Contents

Beyond Correctness: Measuring Cognitive Stability and Confidence Calibration in Human Understanding

Muhammad Rayan Shahid
Independent Researcher
2026

Preprint — Independent Research

Abstract

Correct answers are often treated as evidence of understanding. However, correctness alone cannot distinguish between stable mastery and fragile performance driven by guessing, memorization, or miscalibrated confidence. This paper argues that **static accuracy-based test scores are unreliable indicators of true understanding** because they ignore metacognitive alignment and reasoning stability.

We investigate this claim through the **Human Cognition Measurement System (HCMS)**, a cognition-aware assessment framework designed to measure understanding as a dynamic, multi-signal construct. Rather than evaluating responses solely by correctness, HCMS integrates confidence–accuracy alignment, repeated-trial consistency, and stability under controlled perturbation to infer cognitive robustness.

Across controlled experiments, we show that learners with equivalent accuracy profiles frequently diverge in cognitive stability and calibration. In particular, confidence–accuracy misalignment is strongly associated with degradation in reasoning consistency under perturbation—patterns that accuracy-only metrics fail to detect. These findings suggest that correctness can mask unstable or miscalibrated understanding, leading to misleading assessments of mastery.

HCMS is presented not as a replacement for existing assessment models, but as an **instrument for probing cognitive validity** beyond correctness. By operationalizing metacognitive alignment and stability as measurable signals, this work provides empirical evidence that assessment systems must move beyond static scores to faithfully represent how humans understand, not just what they answer.

---

1. Introduction

Correct answers are commonly treated as evidence of understanding. In educational assessment, automated grading systems, and standardized testing, performance is typically summarized as a static accuracy score. While efficient, this paradigm implicitly assumes that correctness reflects stable mastery. In practice, however, correct responses can arise from guessing, short-term memorization, pattern matching, or brittle reasoning strategies that do not generalize beyond the immediate context.

This limitation has long been acknowledged in learning science and cognitive psychology. Research on metacognition emphasizes that understanding is not only a function of what a learner answers, but also how confident they are, how consistently they reason across trials, and how their performance responds to variation or uncertainty. Despite this, most computational assessment systems continue to equate correctness with comprehension, offering limited insight into the robustness of a learner's knowledge.

As a result, two learners with identical test scores may possess fundamentally different cognitive states. One may exhibit stable, well-calibrated understanding, while another may display overconfidence, inconsistency, or fragile reasoning that degrades under minimal perturbation. Accuracy-based evaluation is unable to distinguish between these cases, raising a critical question: **Are static test scores reliable indicators of true understanding?**

This paper investigates this question by treating cognition as a multi-dimensional construct rather than a single outcome variable. We introduce the **Human Cognition Measurement System**

(HCMS), a cognition-aware assessment framework designed to probe understanding beyond correctness. HCMS integrates multiple behavioral signals—accuracy, self-reported confidence, repeated-trial consistency, and stability under controlled perturbation—to infer cognitive robustness and metacognitive alignment.

Importantly, HCMS is not proposed as a replacement for existing assessment models, but as a complementary measurement instrument. Its purpose is to reveal cognitive properties that correctness-based metrics systematically obscure, particularly in cases where surface performance appears strong.

Through controlled experiments, we demonstrate that learners with similar accuracy profiles often diverge substantially in confidence calibration and reasoning stability. These divergences become pronounced under perturbation, where miscalibrated learners exhibit significant degradation in cognitive consistency despite unchanged accuracy. These findings suggest that correctness alone is an incomplete and potentially misleading proxy for understanding.

The contributions of this work are threefold: - We formalize cognition-aware assessment as the measurement of stability and calibration, not only correctness. - We present HCMS as an interpretable framework for operationalizing these latent cognitive dimensions. - We provide empirical evidence that confidence–accuracy misalignment predicts instability that static test scores fail to capture.

By reframing assessment as cognitive measurement rather than answer verification, this work aims to advance the design of systems that measure **how people understand**, not merely **what they answer**.

---

2. Related Work

Assessment systems have traditionally emphasized correctness-based evaluation, an approach rooted in standardized testing and automated grading. While effective for scalable performance measurement, such methods reduce understanding to a single observable outcome and overlook deeper cognitive properties.

Metacognition and Confidence Calibration

Educational psychology research has long established the importance of metacognition, particularly learners' ability to monitor and regulate their own understanding. Studies on confidence calibration demonstrate that systematic overconfidence and underconfidence are common and directly impact learning outcomes. However, these findings are typically examined in controlled psychological studies and are rarely operationalized in computational assessment systems.

Learner Modeling and Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) and learner modeling approaches, including Bayesian Knowledge Tracing and Item Response Theory, estimate latent mastery states based on observed performance. While effective for modeling knowledge acquisition, these methods generally treat confidence, stability, and metacognitive behavior as implicit or secondary signals.

Learning Analytics and Behavioral Features

Recent work in learning analytics has explored behavioral indicators such as response time, attempt frequency, and revision patterns. Although these approaches provide richer signals than correctness alone, many rely on complex or opaque models that limit interpretability and robustness under real-world noise.

Positioning of HCMS

HCMS differs from prior work by explicitly modeling cognition as a multi-signal measurement problem. Rather than optimizing prediction accuracy, HCMS focuses on interpretability, stability, and metacognitive alignment. Confidence, consistency, and robustness are treated as first-class cognitive signals rather than auxiliary features.

By integrating these dimensions into a unified, transparent framework, HCMS complements existing assessment and tutoring systems while addressing cognitive properties that are underrepresented in current computational approaches.

---

3. Methodology

The Human Cognition Measurement System (HCMS) is designed as a structured measurement framework rather than a predictive model. Its methodology emphasizes interpretability, modularity, and empirical validity in estimating cognitive understanding beyond surface-level performance.

HCMS operates as a multi-phase pipeline, where each phase corresponds to a distinct cognitive signal or validation layer. This structure enables both reproducibility and independent analysis of each cognitive component.

Data Representation

Learner interaction data consists of: - Concept-level response outcomes - Self-reported confidence ratings - Temporal identifiers for repeated attempts

These inputs enable analysis of both performance and metacognitive behavior over time.

Cognitive Feature Extraction

HCMS derives multiple cognitive features from raw interaction data, including: - Response accuracy
- Confidence magnitude
- Confidence–accuracy alignment
- Temporal consistency across repeated trials
- Variance in confidence and response behavior

These features are computed per learner and per concept, forming the basis for higher-level cognitive inference.

Cognitive Inference

Cognitive states are inferred using rule-based and statistical aggregation mechanisms rather than opaque predictive models. Mastery estimates reflect both correctness and stability, while calibration metrics quantify the alignment between confidence and actual performance.

Misconceptions are identified through recurring incorrect or unstable response patterns rather than isolated errors.

Validation and Reliability Checks

To ensure measurement reliability, HCMS applies internal validation mechanisms, including: - Consistency checks across repeated trials
- Correlation analysis between confidence and accuracy
- Stability evaluation under controlled perturbation

Low confidence–accuracy correlation or high variance across trials indicates potential miscalibration or unstable understanding.

Robustness Evaluation

HCMS explicitly tests the robustness of cognitive inference under noisy and adversarial input conditions. Perturbations are applied in a controlled manner to assess whether inferred cognitive states remain stable when surface features vary.

This step distinguishes robust understanding from brittle correctness.

Explainability and Output Generation

All cognitive inferences are accompanied by transparent decision traces, enabling inspection of contributing signals. Final outputs consist of structured cognitive profiles summarizing understanding level, calibration status, and stability.

Each phase produces intermediate artifacts, allowing full reproducibility and independent verification of the measurement process.

---

4. Experiments

The experimental design aims to evaluate whether correctness-based performance sufficiently reflects stable understanding, and whether additional cognitive signals—particularly confidence calibration and consistency—provide meaningful explanatory power beyond accuracy alone.

All experiments were conducted using structured learner response data collected across multiple concept-level tasks. Each learner interaction included a response label, a self-reported confidence score, and a temporal identifier, enabling repeated-trial and stability analysis.

Experimental Objectives

The experiments were designed to answer three core questions:

1. Can learners with similar accuracy exhibit different cognitive stability profiles?

2. Does confidence–accuracy misalignment correlate with degradation under perturbation?

3. Can cognition-aware metrics distinguish robust understanding from brittle correctness?

Experimental Setup

Learners completed multiple trials on semantically consistent tasks targeting the same underlying concept. For each trial, learners provided: - A binary or categorical response
- A normalized confidence rating
- A timestamped interaction record

HCMS processed these inputs through its cognitive inference pipeline, producing per-learner metrics for accuracy, confidence calibration, and consistency.

To isolate cognitive stability, no adaptive feedback was provided during experimental trials.

Cognitive Profiling and Grouping

Learners were grouped based on confidence–accuracy alignment, computed as the correlation between confidence ratings and response correctness across trials.

Two primary cognitive profiles emerged: - **Calibrated learners** — high alignment between confidence and correctness
- **Miscalibrated learners** — systematic overconfidence or underconfidence

Importantly, grouping was performed independently of raw accuracy, allowing comparison between learners with comparable performance scores.

Controlled Perturbation Protocol

To test robustness, learners were exposed to controlled perturbations applied to previously encountered tasks. Perturbations were designed to preserve semantic meaning while introducing minimal variation, such as: - Rephrased prompts
- Altered surface structure
- Injected response noise

Perturbations were applied uniformly across learner groups.

Stability Metrics

Cognitive stability was measured using: - Change in consistency scores across trials
- Variance in confidence ratings
- Deviation in inferred cognitive state

Stability degradation was quantified as the difference between baseline and perturbed conditions.

Evaluation Criteria

The primary evaluation focused on divergence between accuracy-based and cognition-based assessment outcomes. Specifically, we examined cases where accuracy remained stable but cognitive stability deteriorated under perturbation.

Reproducibility

All experimental procedures, metrics, and configurations are deterministic and fully reproducible. Experimental artifacts, including configuration files and summary statistics, are generated automatically by the HCMS pipeline and preserved for independent verification.

5. Results

The results demonstrate a clear divergence between correctness-based evaluation and cognition-aware assessment. While raw accuracy remained similar across multiple learners, HCMS revealed substantial differences in cognitive stability, confidence calibration, and robustness under perturbation.

Accuracy vs Cognitive Stability

Learners with comparable accuracy scores exhibited markedly different consistency and calibration profiles. In several cases, learners achieving high correctness displayed low stability across repeated trials, indicating fragile or surface-level understanding.

Conversely, some learners with moderate accuracy but high confidence–accuracy alignment maintained stable cognitive states across trials.

Confidence–Accuracy Alignment

Confidence–accuracy correlation emerged as a strong differentiating signal. Learners with alignment values below 0.3 consistently exhibited greater variance in confidence and consistency across trials.

Miscalibrated learners demonstrated unstable confidence behavior even when correctness was preserved, suggesting a disconnect between perceived and actual understanding.

Stability Under Controlled Perturbation

Under controlled perturbations, accuracy scores remained largely unchanged across learner groups. However, cognition-aware metrics showed significant divergence.

Miscalibrated learners experienced pronounced degradation in consistency scores and increased variance in inferred cognitive state following perturbation. In contrast, calibrated learners demonstrated minimal stability loss under identical conditions.

Cognitive Profile Differentiation

HCMS-generated cognitive profiles integrated accuracy, calibration, and stability into a unified learner representation.

In multiple cases, learners classified as equivalent by accuracy-only evaluation were assigned different cognitive verdicts by HCMS, including distinctions between stable mastery, partial understanding, and unstable reasoning.

Summary of Findings

Overall, the results show that: - Accuracy does not reliably indicate cognitive stability
- Confidence–accuracy misalignment predicts robustness degradation
- Cognition-aware metrics reveal latent differences masked by static scores

6. Discussion

The findings reinforce the central claim of this work: correctness alone is an insufficient indicator of understanding. Learners who appear equivalent under traditional accuracy-based evaluation often differ substantially in cognitive stability and metacognitive alignment.

The divergence observed under perturbation is particularly revealing. While accuracy remains unchanged, cognition-aware signals expose fragile reasoning structures that are vulnerable to minimal variation. This suggests that correctness may reflect task familiarity rather than conceptual robustness.

Confidence–accuracy alignment emerges as a critical signal in distinguishing stable understanding from brittle performance. Miscalibrated learners demonstrate internal inconsistency even when external performance appears strong, highlighting the importance of metacognitive measurement.

HCMS provides a structured and interpretable means of operationalizing these constructs. Rather than replacing existing assessment systems, it functions as a diagnostic layer capable of revealing cognitive properties that static scores systematically obscure.

---

7. Limitations and Scope

This work focuses on controlled experimental settings and concept-level tasks. While sufficient for isolating cognitive signals, real-world educational environments introduce additional variables such as feedback, motivation, and curriculum structure.

Confidence reporting is self-assessed and may be influenced by individual response styles. Future work will explore alternative calibration mechanisms and behavioral proxies.

HCMS is designed as a measurement framework rather than a predictive learner model. Its scope is diagnostic and analytic, and it is not intended to directly optimize learning outcomes without integration into broader systems.

---

8. Conclusion and Future Work

This work introduced the Human Cognition Measurement System (HCMS), a cognition-aware assessment framework designed to move beyond correctness-based evaluation toward more faithful measurement of understanding. By integrating accuracy, confidence calibration, and consistency, HCMS provides a structured and interpretable representation of learner cognition.

The results demonstrate that learners with similar accuracy profiles can exhibit substantially different cognitive stability and metacognitive alignment. These differences remain hidden in traditional assessment systems but are consistently revealed through cognition-aware metrics, particularly under controlled perturbation.

HCMS contributes a practical methodology for operationalizing cognitive constructs that have long been emphasized in learning science but underutilized in computational assessment. The framework is modular, reproducible, and designed to function both as a research instrument and as a foundation for adaptive learning systems.

Future work will extend HCMS through longitudinal studies, cross-domain validation, and synthetic learner simulations. Additional efforts will explore real-world deployment within educational platforms, enabling validation at scale while preserving interpretability and cognitive fidelity.

References

[1] B. S. Bloom et al. Taxonomy of Educational Objectives: The Classification of Educational Goals. Longmans, Green, 1956.

[2] J. Dunlosky and J. Metcalfe. Metacognition. Sage Publications, 2009.

[3] A. Koriat. When are two heads better than one and why? The confidence–accuracy relationship. Journal of Experimental Psychology: General, 2012.

[4] F. M. Lord. Applications of Item Response Theory to Practical Testing Problems. Routledge, 1980.

[5] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 1995.

[6] R. A. Bjork and E. L. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. Psychology and the Real World, 2011.

[7] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2019.