FOR PARTNERS & PROGRAM DIRECTORS

The science behind the ERS, *in plain terms.*

This brief explains how the ERS produces a score: the research principles it is built on, the pipeline that turns a mentorship session into numbers, and the engineering that keeps those numbers consistent and auditable. It is written for people who want to know what is happening under the hood, without needing a data science background.

BUILT BY PACEFUL

The ERS is developed by **Paceful**, a behavioral signal scoring platform built to turn structured human observation into consistent, auditable measurement. Paceful's engine is designed for settings where the score supports a human decision rather than replaces it. Mentora runs this engine configured specifically for youth mentorship: the signals it looks for, the rubrics it scores against, and the reports it produces are all built around how young players actually develop across a twelve-week cycle.

THE RESEARCH PRINCIPLES IT STANDS ON

01

Structured observation beats unstructured judgment.

Decades of assessment research point the same way: when observers rate against fixed, behaviorally defined criteria, their judgments are more reliable, more comparable, and less biased than free-form impressions. The ERS applies that finding to mentorship. Every mentor observes through the same structure, every week.

02

Behavioral anchors keep ratings honest.

Adjectives drift. "Motivated" means different things to different people on different days. The ERS rubrics are anchored to described behaviors instead: did the player complete agreed actions, ask questions unprompted, respond to correction by adjusting. Behaviors can be observed and checked. Impressions can't.

03

Trends beat snapshots.

Any single observation is noisy. A player can have a bad week for reasons that have nothing to do with development. Repeated measurement against the player's own baseline separates real change from noise, which is why the ERS is built to show direction of travel over twelve weeks rather than grade a single day.

04

Multiple sources beat one.

The score draws on three independent inputs: the mentor's structured notes, the player's own check-ins, and family context. Where inputs agree, confidence rises. Where they disagree, that disagreement is itself useful information, and it goes to the mentor as a prompt to look closer, not to an algorithm to resolve.

The pipeline: five steps from session to score.

STEP 01

Structured capture

Every session produces fixed-format inputs: the mentor's post-session notes, the player's pre-session check-in, and any family context. Free text is allowed, but it sits inside a defined structure, which is what makes reliable extraction possible in the first place.

STEP 02

Signal extraction

A language model reads those inputs and identifies behavioral signals from a fixed, versioned taxonomy for each dimension. Each extracted signal is tied to the specific evidence in the notes that produced it. The model is not forming an opinion of the player. It is detecting whether defined, observable behaviors are present, absent, or changing.

STEP 03

Rubric scoring

Extracted signals are scored against behaviorally anchored rubrics, one per dimension. The criteria are identical for every player and every week, which is the property that makes a "7" in week 2 mean the same thing as a "7" in week 12.

STEP 04

Aggregation and banding

Dimension scores land on a 1-10 scale and roll up into four readiness bands: Below Threshold, Developing, Established, Strong. Banding exists because a family needs a plain-language answer, not a decimal.

STEP 05

Trend modelling and reporting

Scores are stored over time and always read relative to the player's own baseline. Checkpoint reviews at week 6 and week 12 compare where the player is against where they started, and those deltas drive the mentor brief and the parent report.

WHY USE AI AT ALL?

The AI's job is consistency and recall, not judgment.

Mentors are excellent observers and inconsistent raters. That is not a criticism; it is true of every human evaluator ever studied. Applying identical criteria to thousands of observations without fatigue or drift is exactly what machines do well, and it is the only job the model has. It never decides what a score means for a player. Interpreting the trend, adjusting the plan, and talking to the family stay with the mentor and the program, permanently.

How we know the engine does what it claims.

A scoring system working with young people has to be held to a higher standard than "it seems right." Paceful's engine ships with validation and audit infrastructure as core features, not afterthoughts.

VALIDATION INFRASTRUCTURE

DISCRIMINATION & CALIBRATION

A purpose-built validation harness measures whether scores separate what they claim to separate (ROC analysis with confidence intervals) and whether they are calibrated: a score should mean the same thing at the same level, every time.

PER-SIGNAL CONTRIBUTION

For any score, the harness can show which signals drove it and by how much. No score is a black box. If a dimension moves, the evidence behind the movement is inspectable down to the source notes.

SHADOW MODE

New scoring versions run silently alongside the live version before promotion. Outputs are compared before anything changes for real players, so scoring updates never alter a player's record unannounced.

AUDITABILITY

Every score carries a complete audit trail. Records are hash-chained so that any tampering is detectable, and any score in the system can be reproduced from its exact inputs and engine version. If a family or a partner ever asks "why did this number change," there is a definitive, checkable answer.

WHERE VALIDATION STANDS TODAY

We are precise about the difference between what is verified and what is in progress. The pipeline and validation harness have been verified end-to-end on controlled test datasets, which confirms the engine measures what it is designed to measure. Formal outcome validation on real-world program data is an active workstream: a study protocol has been drafted and is deliberately designed for independent review, including external co-authors. We publish claims when the evidence supports them, not before. Until then and after, the ERS operates strictly as decision support.

We do not make clinical or diagnostic claims. The ERS scores observable behavior in a mentorship context, full stop.

We do not claim to predict talent outcomes. The score measures engagement and development habits, not footballing ceiling.

We do not automate decisions. No score triggers any action on its own. Every consequential step runs through a person.

> *Our position is simple: a measurement system earns trust by being open about its evidence, its mechanics, and its limits. This document exists so partners never have to take the score on faith.*

team@mentoramentorships.com · mentoramentorships.com