.3 Method 12: Multi-Modal Attention Models (MMAE)

**Description:** Combines multiple disparate signals—scroll depth, video playback speed changes, session frequency—to infer attention quality and subsequent learning.

**Mechanism:** Attention mechanisms weight different behavioral signals based on their relevance to proficiency prediction, learning optimal combinations from data.

**Characteristics:**

- Integrates diverse behavioral signals
- Attention mechanism identifies relevant features
- Flexible across content types
- Requires multi-modal data collection

### 3.3.4 Method 13: Regression/Classification Predictive Models

**Description:** Direct models (Random Forests, Logistic Regression) trained to predict final exam scores or mastery states early in a course based on "clickstream" features.

**Mechanism:** Standard supervised learning approaches trained on engagement features to predict assessment outcomes.

**Characteristics:**

- Straightforward supervised learning framework
- Can use any combination of engagement features
- Interpretable (especially tree-based methods)
- Requires labeled outcome data for training

**Supporting Evidence:** MOOC research demonstrated that attendance and utilization rates in the first week highly predict eventual course completion.

# 4 Systematic Ranking and Evaluation

This section presents a systematic evaluation of all thirteen methods against defined criteria, resulting in an overall ranking with justification based on evidence from the source material.

## 4.1 Evaluation Criteria

Methods are evaluated on seven criteria derived from the source material's discussion of successful implementations and validation approaches:

1. **Empirical Validity (EV):** Evidence supporting the method's theoretical basis and documented use
2. **Accuracy/Predictive Power (AP):** Reported correlation or prediction accuracy
3. **Theoretical Foundation (TF):** Grounding in established learning science or cognitive theory
4. **Practical Applicability (PA):** Ease of implementation and deployment
5. **Generalizability (GE):** Applicability across different content types and contexts
6. **Validation Capability (VC):** Ability to calibrate and validate estimates
7. **Industry Adoption (IA):** Evidence of use by major learning platforms

## 4.2 Scoring Methodology

Each method receives a score of 1–5 for each criterion:

- **5:** Strong evidence/support in source material
- **4:** Good evidence/support with minor limitations
- **3:** Moderate evidence/support
- **2:** Limited evidence/support
- **1:** Minimal or no evidence/support

**Weighting:** Criteria are weighted based on importance to the framework's goals:

- Empirical Validity: 20%
- Accuracy/Predictive Power: 20%
- Theoretical Foundation: 15%
- Practical Applicability: 15%
- Generalizability: 10%
- Validation Capability: 10%
- Industry Adoption: 10%

## 4.3 Detailed Scoring

### 4.3.1 Heuristic-Based Methods

| **Method** | **EV** | **AP** | **TF** | **PA** | **GE** | **VC** | **IA** | **Weighted** |
|---|---|---|---|---|---|---|---|---|
| XP Models | 3 | 2 | 2 | 5 | 4 | 2 | 4 | 2.95 |
| Time-on-Task | 4 | 3 | 3 | 5 | 4 | 3 | 4 | 3.60 |
| MBES | 3 | 2 | 3 | 4 | 3 | 3 | 3 | 2.90 |
| TWCM | 3 | 3 | 3 | 4 | 4 | 3 | 2 | 3.15 |

Table 2: Heuristic-Based Methods Scoring

**Justifications:**

*XP Models:* High practical applicability (simple implementation) and industry adoption (Duolingo, Khan Academy) but low accuracy—source notes these are for gamification, not proficiency measurement.

*Time-on-Task:* Source documents correlation between time spent and test scores (Khan Academy evidence). High practical applicability with moderate accuracy.

*MBES:* Threshold-based approach aligned with mastery learning but limited evidence of predictive accuracy in source.

*TWCM:* Improves on simple completion by incorporating time quality; moderate support across criteria.

### 4.3.2 Model-Based Methods

| **Method** | **EV** | **AP** | **TF** | **PA** | **GE** | **VC** | **IA** | **Weighted** |
|---|---|---|---|---|---|---|---|---|
| EW-BKT | 4 | 4 | 5 | 3 | 4 | 4 | 3 | 3.90 |
| HLR | 5 | 5 | 5 | 4 | 4 | 5 | 5 | 4.70 |
| PFA + Engagement | 4 | 4 | 4 | 3 | 4 | 4 | 3 | 3.75 |
| IRT Analogy | 3 | 3 | 5 | 2 | 3 | 4 | 2 | 3.10 |
| CLPM | 4 | 4 | 5 | 3 | 3 | 3 | 3 | 3.65 |

Table 3: Model-Based Methods Scoring

**Justifications:**

*EW-BKT:* Strong theoretical foundation (Corbett & Anderson, 1994); adapts proven BKT to engagement data with good generalizability.

*HLR:* Highest scores—source documents 12% retention improvement at Duolingo, strong theoretical basis in forgetting curves, trainable parameters, and clear validation approach.

*PFA + Engagement:* Interpretable logistic regression with engagement covariates; solid theoretical grounding and good accuracy.

*IRT Analogy:* Excellent theoretical foundation but requires assessment validation; lower practical applicability for pure engagement-based estimation.

*CLPM:* Strong fit for video content based on documented correlation between pausing/rewinding and performance; specialized rather than general.

### 4.3.3 ML-Based Methods

| Method | EV | AP | TF | PA | GE | VC | IA | Weighted |
|---|---|---|---|---|---|---|---|---|
| DKT + Engagement | 5 | 4 | 4 | 2 | 5 | 4 | 4 | 4.00 |
| SPRING | 5 | 5 | 4 | 2 | 3 | 5 | 4 | 4.10 |
| MMAE | 4 | 4 | 3 | 2 | 4 | 3 | 2 | 3.30 |
| Regression/Classification | 4 | 4 | 2 | 4 | 4 | 4 | 4 | 3.70 |

Table 4: ML-Based Methods Scoring

**Justifications:**

*DKT + Engagement:* Strong empirical support (Piech et al., 2015; Tong & Ren, 2025); highly generalizable but complex implementation.

*SPRING:* Documented correlation of $\approx 0.55$ with test outcomes; excellent validation against external assessments; specialized for interactive/game content.

*MMAE:* Flexible multi-modal approach but limited documented performance metrics in source.

*Regression/Classification:* Strong for early prediction (MOOC evidence); practical but theoretically less grounded.

## 4.4 Overall Ranking

| Rank | Method | Score | Tier |
|---|---|---|---|
| 1 | Half-Life Regression (HLR) | 4.70 | Tier 1 |
| 2 | Stealth Assessment (SPRING) | 4.10 | Tier 1 |
| 3 | DKT with Engagement Features | 4.00 | Tier 1 |
| 4 | Engagement-Weighted BKT | 3.90 | Tier 2 |
| 5 | PFA with Engagement Covariates | 3.75 | Tier 2 |
| 6 | Regression/Classification | 3.70 | Tier 2 |
| 7 | Cognitive Load Proxy Model | 3.65 | Tier 2 |
| 8 | Time-on-Task & Completion | 3.60 | Tier 3 |
| 9 | Multi-Modal Attention Models | 3.30 | Tier 3 |
| 10 | Time-Weighted Completion | 3.15 | Tier 3 |
| 11 | IRT Analogy | 3.10 | Tier 3 |
| 12 | Heuristic Point Systems (XP) | 2.95 | Tier 4 |
| 13 | Mastery-Based Engagement Scoring | 2.90 | Tier 4 |

Table 5: Overall Method Ranking

## 4.5 Tier Classification

**Tier 1 (Score $\geq$ 4.0):** Methods with strong empirical validation, demonstrated accuracy, and solid theoretical or practical foundations. Recommended for primary implementation.

**Tier 2 (Score 3.6–3.99):** Methods with good support and clear applicability. Recommended for secondary implementation or specific use cases.

**Tier 3 (Score 3.1–3.59):** Methods with moderate support. Suitable for baseline or supplementary roles.

**Tier 4 (Score $<$ 3.1):** Methods with limited evidence for accurate proficiency estimation. Suitable only for gamification or auxiliary purposes.

# 5 Comparative Analysis

This section provides detailed side-by-side comparison of methods, analyzing strengths and weaknesses, use cases, and implementation complexity.

## 5.1 Strengths and Weaknesses Matrix

| Method | Strengths | Weaknesses |
|---|---|---|
| XP Models | Simple implementation; immediate learner feedback; strong gamification value | Does not measure actual learning; easily gamed; no predictive validity for proficiency |
| Time-on-Task | Easy to collect; documented correlation with outcomes; scalable | Cannot distinguish engagement quality; susceptible to passive viewing; moderate accuracy |
| MBES | Interpretable levels; aligns with mastery learning; clear thresholds | Threshold selection arbitrary; discrete rather than continuous; limited validation |
| TWCM | Accounts for engagement quality; more nuanced than simple completion | Requires expected duration calibration; still primarily time-based |
| EW-BKT | Strong theoretical foundation; probabilistic uncertainty; adapts proven model | Parameter estimation complexity; requires substantial interaction data |
| HLR | Documented 12% retention improvement; models forgetting; trainable | Requires practice data for optimal calibration; complexity moderate |
| PFA + Engagement | Interpretable coefficients; combines assessment and engagement; flexible | Requires assessment data for training; parameter estimation needs data |
| IRT Analogy | Strong psychometric theory; principled difficulty calibration | Requires subsequent assessment validation; indirect measurement |
| CLPM | Grounded in cognitive load theory; strong for video; captures effort | Video-specific; requires behavioral pattern interpretation |
| DKT + Engagement | Captures complex patterns; highly flexible; strong empirical support | Requires substantial data; less interpretable; computational cost |
| SPRING | Validated $r \approx 0.55$; designed for interactive content; ECD framework | Specialized for games/simulations; requires structured logging |

*Continued on next page*

| Method | Strengths | Weaknesses |
|---|---|---|
| MMAE | Integrates diverse signals; flexible; attention mechanism | Limited performance documentation; complex implementation |
| Regression/Classification | Straightforward ML; interpretable trees; good early prediction | Theoretically less grounded; requires labeled outcomes |

Table 6: Comprehensive Strengths and Weaknesses Analysis

## 5.2 Use Case Analysis

| Method | Real-time | Low Data | High Accuracy | Interpretable |
|---|---|---|---|---|
| XP Models | ✓ | ✓ |  | ✓ |
| Time-on-Task | ✓ | ✓ |  | ✓ |
| MBES | ✓ | ✓ |  | ✓ |
| TWCM | ✓ | ✓ |  | ✓ |
| EW-BKT |  |  | ✓ | ✓ |
| HLR |  |  | ✓ | ✓ |
| PFA + Engagement |  |  | ✓ | ✓ |
| IRT Analogy |  |  |  | ✓ |
| CLPM | ✓ |  | ✓ | ✓ |
| DKT + Engagement |  |  | ✓ |  |
| SPRING |  |  | ✓ |  |
| MMAE |  |  | ✓ |  |
| Regression/Classification |  |  | ✓ | ✓ |

Table 7: Use Case Suitability Matrix

**Legend:**

- **Real-time:** Suitable for immediate feedback without historical data
- **Low Data:** Functions effectively with minimal training data
- **High Accuracy:** Documented strong predictive performance
- **Interpretable:** Provides understandable proficiency estimates

### 5.3 Implementation Complexity Assessment

| Method | Complexity | Implementation Notes |
|---|---|---|
| XP Models | Simple | Basic counters and aggregations |
| Time-on-Task | Simple | Event logging and duration calculation |
| MBES | Simple | Threshold configuration and level assignment |
| TWCM | Simple | Duration normalization and weighting |
| EW-BKT | Moderate | Bayesian parameter estimation; state tracking |
| HLR | Moderate | Forgetting curve modeling; regression training |
| PFA + Engagement | Moderate | Logistic regression with feature engineering |
| IRT Analogy | Moderate | Item parameter estimation; latent trait modeling |
| CLPM | Moderate | Behavioral pattern recognition; cognitive interpretation |
| DKT + Engagement | Complex | Neural network architecture; sequence modeling |
| SPRING | Complex | ECD framework; evidence extraction pipeline |
| MMAE | Complex | Multi-modal fusion; attention mechanisms |
| Regression/Classification | Moderate | Standard ML pipeline; feature selection |

Table 8: Implementation Complexity Classification

### 5.4 Category Comparison Summary

**Heuristic Methods:** Best for rapid deployment, gamification, and baseline functionality. Provide immediate feedback but lower accuracy. Average weighted score: 3.15.

**Model-Based Methods:** Best for theoretically grounded estimation with interpretable results. Balance accuracy and implementation complexity. Average weighted score: 3.82.

**ML-Based Methods:** Best for highest accuracy when sufficient data is available. Capture complex patterns but require more resources. Average weighted score: 3.78.

# 6 Content-Specific Recommendations

This section provides targeted recommendations for each content type based on the characteristics of available engagement signals and the documented evidence for each method's applicability.

## 6.1 Video Content

### 6.1.1 Recommended Methods

**Primary: Cognitive Load Proxy Model (CLPM)**

*Rationale:* The source material explicitly documents that behavioral indicators from video interaction—specifically pausing and rewinding—are positively correlated with higher exam performance, while frequent fast-forwarding is associated with lower performance. CLPM is designed to interpret exactly these signals as indicators of cognitive engagement.

*Evidence:*

- Guo, Kim, & Rubin (2014) documented how video production affects student engagement
- Yürüm et al. (2022) demonstrated that video clickstream data predicts university students' test performance

**Secondary: DKT with Engagement Features**

*Rationale:* For platforms with sufficient video interaction data, DKT can learn complex temporal patterns in viewing behavior that correlate with learning outcomes.

### 6.1.2 Key Engagement Signals for Video

Based on the source material, the following signals are particularly relevant for video content:

- **Pause frequency and duration:** Positive correlation with learning
- **Rewind/replay actions:** Indicator of effortful processing
- **Fast-forward frequency:** Negative correlation with performance
- **Completion percentage:** Basic engagement indicator
- **Total watch time vs. video length:** Engagement quality proxy

### 6.1.3 Implementation Considerations

1. Implement detailed video player event logging (pause, seek, play, complete)
2. Calculate behavioral ratios (rewinds per minute, completion adjusted for fast-forward)
3. Calibrate expected viewing times for different video lengths and complexities
4. Consider segmenting long videos to capture per-segment engagement

## 6.2 Text/PDF Content

### 6.2.1 Recommended Methods

**Primary: Half-Life Regression (HLR)**

*Rationale:* HLR's strength lies in modeling memory and retention over time, which is particularly relevant for text-based learning where material is often studied in discrete sessions and retention must be tracked across time. The documented 12% retention improvement at Duolingo demonstrates effectiveness for content that requires memorization and recall.

**Secondary: Time-Weighted Completion Model (TWCM)**

*Rationale:* For simpler implementations, TWCM provides a practical approach by comparing actual reading time against expected duration, capturing engagement quality for text content.

### 6.2.2 Key Engagement Signals for Text/PDF

- **Time on page/section:** Primary engagement indicator
- **Scroll depth and patterns:** Coverage and re-reading indicators
- **Session frequency:** Spaced practice indicator for HLR
- **Completion status:** Basic progress indicator
- **Highlighting/annotation (if available):** Active engagement marker

### 6.2.3 Implementation Considerations

1. Calculate expected reading times based on word count and complexity
2. Track scroll position and time spent per section
3. For HLR, maintain history of engagement with each content item
4. Consider text difficulty when calibrating expected engagement

## 6.3 Interactive Content (Simulations, Games, Exercises)

### 6.3.1 Recommended Methods

**Primary: Stealth Assessment (SPRING)**

*Rationale:* SPRING was specifically designed for inferring proficiency from game activity logs using Evidence-Centered Design. The source material documents a correlation of approximately 0.55 between SPRING predictions and test outcomes, validating that action sequences can predict learning without quiz questions.

*Evidence:* Gonzalez-Brenes et al. (2016) demonstrated this data-driven approach for inferring student proficiency from game activity logs.

**Secondary: DKT with Engagement Features**

*Rationale:* DKT's sequence modeling capabilities are well-suited to the rich interaction patterns generated by interactive content.

### 6.3.2 Key Engagement Signals for Interactive Content

- **Action sequences:** Pattern of interactions with the simulation/game
- **Time between actions:** Reflection and planning indicators
- **Success/failure on embedded challenges:** Direct performance signals
- **Exploration patterns:** Coverage and curiosity indicators
- **Help-seeking behavior:** Self-regulation indicator
- **Retry patterns:** Persistence and learning from failure

### 6.3.3 Implementation Considerations

1. Design structured action logging aligned with ECD principles
2. Identify evidence rules linking actions to competencies
3. Include embedded micro-challenges that don't feel like assessment
4. Track both process (how) and outcome (what) measures

## 6.4 Content-Specific Summary Table

| Content Type | Primary Method | Key Signal | Complexity |
|---|---|---|---|
| Video | CLPM | Pause/rewind patterns | Moderate |
| Text/PDF | HLR | Time + session spacing | Moderate |
| Interactive | SPRING | Action sequences | Complex |

Table 9: Content-Specific Recommendation Summary

# 7 Implementation Roadmap

This section provides a phased approach to implementing the Non-Question Proficiency Evaluation Framework based on the method rankings and practical considerations.

## 7.1 Phased Implementation Approach

### 7.1.1 Phase 1: Foundation (Baseline Functionality)

**Duration:** Initial deployment phase

**Methods to Implement:**

1. Time-on-Task & Completion Metrics (Rank 8)
2. Time-Weighted Completion Model (Rank 10)
3. XP Models (Rank 12) — for gamification only

**Rationale:** These heuristic methods provide immediate functionality with simple implementation. They establish data collection infrastructure and provide baseline proficiency signals while more sophisticated methods are developed.

**Deliverables:**

- Event logging for all content interactions
- Basic proficiency scores based on completion and time
- Gamification layer (XP, streaks) for engagement
- Data pipeline for storing interaction logs

**Success Criteria:**

- All content interactions logged with timestamps
- Basic proficiency scores generated in real-time
- Data available for training advanced models

### 7.1.2 Phase 2: Enhancement (Model-Based Methods)

**Duration:** After Phase 1 completion + data accumulation

**Methods to Implement:**

1. Half-Life Regression (Rank 1) — for text content
2. Cognitive Load Proxy Model (Rank 7) — for video content
3. Engagement-Weighted BKT (Rank 4) — for sequential content

**Rationale:** Model-based methods provide theoretically grounded proficiency estimation with interpretable results. HLR offers the best overall performance and enables spaced repetition scheduling.

**Prerequisites:**

- Sufficient engagement data from Phase 1
- Some validated assessment outcomes for calibration
- Video player with detailed event tracking (for CLPM)

**Deliverables:**

- HLR model trained on user engagement data
- CLPM scoring for video content
- Spaced repetition recommendations based on HLR half-lives
- Calibrated confidence estimates for proficiency scores

**Success Criteria:**

- Proficiency estimates correlate with subsequent assessment ($r > 0.40$)
- Retention improvement measurable through A/B testing
- Users receive personalized review recommendations

### 7.1.3 Phase 3: Advanced Capabilities (ML-Based Methods)

**Duration:** After Phase 2 validation + substantial data accumulation

**Methods to Implement:**

1. Deep Knowledge Tracing with Engagement (Rank 3)
2. Stealth Assessment/SPRING (Rank 2) — for interactive content
3. Regression/Classification Models (Rank 6) — for early prediction

**Rationale:** ML-based methods capture complex patterns and provide highest accuracy when sufficient training data is available. DKT complements model-based approaches; SPRING enables proficiency inference from game-like interactions.

**Prerequisites:**

- Large-scale engagement data (thousands of users)
- Validated assessment outcomes for supervised training
- Computational infrastructure for neural network training
- Interactive content with structured action logging

**Deliverables:**

- Trained DKT model for sequence-based proficiency tracking
- SPRING pipeline for interactive content assessment
- Early warning system using classification models
- Ensemble proficiency scores combining multiple methods

**Success Criteria:**

- Proficiency estimates achieve $r > 0.50$ with assessments
- Early prediction enables effective intervention
- Interactive content provides valid proficiency signals without questions

## 7.2 Method Prioritization Summary

| Phase | Methods | Content Focus | Data Needs |
|---|---|---|---|
| Phase 1 | Time-on-Task, TWCM, XP | All | Minimal |
| Phase 2 | HLR, CLPM, EW-BKT | Text, Video | Moderate |
| Phase 3 | DKT, SPRING, Regression | Interactive, All | Substantial |

Table 10: Implementation Phase Summary

## 7.3 Dependencies and Prerequisites

**Infrastructure Requirements:**

- Event logging system capturing all content interactions
- Data warehouse for storing engagement histories
- Assessment system for calibration and validation
- ML training infrastructure (Phase 3)

**Data Requirements:**

- Phase 1: Real-time interaction events
- Phase 2: User histories + some assessment outcomes
- Phase 3: Large-scale labeled data + structured action logs

# 8 Limitations and Future Research

## 8.1 Known Limitations

Based on the source material, several fundamental limitations apply to non-question proficiency estimation:

### 8.1.1 Inherent Uncertainty

The source characterizes this approach as "an estimation problem under uncertainty." Passive study is less predictive of retention than active practice, meaning engagement-based estimates should be treated as **"lower-confidence evidence"** until validated by assessment. The analogy provided is apt: "Estimating proficiency from content engagement is like tracking a hiker's progress by observing their pace and the terrain they cover; while you can reasonably estimate how far they've come, you don't know for certain they've reached the summit until they check in at the peak."

### 8.1.2 Accuracy Ceiling

The documented statistical accuracy range of $r = 0.40$ to 0.65 indicates that engagement-based methods explain only 16–42% of variance in assessment outcomes. This represents a ceiling on precision that may limit high-stakes applications.

### 8.1.3 Industry Practice

The source notes that "most major learning platforms track engagement data extensively but are conservative about using it as the sole proof of proficiency." Khan Academy, for example, does not count video watching toward mastery—they prioritize "learning by doing" to ensure high confidence. This conservatism reflects the limitations of engagement-only estimation.

### 8.1.4 Gaming and Validity Threats

Simple metrics like time-on-task are susceptible to gaming (leaving content open without engagement). More sophisticated methods are needed to distinguish genuine engagement from superficial interaction.

## 8.2 Areas Requiring Further Validation

### 8.2.1 Cross-Platform Generalization

While methods have been validated in specific contexts (Duolingo for HLR, game-based learning for SPRING), further research is needed on generalization across different platforms, content types, and learner populations.

### 8.2.2 Long-Term Predictive Validity

Most documented evidence focuses on immediate or short-term correlations. Long-term retention and transfer of learning require additional validation.

### 8.2.3 Calibration Methods

The source documents the need for validation against external assessments but provides limited guidance on optimal calibration procedures and frequency.

## 8.3 Future Research Directions

Based on the source material's discussion of emerging work:

1. **Cognitive Load Integration:** Tong & Ren (2025) demonstrate the value of integrating knowledge tracing with cognitive load estimation for personalized learning paths.
2. **Multi-Modal Fusion:** Combining diverse behavioral signals (MMAE approach) warrants further investigation.
3. **Confidence Calibration:** Methods for producing well-calibrated uncertainty estimates alongside proficiency scores.
4. **Transfer Learning:** Reducing data requirements for new content/contexts through transfer from existing models.

# 9 Conclusion and Recommendations

## 9.1 Summary of Key Findings

This report has systematically evaluated thirteen methods for estimating proficiency from non-question learning activities. The analysis, based exclusively on documented research evidence and industry practice, yields the following conclusions:

**Feasibility:** Estimating proficiency from engagement data is feasible as an estimation problem, with documented correlations of $r = 0.40$ to 0.65 between engagement metrics and assessment outcomes.

**Top Methods:** The three highest-ranked methods are:

1. **Half-Life Regression (HLR)** — Score: 4.70 — Best overall due to documented effectiveness (12% retention improvement), strong theoretical foundation, and practical applicability
2. **Stealth Assessment (SPRING)** — Score: 4.10 — Best for interactive content with validated correlation of ≈ 0.55
3. **Deep Knowledge Tracing with Engagement** — Score: 4.00 — Best for capturing complex temporal patterns with sufficient data

**Content-Specific Recommendations:**

- **Video:** Cognitive Load Proxy Model leveraging pause/rewind behavioral signals
- **Text/PDF:** Half-Life Regression for retention modeling
- **Interactive:** Stealth Assessment (SPRING) for action sequence inference

## 9.2 Final Recommendations

For the NEXS-399 Non-Question Proficiency Evaluation Framework:

1. **Adopt a multi-method approach** that matches methods to content types rather than seeking a single universal solution.
2. **Implement in phases**, starting with simple heuristics to establish data collection, then progressing to model-based and ML-based approaches as data accumulates.
3. **Prioritize HLR** for text content and general proficiency tracking due to its documented effectiveness and trainability.
4. **Invest in CLPM** for video content, implementing detailed player event logging to capture behavioral indicators.
5. **Design interactive content** with SPRING-style evidence-centered logging from the outset.
6. **Treat engagement-based estimates as lower-confidence evidence** and use periodic assessment to calibrate and validate models.
7. **Communicate uncertainty** to stakeholders—these methods provide valuable signals but not definitive proficiency determination.

### 9.3 Next Steps

1. Implement Phase 1 event logging and baseline heuristic methods
2. Design calibration study with embedded assessments
3. Develop HLR model for text content
4. Implement video player instrumentation for CLPM
5. Plan interactive content with structured action logging
6. Establish A/B testing framework for method validation

## 10 References

The following references are drawn from the source material's documented peer-reviewed citations:

1. Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243.

2. Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.

3. Gonzalez-Brenes, J. P., Huang, Y., & Brusilovsky, P. (2016). A data-driven approach for inferring student proficiency from game activity logs. *Proceedings of the Workshop on Games and Learning Alliance*.

4. Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. *Proceedings of the First ACM Conference on Learning @ Scale*, 41–50.

5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28.

6. Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1848–1858.

7. Tong, X., & Ren, Y. (2025). Deep knowledge tracing and cognitive load estimation for personalized learning path. *Journal of Educational Technology*.

8. Yürüm, O. T., et al. (2022). The use of video clickstream data to predict university students' test performance. *Computers & Education*, 177, 104366.

*End of Report*

Document Version: 1.0
Date: January 8, 2026
Status: Final Report