



References:

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41(3), 586–598. <https://doi.org/10.1037/0022-3514.41.3.586>

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406. <https://doi.org/10.1037/0033-295X.100.3.363>

Gouttebarga, V., Bindra, A., Blauwet, C., Campriani, N., Currie, A., Engebretsen, L., Hainline, B., Kroshus, E., McDuff, D., Mountjoy, M., Purcell, R., Putukian, M., Reardon, C. L., Rice, S. M., & Budgett, R. (2021). International Olympic Committee (IOC) Sport Mental Health Assessment Tool 1 (SMHAT-1) and Sport Mental Health Recognition Tool 1 (SMHRT-1): Towards better support of athletes' mental health. *British Journal of Sports Medicine*, 55(1), 30–37. <https://doi.org/10.1136/bjsports-2020-102411>

Gulliver, A., Griffiths, K. M., & Christensen, H. (2012). Barriers and facilitators to mental health help-seeking for young elite athletes: A qualitative study. *BMC Psychiatry*, 12, Article 157. <https://doi.org/10.1186/1471-244X-12-157>

Kuzmits, F. E., & Adams, A. J. (2008). The NFL combine: Does it predict performance in the National Football League? *Journal of Strength and Conditioning Research*, 22(6), 1721–1727. <https://doi.org/10.1519/JSC.0b013e318185f09d>

Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, 40(6), 998–1009. <https://doi.org/10.1002/ejsp.674>

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1), 75–98. <https://doi.org/10.1037/0033-2909.116.1.75>

Lyons, B. D., Hoffman, B. J., & Michel, J. W. (2009). Not much more than g? An examination of the impact of intelligence on NFL performance. *Human Performance*, 22(3), 225–245. <https://doi.org/10.1080/08959280902970401>

MacLean, A., Young, R. M., Bellotti, V. M. E., & Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Human–Computer Interaction*, 6(3–4), 201–250. <https://doi.org/10.1080/07370024.1991.9667168>

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>

Plews, D. J., Laursen, P. B., Stanley, J., Kilding, A. E., & Buchheit, M. (2013). Training adaptation and heart rate variability in elite endurance athletes: Opening the door to effective monitoring. *Sports Medicine*, 43(9), 773–781. <https://doi.org/10.1007/s40279-013-0071-8>

Rice, S. M., Parker, A. G., Mawren, D., Clifton, P., Harcourt, P., Lloyd, M., Kountouris, A., Smith, B., McGorry, P. D., & Purcell, R. (2020). Preliminary psychometric validation of a brief screening tool for athlete mental health among male elite athletes: The Athlete Psychological Strain Questionnaire. *International Journal of Sport and Exercise Psychology*, 18(6), 850–865. <https://doi.org/10.1080/1612197X.2019.1611900>

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>

Saw, A. E., Main, L. C., & Gastin, P. B. (2016). Monitoring the athlete training response: Subjective self-reported measures trump commonly used objective measures: A systematic review. *British Journal of Sports Medicine*, 50(5), 281–291. <https://doi.org/10.1136/bjsports-2015-094758>

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5, Article 258. <https://doi.org/10.3389/fpubh.2017.00258>

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>

Strathern, M. (1997). "Improving ratings": Audit in the British university system. *European Review*, 5(3), 305–321.
Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO-5 Well-Being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, 84(3), 167–176. <https://doi.org/10.1159/000376585>

White-Paper

The Missing Stat Line

Why the Mind is Ignored by Modern Analytics, and How Properly Capturing It Redefines the Athletic Edge

John Armistead
Co-Founder of AthletesTogether
October 1, 2026. A White Paper.

Abstract

Every part of an athlete's development has a central quantifiable metric except the mental part. Where sport measures the mind at all, it does so once, through snapshot assessments like the Wonderlic and S2, or daily, through tools that capture how athletes feel rather than what they do. This framework paper argues that the missing measure is a daily, action-based record of mental training: what athletes do for their mental game each day, and how well they do it.

Drawing on psychometric principles, it proposes seven requirements any such measure should meet, then describes one implementation, AthletesTogether's Mindset Increase and My Baseline, including the design paths rejected along the way.

It closes with five testable propositions and a research agenda for determining whether daily mental training translates into better wellbeing and performance.

The Missing Stat Line

The Snapshot Problem

In my last paper, I argued that the interface is the product.

This paper is about the number the interface is built around.

Every February, the NFL Scouting Combine brings several hundred of the best football players in the world to Indianapolis and asks them to prove, under a stopwatch, what years of film have already shown. The 40-yard dash, the vertical, the bench press, and the hand measurement each produce a number. Each number moves draft stock, and draft stock is ultimately worth millions of dollars.

Then the prospects sit down with the people who will decide their futures. In a series of short interviews, coaches, general managers, and scouts try to get inside a player's head. They ask unusual questions, watch how he thinks on his feet, and leave with an impression of who he is. Those impressions become scouting reports, and the reports carry real weight on draft day.

Notice the inversion. The traits that film already captures well, like speed, power, and explosiveness, get measured again with laboratory precision. The trait film captures worst, how an athlete thinks, focuses, and responds under pressure, is left largely to conversation and instinct.

To be fair, the league has tried to put a number on the mind. For decades, every prospect sat for the Wonderlic, a 12-minute general cognitive test. Its scores never reliably predicted NFL performance (Kuzmits & Adams, 2008; Lyons et al., 2009), and before the 2022 Combine the league stopped administering it, along with every other psychological assessment. Individual teams now contract their own instruments, such as the S2 Cognition test and the Athletic Intelligence Quotient.

These tools are more sophisticated, but they share the Wonderlic's structural limitation: they are taken once. In 2023, C.J. Stroud's leaked S2 score ranked last among that year's quarterback prospects. He was named Offensive Rookie of the Year.

The problem, then, is not that the mind goes unmeasured. It's that the mind is measured the way we measure height: once, at the door, as if it will never change. The rest of an athlete is measured the way we measure readiness, which is continuously. Sleep, heart rate variability, GPS load, and force-plate jumps are all tracked daily. A strength coach knows whether an athlete is fresher than last week. Nobody knows whether that athlete is sharper.

I lived inside that gap. At Richmond and later at USC, I knew my 40 time and my lifts. At the Olympic Trials, I knew my push times. What no one could tell me, not a coach, not a trainer, not me, was whether on a given Tuesday I was mentally better than I had been a month before. I didn't know whether my focus was improving, whether I was recovering from a bad rep faster, or whether a rough morning at home would cost me that afternoon.

That missing number has consequences. In high-performance sport, what gets measured gets managed, and what doesn't get measured gets nothing. Without a number there is no baseline, no reps designed to move it, no coaching attention, and no line in the budget. The mind isn't neglected because anyone decided it doesn't matter; ask any coach and they'll say it matters most. It's neglected because it never shows up on the sheet. The obvious objection is that the mind is too complex to quantify. Sleep, stress, relationships, and a hundred other daily inputs shift mood, attention, and processing speed from one day to the next. It is, the argument goes, simply too dynamic a system to reduce to a number.

Until it isn't.

The Measurement Landscape

Wearables like WHOOP and the Oura Ring have made heart rate variability one of the most closely watched numbers in sport. HRV is the variation in time between consecutive heartbeats, and it reflects the balance of the autonomic nervous system (Shaffer & Ginsberg, 2017). It shifts with sleep, stress, hydration, illness, and training load, often for reasons no one can identify on a given morning. By the logic of the objection that closed Part 1, it should be far too noisy to be useful.

The Missing Stat Line

The Measurement Landscape (cont.)

Instead, sport science tracks it every day, precisely because it moves. A single HRV reading means little. A trend against the athlete's own baseline reveals a great deal about how the body is adapting and recovering (Plews et al., 2013). The fluctuation isn't noise hiding the signal. The fluctuation is the signal.

That raises the question at the center of this paper. If complexity isn't the barrier, why has no one built the equivalent number for the mind?

The answer lies in how the data is generated. Every measurement in this space comes from one of three sources. Sensed data is collected passively: a device records the body while the athlete goes about their life. Reported data is the athlete describing their own state, whether on a clinical questionnaire or a morning mood slider. Performed data is generated when the athlete does something and the doing itself is scored, like a lift, a sprint, or a cognitive task. Crossing these sources with how often each is collected produces the landscape in Figure 1.

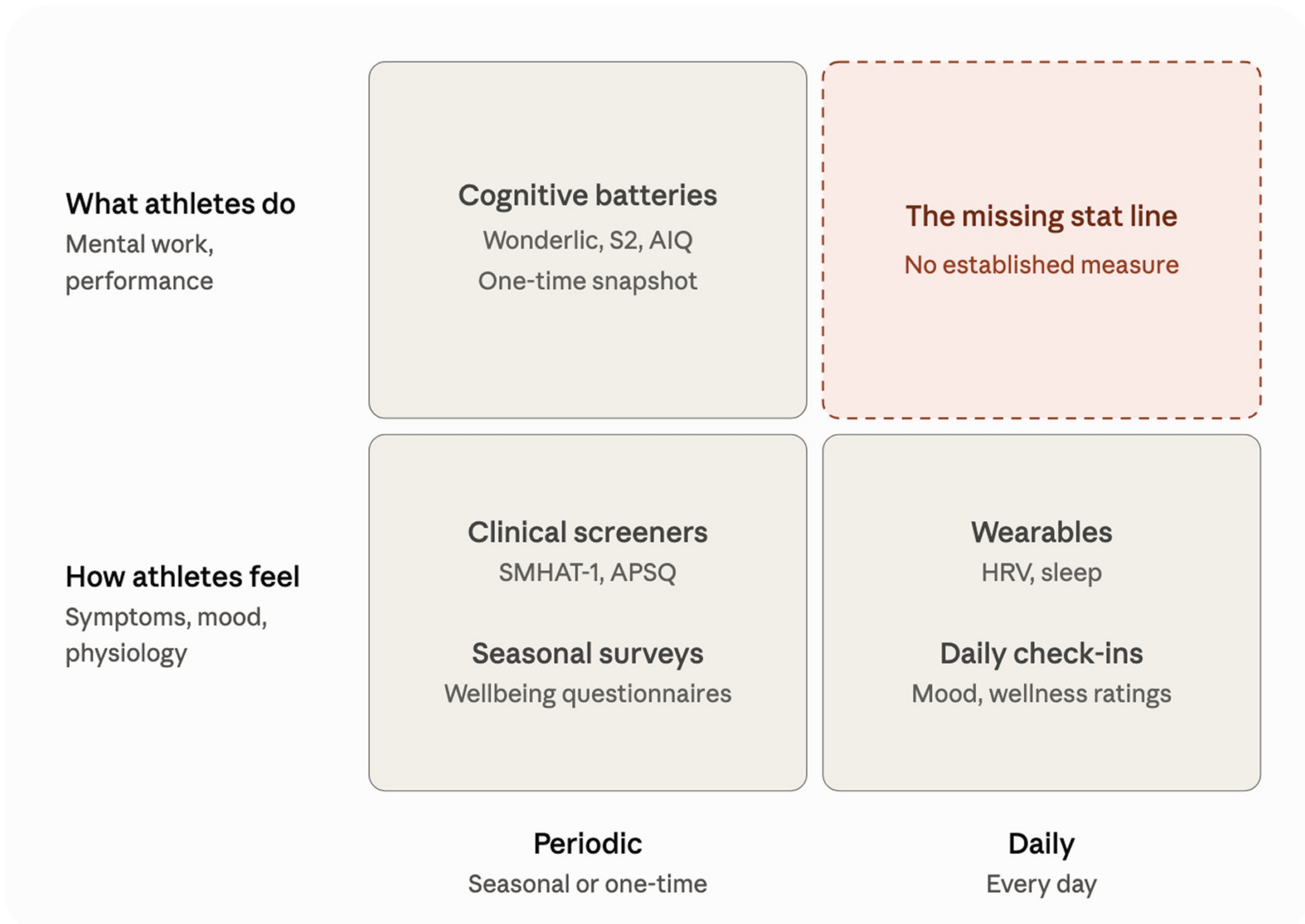


Figure 1

Clinical screeners.

Instruments such as the IOC's SMHAT-1 (Gouttebarga et al., 2021), the Athlete Psychological Strain Questionnaire (Rice et al., 2020), and the GAD-7 (Spitzer et al., 2006) are well validated for what they were built to do: identify athletes who may need clinical support.

Nothing in this paper argues against using them. But they are designed to detect illness, not development. A psychologically healthy athlete scores in the normal range and stays there, a ceiling effect that makes the number unable to register growth. A screener can tell you an athlete isn't struggling. It can't tell you they're getting better.

Seasonal surveys.

Wellbeing measures such as the WHO-5 (Topp et al., 2015), along with the preseason and postseason questionnaires many programs build themselves, widen the lens beyond illness. But they are administered a few times a year. They ask athletes to recall weeks or months of experience, which introduces recall bias, and long batteries create response burden that erodes data quality. Most importantly, they produce a report for staff, not feedback for the athlete. By the time the results are analyzed, the stretch of season they describe may already be over.

Cognitive batteries.

Part 1 covered these: the Wonderlic, S2, and AIQ. Because they generate performed data, they are the closest existing tools to a true mental stat line. But they are administered once, usually at high stakes, and they measure capacity, not training.

Wearables.

Sensed data solves the frequency problem completely; it arrives every day without the athlete doing anything. But it measures the body's response to stress, not the athlete's mental skill. Inferring focus, composure, or confidence from HRV is a construct validity problem. A low reading might reflect a hard session, a poor night of sleep, or a stressful week, and it says nothing about whether the athlete handled that stress well.

Daily check-ins.

Mood and wellness ratings bring reported data up to daily frequency, and the evidence for them is stronger than skeptics assume: self-reported measures have tracked athletes' training responses more sensitively than many objective markers (Saw, Main & Gustin, 2016). But a check-in captures state, not work. A low mood after a brutal training day isn't a failure, and a high one isn't an accomplishment. The athlete reports how they are; the number can't reflect what they did about it.

The Missing Stat Line

The Measurement Landscape (cont.)

Streaks.

The closest thing to the empty quadrant is the streak. Meditation and wellness apps count daily sessions, which makes them both daily and action-based. But a streak measures attendance, not quality. Ten distracted minutes count the same as ten focused ones, and nothing in the number shows whether the athlete is getting better at the skill.

That leaves the empty corner of Figure 1. The field can capture how athletes feel, occasionally or daily, and it can test what they're capable of, once. It has no established way to measure what athletes do for their mental game every day, and how well they do it.

This gap matters because of how physical training works. In a weight room, measurement and training are the same act. A logged set is both the rep and the data point; no one runs a separate assessment to find out whether the athlete trained. Progress shows up through accumulated, quantified repetition long before any test is run. Mental training has no equivalent. There is no rep that is also a data point.

Without one, the only signal left to see is breakdown. When an athlete struggles badly enough, a screener flags it and a clinician steps in, so the industry has organized itself around rescue: crisis resources, clinical referrals, reactive care. That care matters. But no strength coach waits for a torn ACL to start an athlete lifting, because strength progress is visible long before anything breaks. The mental game will keep being treated as crisis aversion until its progress can be seen the same way.

Requirements for a Mental Box Score & Psychometric Progression

Measuring the mind falls within psychometrics, the branch of psychology concerned with measuring what can't be observed directly. A tape measure can capture a vertical jump and a scale can capture body weight, but there is no instrument to hold up against focus, composure, or resilience. These are latent constructs: real qualities that can only be inferred from observable behavior. Psychometrics exists to build instruments that turn observable responses into numbers, then to prove those numbers actually track the quality underneath.

Over the past century, the field has converged on three questions every instrument must answer.

The first is reliability: is the measure consistent? A reliable instrument returns the same result when nothing has changed. A hand-timed 40 is less reliable than a laser-timed one, because two scouts with stopwatches will disagree about the same run. For a daily training metric, reliability takes a specific form. The score is supposed to vary, because athletes do different amounts of work on different days. What it must not do is vary for any other reason. The same work, done the same way, should earn the same credit.

The second is validity: does the measure capture what it claims to? Reliability alone isn't enough. A scale that reads 15 pounds heavy is perfectly consistent, and consistently wrong. Validity asks whether a number means what its designers say it means and whether it relates to the outcomes that matter. The Wonderlic produced consistent scores; it failed on validity, because those scores didn't predict performance.

The third is responsiveness: can the measure detect change? A scale that reads only in 10-pound increments is useless for tracking a weight cut, because real progress happens below its resolution. Clinical screeners, as Section 2 showed, fail here through their ceiling effects.

Contemporary measurement frameworks such as COSMIN (Mokkink et al., 2010) add two practical considerations: interpretability, meaning whether users can understand what a score means, and feasibility, meaning whether people can realistically complete it. A measure no one understands or uses measures nothing.

Meeting these criteria is difficult for any psychological construct. Meeting them every day, for something as multidimensional as an athlete's mental game, is harder still. The seven requirements below translate these criteria into the specific demands of a daily mental training metric. They emerged from our own iterative design work and are offered as a practice-derived framework: each is grounded in existing research, and each is stated so that any approach, including ours, can be evaluated against it.

The Missing Stat Line

Requirements for a Mental Box Score & Psychometric Progression (cont.)

1. Daily. The measure updates at least once per day and reflects that day's activity. This is responsiveness applied to time. Habits form through consistent repetition over weeks (Lally et al., 2010), and a measure that updates seasonally can't move at the speed habits are built. Section 2's HRV case shows that daily fluctuation is information, not noise. Without this requirement, the measure suffers from lag: by the time the number arrives, the behavior it describes is weeks old, and the athlete has nothing to act on.

2. Action-based. The measure counts the mental training an athlete completes, the way a training log counts sets, rather than scoring how the athlete feels. This is a validity requirement: if the construct is mental training, the measure must capture training. Section 2 showed that reported state is confounded with training load and life circumstances; a low mood after a brutal session says nothing about whether the athlete did the work. Without this requirement, the measure conflates state with effort, crediting easy days and penalizing hard ones regardless of what the athlete actually did.

3. Quality-weighted. Doing the work well counts for more than simply showing up. Research on expert performance holds that improvement comes from focused, effortful practice, not accumulated hours (Ericsson, Krampe & Tesch-Römer, 1993). Section 2's analysis of streaks shows what happens when quality is ignored: ten distracted minutes count the same as ten focused ones. Without this requirement, the measure rewards attendance over improvement and becomes trivially easy to game.

Together, the second and third requirements resolve a tension that runs through any effort-based measure. Counting completed work shows that an athlete is busier; on its own, it doesn't show that they are sharper. Strength and conditioning solved this long ago by tracking two kinds of numbers. The training log, a process measure, shows that the work happened. Personal records, an outcome measure, show that the work paid off. A coach with only the log can't tell whether the training works, and a coach with only the PRs has nothing to guide daily coaching. A mental box score needs both: the volume of work completed and the quality of that work measured against the athlete's own history. That last phrase matters.

Training is about within-person change, so the comparison should be self-referenced rather than normative. The question isn't how an athlete ranks against a draft class, but whether they are better than they were last month.

4. Non-clinical. The measure tracks training and development, never illness, and never diagnoses. This sets the boundary of the construct, which makes it a validity requirement: a training measure that drifts into symptom detection is no longer measuring training. It is also a practical requirement. Stigma remains one of the most consistent barriers to athletes seeking mental health support (Gulliver, Griffiths & Christensen, 2012), and tools framed around illness reach fewer athletes than tools framed around performance. Without this requirement, the measure inherits both the ceiling effect of clinical screeners and the reluctance athletes bring to them.

5. Honest, not punitive. The measure neither shames a bad day nor inflates progress into a participation trophy. Self-determination theory holds that feedback supporting an athlete's sense of competence sustains motivation, while controlling or shaming feedback erodes it (Ryan & Deci, 2000). But a score that can never fall carries no information; progress has to be real and earned. Without this requirement, the measure fails in one of two directions: athletes disengage from a number that punishes them, or they stop trusting a number that flatters them.

6. Two coherent horizons. The measure gives athletes both a daily target to act on and a long-term number that reflects accumulated development, and the two never contradict each other. Goal-setting research shows that near-term goals build self-efficacy and sustain effort toward long-term ones (Bandura & Schunk, 1981). A single number can't serve both purposes: one that resets daily has no memory, and one that accumulates slowly gives no reason to act today. Without this requirement, the measure offers either no daily motivation or no sense of who the athlete is becoming.

7. Private and burden-free. Individual-level data stays with the athlete, organizations see aggregate trends, and the measure requires no data entry beyond the training itself. Privacy is a validity requirement as much as an ethical one: athletes who know they are being watched shift their behavior toward what they think observers want, an effect known as social desirability bias. Low burden is a feasibility requirement: every added step reduces completion, and incomplete data is biased data. Without this requirement, athletes perform for an audience, staff inherit new administrative work, and the measure gradually stops reflecting reality.

The Missing Stat Line

Requirements for a Mental Box Score & Psychometric Progression (cont.)

These requirements don't sit together comfortably, and any honest design has to choose how to balance them. Four tensions matter most.

Honest pulls against non-punitive. A score that can never go down can't signal a problem, and a score that punishes bad days drives athletes away precisely when they most need support. Every design picks a point between the two. Daily pulls against burden-free. In most measurement, frequency and burden rise together. That changes only when the measurement is the training: when, as with a logged set in the weight room, completing the rep generates the data point. Under that design, daily frequency costs the athlete nothing extra.

Action-based pulls against quality-weighted. Counting completed work is simple and highly reliable. Judging its quality is more meaningful, but it introduces scoring judgment, and with it the potential for error. Every quality criterion has to be specified precisely enough to be applied identically every time.

Non-clinical pulls against duty of care. A training measure must never diagnose, but it may still surface signs that an athlete is withdrawing. A non-clinical tool needs a clear pathway for connecting athletes with qualified professionals, even though it never makes that clinical judgment itself.

Figure 2 applies these requirements to the approaches surveyed in Section 2. Each meets some; none meets most. Clinical screeners and seasonal surveys fail on frequency and responsiveness. Wearables and daily check-ins solve frequency but measure the wrong construct. Cognitive batteries come closest on construct but are administered once.

The requirements are demanding and, as the tensions show, partly in conflict. No measure will satisfy all seven perfectly. The next section describes one attempt to balance them, along with the design paths we rejected along the way.

Case Study

How AthletesTogether is Building It

This section describes AthletesTogether's own implementation, written by the people who built it. It's offered as one attempt to meet the requirements in Part 3, not the only one, and the reasoning behind each decision is laid out so readers can judge it for themselves.

To keep that reasoning visible, each component below follows the same structure: the decision we made, the requirement it answers, and what it cost us. Design researchers call this documenting design rationale, which means laying out the questions a design had to answer, the options considered, and the criteria used to choose between them (MacLean et al., 1991). Here, the criteria are the seven requirements from Part 3.

The system has four parts. Athletes complete mental training in the app. Every completed activity earns credit toward a daily score, Mindset Increase. Daily scores accumulate into a long-term number, My Baseline. Once a week, the Film Session recaps what the athlete built. Alongside all of it, coaches and staff see aggregate trends for their team.

Mindset Increase: the daily number

Every athlete starts every day at +0%.

From there, the score builds through what the athlete does, up to a cap of 100%. Seven performance games train attention, reaction time, memory, and cognitive flexibility. Six wellness activities train skills like reframing, control orientation, and earned confidence. The Daily Warm Up pairs a mood check with a 60-second themed audio session and an intention for the day. Habits, journaling, Playbooks, and meditations round out the inputs.

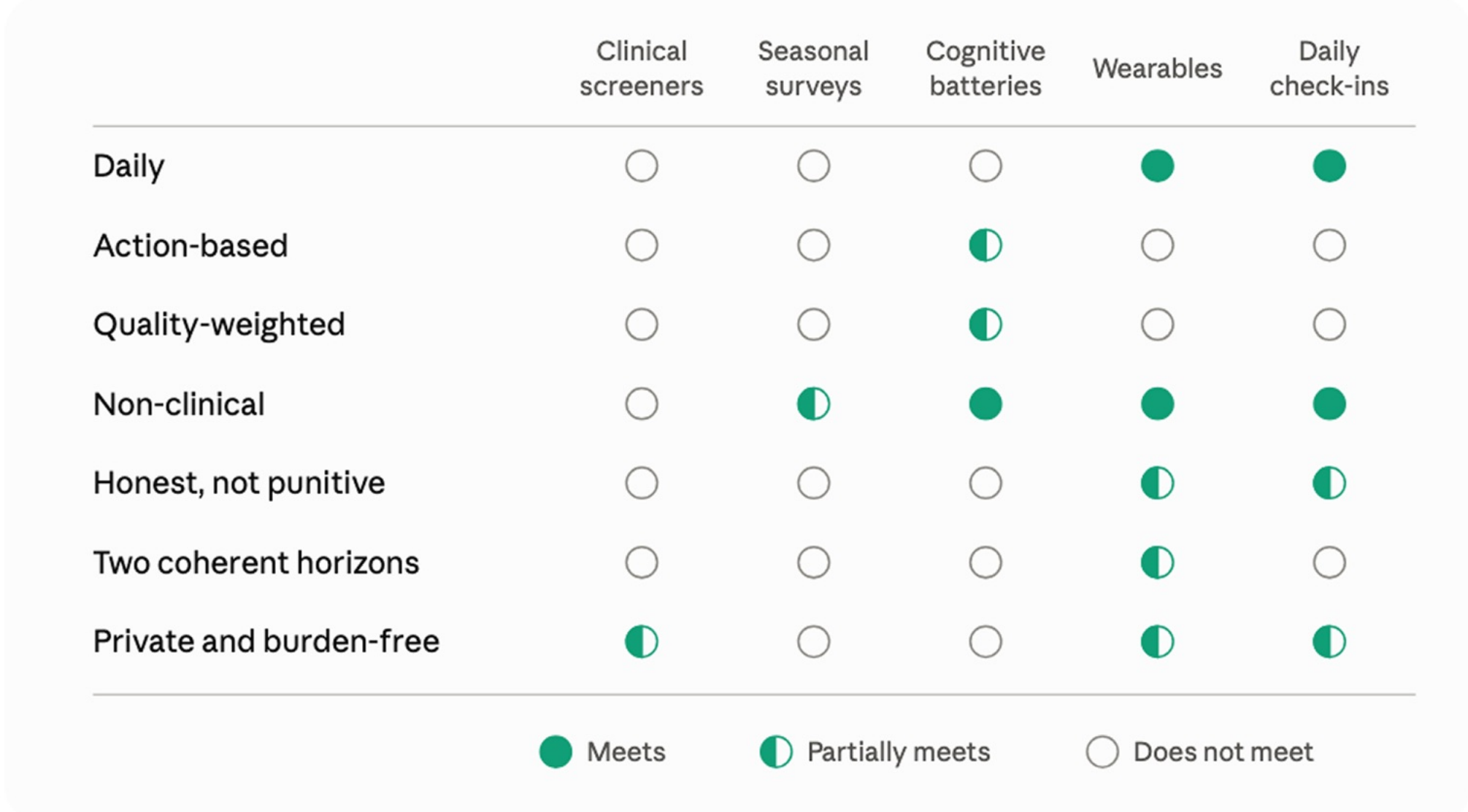


Figure 2

The Missing Stat Line

Case Study How AthletesTogether is Building It

The daily reset answers the daily requirement. Yesterday doesn't carry over, good or bad, so every morning is a full opportunity. The inputs answer the action-based requirement: the score counts mental work completed, the way a training log counts sets.

That includes the mood check in the Daily Warm Up. The rating is recorded for context, but it doesn't affect credit. An athlete who reports a rough morning and does the work earns the same as one who reports a great one. If mood moved the score, we'd be measuring how athletes feel, which is exactly what Part 2 argued against.

The cost is coverage. Mindset Increase only sees what happens in the app. An athlete who does excellent work with their team's sports psychologist, or journals in a notebook, gets no credit for it. The score measures a slice of an athlete's mental training, not all of it, and we'd rather say that plainly than have someone point it out.

Not every rep counts the same.

Quality enters the score in two ways. In the performance games, credit rises when an athlete beats their own historical baseline. A personal best counts for more than a routine round, because improvement relative to yourself is the closest thing the mind has to a PR.

In the wellness activities, each response is evaluated against defined criteria, and credit scales with how well it meets them. The Controllables Filter shows how this works. The athlete lists three to five stressors taking up space and sorts each one into the Arena (things they can control) or the Noise (things they can't). Each classification is compared to its best fit, with a short rationale the athlete can learn from. Strong filtering earns roughly 10%. Weak filtering earns roughly 3%. Both count; one counts more.

This answers the quality-weighted requirement, and it resolves the sharper-versus-busier tension from Part 3. Volume shows the athlete trained. Quality, measured against their own history, shows whether the training is paying off.

The cost is judgment. Counting completed work is perfectly reliable. Judging its quality isn't. Any scoring system can misjudge a response, which is why standardized scoring systems and algorithms are put in place by licensed and verified sports psychologists, neuroscientists, and psychiatrists.

My Baseline: the long-term number

Daily work has to add up to something.

My Baseline is the long-term number, and it's built on a deliberately different scale. Each day contributes points based on that day's Mindset Increase: the daily percentage divided by 20, rounded, so each day is worth between 0 and 5 points. The number updates every day and climbs a ten-tier ladder called The Climb, from Rookie to GOAT.

The separate scale is intentional. If My Baseline used the same 0–100% scale as the daily score, athletes would read it as an average of their days, and one great day would look like real development. Different scales signal different meanings: one tells you what you did today; the other tells you who you're becoming. That's the two-coherent-horizons requirement. The two numbers measure different things, but they never contradict each other, because one is built entirely from the other.

This is version one, and it's deliberately simple. My Baseline doesn't apply quality adjustments of its own yet, but quality still flows through it, since every day's points come from an already quality-weighted daily score. More sophisticated logic is planned for later versions.

The cost is cognitive load. Two numbers ask more of an athlete than one. We mitigate that with plain explainer copy and a direct path from the daily score to the Film Room, where My Baseline lives, so athletes always see how the two connect.

The day-one athlete

Picture a new athlete on their first day. They do everything: a Daily Warm Up, two performance games, a Controllables Filter, a WIN List. They hit 100% on Mindset Increase.

Their My Baseline goes up by 5 points. They're still near the bottom of The Climb.

The Missing Stat Line

Case Study How AthletesTogether is Building It

That's the point. One day is a full effort, but it isn't a history. The daily score honors what they did today. The long-term number waits for them to do it again.

My Baseline

+21 +1 today

Starter



Every session you complete adds to this. My Baseline shows how much you've improved since you joined.

The Climb, and the honesty problem

Part 3 named a tension between honest and non-punitive measurement. The Climb is where we had to choose.

Once an athlete reaches a tier, it stays unlocked. The My Baseline number itself responds to inactivity. If an athlete stops training, the number gradually declines, but the tier they earned stays. The tier honors what they've accomplished; the number tells the truth about where they are now.

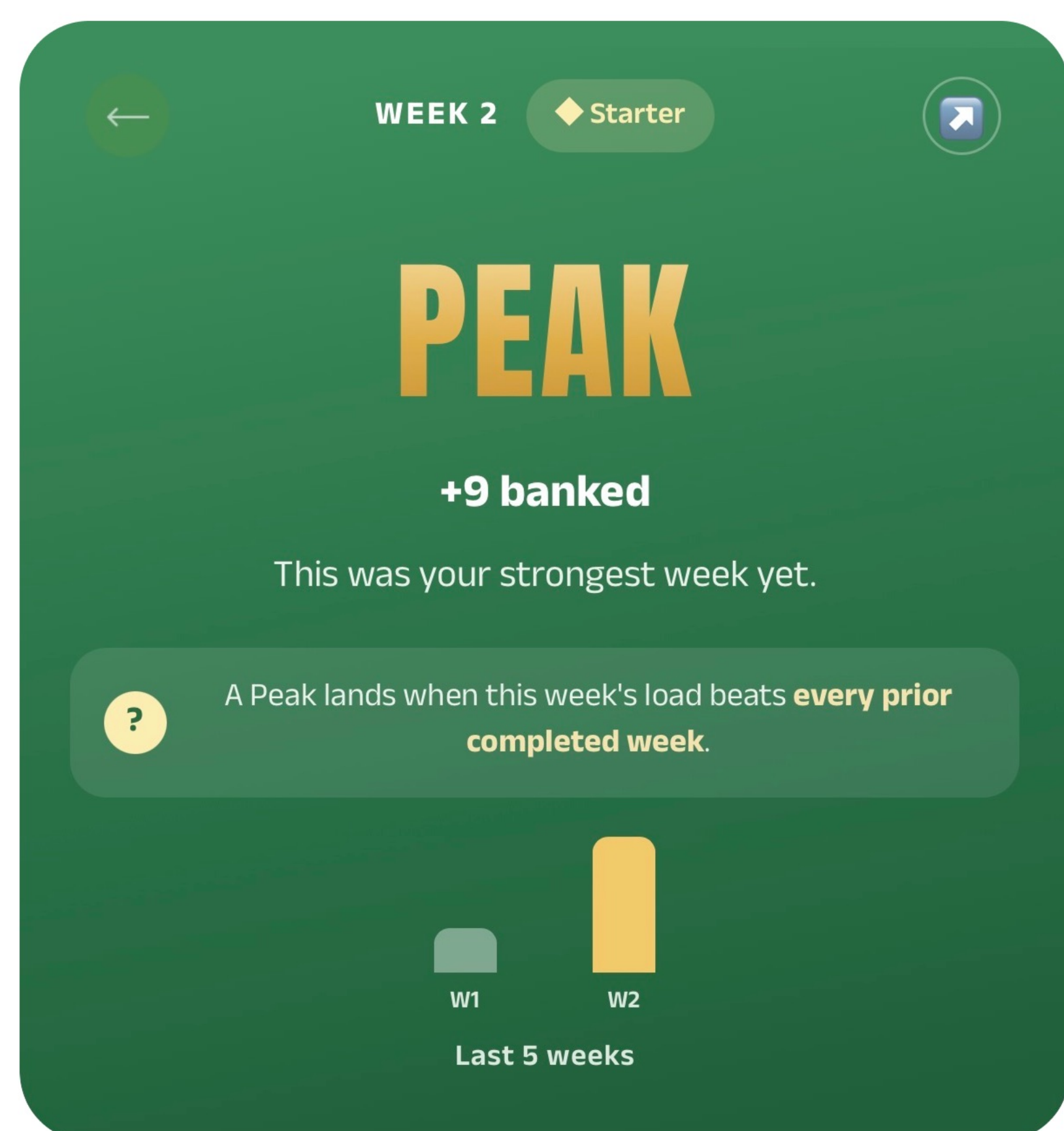
We also made a deliberate messaging choice. We don't frame the app around "you can't lose." That reads as a participation trophy, and athletes know the difference between a score they earned and one they were handed. Progress has to be real to mean anything. Tier thresholds are calibrated so the upper tiers are reachable within a college career, which keeps the top of The Climb demanding but within reach.

The Film Session

Every Sunday at 5pm, athletes get their Film Session: a recap of the week they just put in, with a name for that week. It isn't a score reveal. My Baseline has been moving all week. The Film Session is a moment to sit down and watch the tape.

That's deliberate. Habits form around anticipated, recurring moments, and the Film Session gives each week a natural endpoint and the next week a starting line. In my last paper, I argued that the interface determines whether athletes use a tool at all. The Film Session is that argument applied: a ritual worth coming back for.

The cost is timing. A fixed moment can't fit every schedule, and plenty of athletes compete on Sunday afternoons. The Film Sessions stay available and are logged as history for re-referencing.



What we rejected

The design above is the product of options we built, tested, or seriously considered, and then walked away from. The following four features shaped it most.

The Missing Stat Line

Case Study How AthletesTogether is Building It

A daily readiness score.

Our original daily metric was a mental readiness score. The appeal was obvious. Athletes already trust recovery scores from their wearables, and a mental equivalent felt intuitive: wake up, check your number, know where you stand. We replaced it because it measured how athletes felt, not what they did. A low readiness score after a brutal training week isn't a failure, and it gives the athlete nothing to act on. It broke the action-based requirement, and it ran straight into the argument of Part 2. We didn't just make that argument in this paper. We rebuilt our own product around it.

Weekly wins and losses.

An earlier version of the Film Session gave each week a result (win, draw, or loss) inside a twelve-week season, with promotion and relegation between tiers. The appeal is real. Athletes understand winning and losing instinctively, and competition motivates. But the threshold for a win was arbitrary, and a record like 15–4 says almost nothing about whether an athlete is developing. Athletic departments were also uneasy about a tool where their athletes could "lose" a week of mental training. And it added a third, partially related number alongside the daily score and the long-term baseline. It broke the honest-not-punitive and two-coherent-horizons requirements. We kept the ritual and dropped the scoreboard.

One combined score.

The simplest design would have been one number that blended today's effort with long-term progress. Simplicity matters, and one number is easier to read than two. But blending them muddies both. A great day inflates the long-term picture, and a long history makes today's effort feel invisible. It broke the two-coherent-horizons requirement.

Weekly-only updates.

We also considered updating the long-term number only once a week, at the Film Session, so the reveal would feel like an event. Anticipation is a powerful engagement tool. But making athletes wait until Sunday to see any movement removes the reason to come back on a Tuesday. So My Baseline moves a couple of points a day instead. Weekly-only updates would have broken the daily requirement, and they cut against the adherence argument at the center of my last paper.

The team layer

A mental box score matters beyond the individual athlete. Coaches, sports psychology staff, and club directors have historically had almost no data on the mental side of their programs. The AthletesTogether admin portal gives them that, but only in aggregate. Staff see team-level trends: participation, activity patterns, and how their group's training is moving over time. They never see an individual athlete's scores, journal entries, or responses.

This is a measurement decision as much as an ethical one. If athletes knew their coach could read their Controllables Filter, they'd write what they think the coach wants to see. That's social desirability bias, and it would quietly corrupt every number in the system. Privacy is what keeps the data honest. For organizations working with athletes under 18, parents and guardians consent to the same model: individual data stays with the athlete, and staff see the team.

The portal also requires no new input. Every module is computed from what athletes already do in the app. Staff don't log events, enter schedules, or fill out forms. That's the burden-free half of the requirement, and it matters: a tool that creates work for staff is a tool staff stop using.

The cost is the hardest trade-off in the system. Aggregate-only means a coach can't see that one specific athlete is struggling. That's the tension between non-clinical design and duty of care from Part 3. A design is only as good as its use in the real world. The next section turns from design to evidence: what we've observed so far, and what remains to be tested."

From Framework to Evidence

This paper is a framework, not a study. It identifies a gap in how sport measures the mind, sets out what a measure built to fill that gap would require, and describes one attempt to build it. What it doesn't do is prove that the approach works. That's the next step, and this section lays out what it would take.

Some early observations from the field have shaped the questions below. These are observations from building and operating the platform, not findings. They come from usage patterns and athlete feedback, not from controlled study.

The Missing Stat Line

From Framework to Evidence (cont.)

The clearest came from notifications. When we replaced fixed daily reminders with personalized ones that surfaced something specific to each athlete's own training, athletes came back more often. The pattern we saw, and that athletes described to us, looked like a loop: a personal stat sparks curiosity (Loewenstein, 1994), curiosity brings the athlete in, training produces visible progress, and progress makes the experience worth repeating. Over time, it becomes routine. We think that loop is real. We haven't proven it.

The broader point is the one this paper has made from the start: athletes engage with a number that moves, and a number that reflects their own work. What we don't yet know is what that engagement does for them.

Five testable propositions

The framework in Part 3 makes claims that can be tested. We state them as propositions so any researcher or organization can test them, with or without us.

Proposition 1: Responsiveness. A daily, action-based measure will detect within-athlete change earlier than periodic instruments do. This can be tested by tracking both kinds of measures concurrently over a season and comparing when each registers change.

Proposition 2: Convergence. Accumulated, quality-weighted mental training will be positively associated with validated measures of wellbeing over time. This can be tested by collecting established wellbeing instruments, such as the WHO-5 (Topp et al., 2015), at regular intervals alongside daily training data.

Proposition 3: Performance. Sustained mental training, measured by both volume and quality, will predict performance outcomes beyond what physical training load explains. This is the hardest proposition to test and the most important. It requires longitudinal data that pairs mental training with sport-specific performance measures.

Proposition 4: Adherence. Measures that give athletes daily feedback on their own progress will sustain engagement longer than measures that report only to staff. This can be tested by comparing engagement under each condition.

Proposition 5: Honesty. Athletes will train and report more candidly when staff see only aggregate data than when staff can see individuals. This can be tested by comparing response patterns under different visibility conditions.

What the framework can't settle on its own

The biggest open question is whether daily mental training translates into better athletes. Every proposition above leads back to it. Until it's tested, the honest claim is that a mental box score is measurable and usable, not that it makes anyone better.

The framework also carries a built-in risk. Any measure that becomes a target invites gaming; as Strathern (1997) put it, when a measure becomes a target, it ceases to be a good measure. Quality weighting and self-referenced baselines make that harder, but they don't make it impossible, and any system built on this framework has to keep testing whether its numbers still mean what they're supposed to.

What comes next

We're beginning to test these propositions ourselves. Our study with Austin FC examines wellbeing and performance before and after platform use, with results expected in November 2026, to be published as a follow-up to this paper. Structured onboarding with additional partners will give us cleaner comparisons over the coming season.

But these questions are bigger than one company. We'd welcome researchers and organizations who want to test them with us, or without us.



The Missing Stat Line

The mind is measured once, like height. What sport does measure daily captures how athletes feel, not what they do for their mental game. This paper argued for a different kind of measure, one that counts daily mental work, weights its quality, and builds over time into a record of who an athlete is becoming. It laid out the requirements such a measure has to meet and described one attempt to meet them.

The evidence so far is early. Athletes are using it, and they're telling us it matters to them. Whether that training makes them better on the field is the question ahead, and we've laid out how we intend to answer it.

But picture the alternative to what exists now. A prospect walks into Indianapolis with more than a 40 time and a 15-minute impression formed across a table. They walk in with a season of mental training data: every rep, how well they did it, and how far they've come.

The future of athlete mental health depends on our ability to create it.

