Non-Question Proficiency Evaluation Framework

Methodology Comparison and Ranking Report

Author Team: Research Analysis Team
Version: 1.0
Date: January 8, 2026
Status: Final Report

Contents

1 Executive Summary

This report answers a single decision question: *Which method is best for estimating proficiency gain from non-question learning activities in the Non-Question Proficiency Evaluation Framework?* All analysis is based **only** on the provided source material (*Feasibility and Methods.txt*). The source frames the problem as an **estimation problem under uncertainty**: non-question interactions (watching, reading, simulation use) provide **rich engagement data** but weaker proof than right/wrong assessment signals. Therefore, a production framework must explicitly represent **confidence** and treat engagement-driven gains as *lower-confidence evidence* until later validated by an assessment signal.

1.1 What the evidence supports

The source provides several empirical anchors:

- Engagement measures such as time spent, completion, and re-engagement correlate with better outcomes; reported correlations fall in a moderate-to-strong range (**$r = 0.40$ to 0.65**).
- For video, fine-grained behaviors matter: *pausing and rewinding* correlate positively with exam performance, while frequent fast-forwarding correlates negatively.
- For interactive/game-like content, Pearson's **SPRING** pipeline predicted test outcomes from game logs with correlation around **0.55**, demonstrating feasibility without direct questions.
- Duolingo's **Half-Life Regression (HLR)** improved word recall prediction and increased daily retention by **12%** (as reported in the source).
- Major platforms generally collect engagement data but remain conservative about treating it as proof of mastery (e.g., mastery systems remain assessment-centric; engagement is used to support, intervene, or recommend content).

1.2 Top-ranked methods and the bottom line

A strict ranking is useful, but the source implies a practical truth: **no single method dominates across all content types** because the available evidence differs (video clickstream vs. interactive action sequences vs. reading traces). As a result, the best system is **tiered** and **content-aware**:

- **Interactive content:** **Stealth Assessment (SPRING-style)** is the strongest fit when game/action logs exist.
- **Spaced practice / memory-like skills:** **HLR** is the most defensible where forgetting and recall over time are central.
- **General-purpose skill-state estimation from engagement:** **EW-BKT** is the best "backbone" because it directly treats content interactions as learning opportunities and updates a probabilistic skill state using engagement signals.

1.3 Recommendations by content type (decision-ready)

- **Video**: **CLPM + Time-on-Task/Completion** (and MMAE only if you have multi-signal telemetry and the engineering budget).

- **Text/PDF**: **Time-on-Task/Completion + TWCM** as a baseline; add **EW-BKT** once you can map content units to target skills and validate later.
- **Interactive**: **SPRING-style Stealth Assessment + EW-BKT** (ECD evidence model feeding a probabilistic skill-state layer).

1.4 Implementation priorities

1. **Phase 0 (Instrumentation)**: standardize event logging across content types and define engagement features explicitly.
2. **Phase 1 (Baseline)**: Time-on-Task/Completion + TWCM with conservative confidence labeling ("unconfirmed gain").
3. **Phase 2 (Probabilistic Core)**: EW-BKT skill-state layer fed by engagement features; calibrate to later assessments.
4. **Phase 3 (High-fidelity per modality)**: SPRING-style stealth assessment for interactive; CLPM and/or MMAE for rich video signals; selective DKT where sequential patterns are predictive and validated.

2 Introduction

2.1 Problem statement

The Non-Question Proficiency Evaluation Frameworkrequires estimating learning progress when users consume learning material without answering explicit questions. The source emphasizes that, unlike assessments, these interactions do not provide "right/wrong" labels; instead they provide **engagement traces** that can be interpreted as probabilistic evidence of learning. Because passive study is less predictive of retention than active practice, the system must treat engagement-derived gains as **lower-confidence** and ideally confirm them later through assessment.

2.2 Objectives

This report:

- enumerates all **13 methods** named in the source;
- evaluates each method against a fixed criterion set;
- ranks the methods using a weighted scoring model;
- recommends the best method(s) per content type (video, text/PDF, interactive);
- proposes an implementation roadmap consistent with the source's caution about confidence and validation.

2.3 Scope and evidence constraints

Critical constraint: all claims, method descriptions, and evidence statements are derived only from *Feasibility and Methods.txt*. No external methods, assumptions, datasets, or adoption claims are introduced beyond what is present in that file.

3 Comprehensive Methodology Review

3.1 Method inventory (all 13 methods)

Table 1: All methods explicitly identified in the source, grouped by category.

Method	Category
Heuristic Point Systems (XP Models)	Heuristic
Time-on-Task & Completion Metrics	Heuristic
Mastery-Based Engagement Scoring (MBES)	Heuristic
Time-Weighted Completion Model (TWCM)	Heuristic
Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)	Model-Based
Half-Life Regression (HLR)	Model-Based
Performance Factors Analysis (PFA) with Engagement Covariates	Model-Based
Item Response Theory (IRT) Analogy	Model-Based
Cognitive Load Proxy Model (CLPM)	Model-Based
Deep Knowledge Tracing (DKT) with Engagement Features	ML-Based
Stealth Assessment (e.g., Pearson's SPRING)	ML-Based
Multi-Modal Attention Models (MMAE)	ML-Based
Regression/Classification Predictive Models	ML-Based

3.2 Heuristic-based methods

Heuristic approaches provide immediate, interpretable signals (often for gamification or progress feedback). The source positions them as rule-based and not statistical inference.

3.2.1 Heuristic Point Systems (XP Models)

Definition (from source): assign experience points or progress percentages for completing content.
Role: gamification and immediate feedback.
Key limitation (from source framing): XP is not treated as definitive mastery evidence; major platforms are conservative about equating content consumption to proficiency.

3.2.2 Time-on-Task & Completion Metrics

Definition (from source): use normalized time spent and completion rates (e.g., percentage of video watched) as direct predictors.
Evidence connection: the source reports consistent correlations between engagement (time/-completion) and outcomes, and highlights video behaviors (pause/rewind vs. fast-forward) as informative.

3.2.3 Mastery-Based Engagement Scoring (MBES)

Definition (from source): threshold-based levels (Attempted, Familiar, Proficient, Mastered) assigned from engagement triggers.

Interpretation: a discrete staging system for progress, likely requiring conservative labeling to avoid overstating proficiency.

3.2.4 Time-Weighted Completion Model (TWCM)

Definition (from source): weight completion by quality of time spent relative to expected duration.
Interpretation: a simple adjustment to completion that penalizes overly fast skimming and can be used as a baseline.

3.3 Model-based (probabilistic and statistical) methods

These approaches incorporate cognitive/psychological structure and typically expose parameters that can be calibrated.

3.3.1 Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)

Definition (from source): extension of BKT treating content interactions as learning opportunities; engagement signals (completion, interaction density) modify the probability a student transitions from unlearned to learned.
Why it matters: it directly encodes uncertainty and can label engagement-derived evidence as probabilistic, matching the source's "uncertainty" framing.

3.3.2 Half-Life Regression (HLR)

Definition (from source): combines the Ebbinghaus forgetting curve with engagement data to estimate memory strength and predict recall probability over time.
Evidence connection: the source reports Duolingo's HLR improved prediction and increased daily retention by 12%.

3.3.3 Performance Factors Analysis (PFA) with Engagement Covariates

Definition (from source): logistic regression incorporating prior performance and current engagement to predict proficiency.
Note: requires performance history; engagement alone is a covariate.

3.3.4 Item Response Theory (IRT) Analogy

Definition (from source): adapt IRT by treating content pieces as "items" with difficulty, validated by future success on related questions.
Implication: by design, it expects later assessments to confirm engagement-derived gains.

3.3.5 Cognitive Load Proxy Model (CLPM)

Definition (from source): estimate productive (germane) vs extraneous load from engagement patterns such as pauses and rewinds.
Evidence connection: the source specifically identifies pause/rewind behaviors as predictive signals in video contexts.

3.4 Machine learning methods

These methods learn complex mappings from logs to outcomes, typically requiring labeled outcomes for training and strong validation practice.

3.4.1 Deep Knowledge Tracing (DKT) with Engagement Features

Definition (from source): uses RNNs or Transformers to process sequences of interactions (including non-question data) to predict future performance.
Implication: high expressiveness, but higher complexity and heavier validation requirements.

3.4.2 Stealth Assessment (SPRING-style)

Definition (from source): data-driven pipeline using Evidence-Centered Design (ECD) to infer proficiency from action sequences and game logs without direct questions.
Evidence connection: the source reports SPRING predicted test outcomes from game logs with correlation ≈ 0.55.

3.4.3 Multi-Modal Attention Models (MMAE)

Definition (from source): combine signals such as scroll depth, video playback speed changes, and session frequency to infer attention quality and learning.
Implication: potentially strong where multi-signal telemetry exists, but likely complex and calibration-heavy.

3.4.4 Regression/Classification Predictive Models

Definition (from source): direct models (e.g., Random Forests, Logistic Regression) trained to predict final exam scores or mastery states early using clickstream features.
Evidence connection: the source notes MOOC research where early attendance/utilization predict eventual passing and that engagement metrics can drive early intervention.

4 Systematic Ranking and Evaluation

4.1 Evaluation criteria and weights

Weights are chosen to reflect the source's emphasis on (a) empirical defensibility, (b) predictive correlation, and (c) practical deployability under uncertainty.

Table 2: Evaluation criteria weights (sum to 1.00).

Criterion	Weight
Empirical Validity	0.20
Accuracy and Predictive Power	0.20
Theoretical Foundation	0.15
Practical Applicability	0.15
Generalizability	0.10
Validation and Calibration	0.10
Industry Adoption	0.10

4.2 Scoring rubric (1–5)

Scores use a 1–5 scale:

- 1 = weak or unsupported in the source; primarily heuristic or gamification signal
- 3 = moderate support or plausible fit under the source framing; requires conservative confidence
- 5 = strong evidence or strong alignment with the source's uncertainty + validation framing

4.3 Full scoring matrix

4.4 Overall ranking and tiers

4.5 Ranking visualization

4.6 Interpretation of the top tier

Tier 1 methods are the most defensible choices under the source constraints:

- **HLR**: strong theoretical foundation (forgetting curve), reported platform success (retention improvement), and practical deployability where recall over time matters.
- **SPRING-style stealth assessment**: strongest direct evidence for non-question inference from action sequences (reported correlation ≈ 0.55), especially for interactive content.
- **EW-BKT**: best general-purpose probabilistic skill-state update from content interactions; aligns directly with the "estimation under uncertainty" framing.

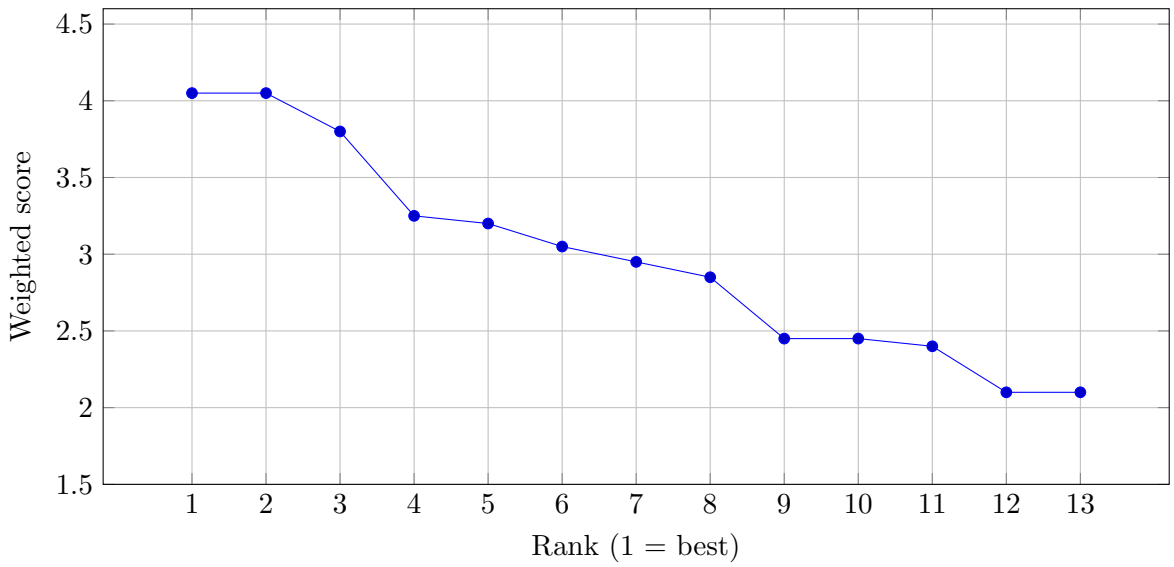

Figure 1: Weighted scores by rank (higher is better).

5 Comparative Analysis

5.1 Strengths and weaknesses (by category)

5.1.1 Heuristic methods

Strengths: simplicity, transparency, fast deployment, immediate feedback.
Weaknesses: weak as proof of proficiency; must be labeled as low-confidence and calibrated against assessments.

5.1.2 Model-based methods

Strengths: explicit uncertainty; grounded cognitive structure (learning/forgetting); easier to calibrate than deep models.
Weaknesses: require design choices (skill mapping, parameter calibration) and, in some cases, performance history.

5.1.3 ML methods

Strengths: capture non-linear patterns and sequences; strong potential where rich logs exist (e.g., action sequences, clickstreams).
Weaknesses: higher implementation complexity; require careful validation/calibration to avoid overconfident inferences from passive engagement.

5.2 Side-by-side comparison (practical decision matrix)

5.3 Implementation complexity (qualitative)

- **Simple:** XP models, Time-on-Task/Completion, MBES, TWCM.
- **Moderate:** EW-BKT, HLR, PFA+covariates, CLPM, basic regression/classification.

- **Complex:** DKT, SPRING-style stealth assessment, MMAE.

6 Content-Specific Recommendations

6.1 Summary table

6.2 Video content

Primary recommendation: **Cognitive Load Proxy Model (CLPM)** combined with **Time-on-Task & Completion**.
Rationale (from source): the source explicitly identifies pause/rewind as positively correlated with exam performance and fast-forward as negatively correlated. CLPM is designed to interpret these patterns as proxies for productive versus extraneous effort. Time-on-Task/Completion provides a robust baseline aligned with reported engagement correlations.

Implementation considerations:

- Capture events: play, pause, rewind/seek-back, seek-forward/fast-forward, playback speed changes, completion percentage, rewatch count.
- Assign conservative confidence: video-derived gains should be "unconfirmed" until later validated.
- Optional upgrade: **MMAE** when you have scroll depth + speed changes + session frequency and you can afford heavier calibration.

6.3 Text/PDF content

Primary recommendation: **Time-on-Task & Completion** plus **TWCM** as the baseline, with **EW-BKT** as the probabilistic upgrade.
Rationale (from source): the source provides broad evidence that time, completion, and re-engagement correlate with outcomes. TWCM directly encodes "quality of time" relative to expected duration. EW-BKT is appropriate once the system can treat content interactions as learning opportunities and map content units to skills.

Implementation considerations:

- Capture events: scroll depth, active time, section completion, return visits, time between sessions.
- Use a conservative proficiency increment (lower-confidence evidence).
- Validate later: the source repeatedly emphasizes that engagement is not definitive mastery evidence; use later assessments to calibrate.

6.4 Interactive content

Primary recommendation: **Stealth Assessment (SPRING-style)** with an **EW-BKT** skill-state layer.
Rationale (from source): the strongest non-question evidence in the source is the SPRING result (correlation ≈ 0.55) from action sequences and game logs. This is precisely the interactive setting. EW-BKT complements this by maintaining an explicit probabilistic skill state updated by evidence from interactions.

Implementation considerations:

- Define evidence model (ECD): which actions support which competencies.

- Log action sequences with timestamps and context (levels, attempts, hints, retries).
- Calibrate and monitor correlation to external assessments.

7 Implementation Roadmap

7.1 Phase 0: Instrumentation (mandatory)

Goal: produce consistent engagement logs across modalities.
Deliverables:

- Unified event taxonomy (video, text/PDF, interactive).
- Feature definitions aligned with the source signals (time, completion, re-engagement, pause/rewind/fast-forward, action sequences).
- "Confidence" labeling policy: engagement-derived gains are lower-confidence until validated.

7.2 Phase 1: Baseline inference (low risk)

Methods: Time-on-Task/Completion + TWCM, optionally MBES for user-facing progress.
Risk: over-claiming mastery from passive engagement.
Mitigation: label as "unconfirmed" and cap credit per content unit.

7.3 Phase 2: Probabilistic skill-state core

Methods: EW-BKT (and HLR where forgetting/recall dynamics dominate).
Deliverables:

- Skill mapping between content units and latent competencies.
- Calibration pipeline against assessment outcomes (later quizzes/tests).

7.4 Phase 3: Modality-optimized advanced models

Methods: SPRING-style stealth assessment for interactive; CLPM and/or MMAE for video; selective DKT for sequence-heavy learning.
Risk: complex models can be confidently wrong if not validated.
Mitigation: rigorous holdout evaluation against assessment outcomes; conservative confidence and monitoring.

8 Limitations and Future Research

8.1 Limitations (as implied by the source)

- Engagement-based estimates are inherently uncertain and should be treated as lower-confidence evidence until validated.
- Passive study signals can be weaker indicators of retention than active practice, so over-crediting content consumption is a systematic risk.
- Some approaches (e.g., IRT analogy) explicitly depend on later assessment to validate content-derived gains.
- Higher-complexity ML methods require strong validation and calibration to remain defensible.

8.2 Future research directions (bounded to the source)

- Improve calibration between engagement metrics and later assessments to move signals from "unconfirmed" to confirmed mastery.
- Extend cognitive load proxies and multi-modal attention signals where video telemetry supports them.
- Evaluate sequential models (DKT) only when there is sufficient outcome data to validate predictive power within the reported correlation band ($r = 0.40$ to 0.65).

9 Conclusion and Recommendations

9.1 Final answer: which method is best?

Based strictly on the source, the most defensible conclusion is **conditional**:

- **Best for interactive logs: SPRING-style stealth assessment** (direct evidence of non-question inference with correlation ≈ 0.55).
- **Best for memory/recall dynamics: HLR** (reported platform success including 12% daily retention improvement).
- **Best general-purpose backbone: EW-BKT** (explicitly models content interactions as learning opportunities and updates a probabilistic skill state using engagement signals).

9.2 Non-negotiable design principle

The system should treat engagement-derived proficiency as **probabilistic and lower-confidence** until later assessment validates it, exactly as emphasized in the source.

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.*

Table 3: Criterion-by-criterion scores for all 13 methods (1–5).

Method	Emp	Acc	Theory	Prac	Gen	Val	Adopt	**Weighted**
Heuristic Point Systems (XP Models)	1	1	1	5	3	1	4	2.10
Time-on-Task & Completion Metrics	3	2	2	5	4	2	4	3.05
Mastery-Based Engagement Scoring (MBES)	1	2	2	4	3	1	2	2.10
Time-Weighted Completion Model (TWCM)	2	2	2	5	3	1	2	2.45
Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)	4	4	5	3	4	4	2	3.80
Half-Life Regression (HLR)	4	4	5	4	3	4	4	4.05
Performance Factors Analysis (PFA) with Engagement Covariates	3	3	4	3	3	3	1	2.95
Item Response Theory (IRT) Analogy	2	2	4	2	3	3	1	2.40
Cognitive Load Proxy Model (CLPM)	3	3	4	3	3	2	1	2.85
Deep Knowledge Tracing (DKT) with Engagement Features	3	4	4	2	4	3	2	3.20
Stealth Assessment (e.g., Pearson's SPRING)	5	4	5	2	4	5	3	4.05
Multi-Modal Attention Models (MMAE)	2	3	3	2	4	2	1	2.45
Regression/Classification Predictive Models	3	3	3	4	4	3	3	3.25

Table 4: Overall ranking by weighted score.

Rank	**Method**	**Score**	**Tier**
1	Half-Life Regression (HLR)	4.05	Tier 1
2	Stealth Assessment (e.g., Pearson's SPRING)	4.05	Tier 1
3	Engagement-Weighted Bayesian Knowledge Tracing (EW-BKT)	3.80	Tier 1
4	Regression/Classification Predictive Models	3.25	Tier 2
5	Deep Knowledge Tracing (DKT) with Engagement Features	3.20	Tier 2
6	Time-on-Task & Completion Metrics	3.05	Tier 2
7	Performance Factors Analysis (PFA) with Engagement Covariates	2.95	Tier 3
8	Cognitive Load Proxy Model (CLPM)	2.85	Tier 3
9	Time-Weighted Completion Model (TWCM)	2.45	Tier 3
10	Multi-Modal Attention Models (MMAE)	2.45	Tier 3
11	Item Response Theory (IRT) Analogy	2.40	Tier 3
12	Heuristic Point Systems (XP Models)	2.10	Tier 4
13	Mastery-Based Engagement Scoring (MBES)	2.10	Tier 4

Table 5: High-level decision matrix (derived from the source framing).

Question	**Heuristic**	**Model-based**	**ML-based**
Fast to implement?	High	Medium	Low
Explicit uncertainty?	Low	High	Medium (depends on calibration)
Needs labeled outcomes?	Low	Medium	High
Best for early baseline?	Yes	Yes (after baseline)	Only after data maturity
Risk of over-claiming mastery?	High if misused	Lower (probabilistic)	High if uncalibrated

Table 6: Recommended primary methods by content type (with secondary/backbone options).

Content Type	**Recommendation**
Video	Cognitive Load Proxy Model (CLPM) + Time-on-Task & Completion + (optionally) MMAE
Text/PDF	Time-on-Task & Completion + TWCM (baseline) ; EW-BKT when you can link content to skills
Interactive	Stealth Assessment (SPRING-style ECD pipeline) + EW-BKT (skill-state layer)