Does Whoop measure or construct recovery, and what are the implications?

Sep 20, 2026 · Oliver Chen
Stanford University · Class of 2030

Memo #1 in a series on human-state intelligence and the business implications behind society

The short answer

WHOOP constructs recovery because it is fundamentally impossible to measure a “Recovery Unit.” It is worth noting that the WHOOP band effectively measures overnight heart-rate variability (HRV) nearly as well as a clinical chest strap. However, this clean signal is then folded into a 0–100% score with weights the company has never disclosed, mixing in various factors like sleep staging that only "fairly" agrees with a sleep lab. WHOOP also tunes its own training advice to move the score's own main input.

In ten years the company has never published a peer-reviewed test of the score against the thing it says the score predicts: how ready you are to perform that day. The one independent study that comes close found that WHOOP’s raw HRV tracked athlete stress effectively, but it was noted that the Recovery score built on top of it tracked nothing (Lundstrom et al. 2023).

That is not necessarily a scandal on WHOOP’s side. In fact, in today’s market it might even be the right product decision. Unlike directly measured signals like heart rates or skin temperature, nobody has ever agreed on a set way to measure recovery. Anyone selling a recovery number is thus constructing one. What makes WHOOP worth a memo specifically is the business itself. As of March 2026, WHOOP reports a $10.1 billion valuation, cash-flow positive, and 2.5 million members opening the app eight times a day to read a colour based on a recovery score. (Whoop, Mar 2026). The score’s value doesn’t come from being objectively correct, but rather from being believed, and I’ll argue that validating it could make it worth less to WHOOP, not more. Final position and two dated calls at the end of the memo.

What the score is

WHOOP computes recovery once from what the band saw overnight, a score the user can see immediately when they wake up. WHOOP’s developer documentation lists that the factors used to determine the score are the user’s resting heart rate (RHR), HRV, respiratory rate, sleep duration and quality, skin temperature, and blood oxygen (Whoop 101) – some of which are measured more accurately than others. Whoop's own explainer from July 2026, however, lists a slightly different set. They specify on HRV, which is listed to have "the greatest influence," sleeping heart rate, Sleep Performance, and an integrated "illness prediction model" (Whoop, Jul 2026). Either way, the company has not yet stated consistently what goes into a number 2.5 million people wake up to, and, more importantly, has not released the specific weighting of each factor used to calculate the final score.

The HRV metric is the Root Mean Square of Successive Differences (RMSSD), which is taken from a five-minute window in the last slow-wave-sleep period of the night (API docs; Whoop research FAQ). This, alongside the other inputs, are composed by undisclosed weights into a single 0-100% score. As noted earlier, WHOOP's weights are proprietary, though this is also reflected in competing companies where a 2025 review of 14 composite scores from 10 brands found that none disclosed the formulas they use (Doherty et al. 2025). The baseline is personal as WHOOP grades your HRV against your own recent history, and the length of the rolling window, the span of past days that constitutes your baseline, has never been stated. For the colour-code, green is 67% and up, yellow 34–66%, and red is 33% and below (Whoop 101).

WHOOP's definition for the meaning of this score is not modest. The company states that it is "a daily measure of how prepared your body is to perform" (Recovery 101). In a podcast, WHOOP's VP of Data Science, Emily Capodilupo, explains the design target even more plainly, noting:

> *"The goal of the Recovery Score is to measure how ready your body is to adapt to a training stimulus. This is highly correlated with how ready you are to perform." (Whoop podcast #84)*

A significant part of the memo tests this exact claim, so hold on to it.

What "measuring" would require

To say a score measures recovery, you need to have something to check it against. More specifically, I define measuring as a device detecting a static, objective reality that exists in the physical world, independent of whether the device is looking or not. The way recovery is looked at thus far is typically performance-shaped—how high you jump on a force plate the next morning (the most common neuromuscular fatigue test in elite sport), a validated questionnaire like the RESTQ, or even a lab marker that has known repeatability. Whoop's own definition picks the first kind, because the recovery score is supposed to track "how ready you are to perform."

Sports science explicitly highlights that there is no single criterion for measuring recovery. The 2018 consensus statement on recovery and performance calls the concept of recovery "multifaceted," noting various branches such as "physiological" and "psychological," furthermore warning that any "single physiological or psychological parameter will only highlight an isolated aspect" (Kellmann et al. 2018). A review of 56 studies that measured athletes both ways found that "subjective and objective measures of athlete well-being generally did not correlate," and that the subjective ones "reflected acute and chronic training loads with superior sensitivity and consistency than objective measures" (Saw, Main & Gastin 2016). In other words, the evidence showed that asking an athlete how they feel beats measuring their heart.

Not having a way to objectively measure recovery does not excuse WHOOP. Instead, the implications of our original question are sharpened. With no criterion existing, a 0–100% number with a colour attached is a decision the company makes to let you see. That is neither a new discovery about your body, nor an objective marker detected by the strap itself. In turn, WHOOP's recovery score must also be judged not merely as a clinical product, but also as a behavioural one. Every section below follows from that thought.

What studies have shown, and what gets lost

The band itself measures biomarkers, and for signals like HRV and RHR, it is great at doing its job. An independent testing at Ohio State, which consisted of 13 adults over 536 nights using WHOOP against a clinical ECG, put Whoop 4.0's overnight HRV at a concordance of 0.94 (error about 8%) and resting heart rate at 0.91 (Dial et al. 2025). Although WHOOP's major competitor, OURA, performed better with its Gen 4 ring (concordance of 0.99), this is a real measurement of a real signal. It is also the input that WHOOP says matters most in their weighting.

The recovery score is where things get cloudy. There are three different places in which information could get inaccurate: in the composite weighting of various factors to determine the score, in turning daily measurements into a score, or in the implications of the recovery score itself. My concern is predominantly in the composite, though the studies below do not tell us where, or whether, information was lost. Refer to the table below for the various tests conducted, and how implications from raw signals versus recovery scores compare.

What was tested	Results	Funding source
Overnight HRV vs ECG, Whoop 4.0 (Dial 2025)	Concordance 0.94, error 8.2%	Independent
Resting heart rate vs ECG, Whoop 4.0 (Dial 2025)	Concordance 0.91, error 3.0%	Independent
Sleep stages vs sleep lab, Whoop 4.0, 62 adults (Schyvens 2025)	Agreement "fair" (kappa 0.37); deep sleep overestimated by 31 min, 60% of wake missed	Independent
Recovery score vs RESTQ questionnaire, 23 D1 swimmers, 6 weeks (Lundstrom 2023)	HRV alone tracked stress (r = −0.46); the Recovery score tracked no stress or recovery variable in either sex	No funding
Season-average Recovery vs golf performance, 389 male professional golfers (Grosicki et al.)	Higher Recovery associated with better performance, between golfers and within golfers across seasons	WHOOP-affiliated authors
Overnight HRV vs daily wellbeing, 39 adults, ~85 days each, within-person (Ungaro 2026)	Days people felt "energized" had lower HRV (64.1 vs 66.9 ms); no link to stress	PepsiCo/Gatorade

WHOOP has released tests of their Recovery score against performance. In a peer-reviewed study of 389 male professional golfers, higher average Recovery was associated with better golf performance, both between players and within the same player across seasons. All authors listed WHOOP affiliations. The important detail is that the researchers used seasonal averages. This tells us something about recovery and performance over a season, but not how accurately this morning's score predicts this afternoon's round (Grosicki et al.). WHOOP also ran a basketball case study in 2016, comparing seven Division 1 players' shooting across 24

games against their Recovery score that same morning (Whoop, 2016). It is the closest anyone has come to a day-N test, but it was never peer-reviewed, never compared the score against HRV alone, and the full report is no longer available, with only a summary article remaining.

Another study of 23 Division 1 swimmers, independent of the WHOOP organization, looked at how stress, recovery and metabolism corresponded with WHOOP's HRV measurements and recovery scores. HRV alone correlated significantly with general and total stress in men, but not in women analyzed separately. The recovery score, however, was found to have no significant associations with the questionnaire's stress or recovery measures in either sex, consistently returning flat coefficients no further from zero than 0.20. The important detail is that the researchers did not test whether the HRV and Recovery correlations differed significantly from each other. This tells us that some associations found in HRV were absent from the final score, but not that the calculation removed useful information. Recovery also correlated with the ratio of measured to predicted resting metabolic rate in men, so it did show an association with one metabolic measure (Lundstrom et al. 2023).

RESTQ measure	HRV (M)	HRV (F)	Recovery (M)	Recovery (F)
General Stress	-0.72	0.31	-0.02	0.20
Sport-specific stress	-0.60	-0.27	-0.17	0.07
Total stress	-0.74	0.11	-0.07	0.20

n = 10 males, 13 females

The first obvious suspect is what gets mixed in when calculating the recovery score. Independent sleep-lab testing found only "fair" sleep/wake agreement for Whoop 4.0, which overestimated deep sleep by an average of 31 minutes and misclassified approximately 60% of wake epochs as sleep during overnight testing (Schyvens et al. 2025). Sleep is one of the factors that goes into calculating the score additional to the HRV. What this study does not tell us is how much those errors move the final score.

Researchers have also raised concerns regarding double-counting in scores that use both sleep and HRV, which is exactly what WHOOP does. In a review of 14 composite scores from 10 brands, combining sleep and HRV was identified as a potential source of redundancy, due to the fact that the same bad night lowers sleep quality and HRV together. However, it was also found that none of the manufacturers disclosed their exact formulas for composites. The important detail is that with the formula being proprietary, the researchers could not test whether WHOOP's calculation actually penalizes the same disturbance twice, and whether this makes WHOOP's score disproportionately negative (Doherty et al. 2025).

Another factor which may impact the recovery score's accuracy is single night data. Overnight HRV has day-to-day variation of around 10–20% (Buchheit 2014), which is why researchers typically average a week in past studies (Plews et al. 2014). WHOOP, though, scores every night individually against your overall baseline as if each were a separate signal. And even then, HRV doesn't just point one way. When 21 triathletes were deliberately overloaded for three weeks, the overreached group's overnight RMSSD went *up* while their

running performance fell 9% (Le Meur et al. 2013). In turn, a green morning doesn't imply an athlete is ready, reinforcing the idea that the Recovery score is not reliable in determining how wrecked or recovered one is.

There is one caveat I owe the reader, which is that the swimmer study is one small study. 23 athletes and two-week average measurements means that the sample size is very small. Also, it only looked into stress and metabolism, not pure 'athletic readiness.' The golf study adds performance evidence, but only at a seasonal level. The question Whoop's own definition leaves me with is more specific to how the recovery score is constructed, and how effective it is: does Recovery on day N predict performance on day N, within the same person, better than HRV alone? The existing studies that I discussed above do not settle that question.

The loop: grading its own homework

This is the part I think matters most, and the part almost nobody covers.

Recovery is not just displayed. It sets the day's training target. In Whoop's words: "Each morning, WHOOP analyzes overnight sleep data, nocturnal HRV, and resting heart rate to produce a Recovery score, then converts that into a recommended strain range for the day" (Strain Coach). The zone boundaries behind that target "were developed by analyzing nearly a million days of Strain and Recovery data and their impact on next-day HRV" (Whoop training zones). And HRV "has the greatest influence" on the Recovery score (Whoop, Jul 2026).

WHOOP also uses the following day's HRV to calculate the next Recovery score. Each morning, the score goes into setting a recommended daily Strain range, which was developed by looking at how training related to the following day's HRV. That HRV then goes into calculating the next morning's score. The important detail is that HRV helps determine both the training recommendation and the score used to assess recovery after following it. In turn, a higher score can make the advice appear to have worked, without telling us whether the athlete is better prepared to perform.

Is that a flaw? Partly no. A daily Strain target that accounts for the following day's HRV can be useful, due to the fact that HRV may provide information about the athlete's physiological state. The important detail is that there are various other composite factors, which limits what an improvement in that measure can tell us. Furthermore, the recovery score doesn't correlate effectively with actual readiness. In turn, the recommendation may help preserve HRV without establishing whether the athlete is better recovered or ready to perform. That matters because it means a member's experience of "this thing works" isn't evidence that the score is making an athlete more ready. It's evidence that the system moves the variable it also grades. Those feel identical from inside the app.

And the feature works whether or not the score is true. Tell 164 students they had 16% or 29% REM sleep, both numbers invented, and they perform to the number they were given on attention and verbal-fluency tests (Draganich & Erdal 2014). Give 63 people with insomnia sham "bad night" readouts from a wrist tracker and their reported alertness drops (effect size 0.79) while their measured vigilance does not move (0.12) (Gavriloff et al. 2018). Interview 17 people who have worn a Whoop or Oura for three months or more and they describe changing training, sleep and food to move the score, while also saying "you can't really capture the complexities of a human on a device" (Ibrahim et al. 2024). The loop doesn't need the score to be right. It needs it to be believed.

What Whoop is without the score

If we look at WHOOP without its recovery score, the strap becomes a screenless band that reports numbers an Apple Watch can already show. Members don't open the app "over eight times per day" (Whoop, Mar 2026) to view the respiratory-rate decimals. They do it for the verdict, and that is why the score is the product.

And the product itself is proving to be large. Whoop's Series G in March 2026 raised $575 million at a $10.1 billion post-money valuation, led by Collaborative Fund, with Abbott and Mayo Clinic on the cap table. The company also disclosed that they have, actively, "over 2.5 million members," and 2025 bookings up 103% to a "$1.1B run rate." WHOOP has also reported to be "cash flow positive" in 2025 (Whoop press release; TechCrunch). One factor that Whoop has never disclosed, in ten years, is churn. As of today, the only public benchmark is Oura's S-1, which was filed this month. Oura reported $907.9 million of revenue in FY 2025, with around 5.0 million paid members and roughly 85% weighted twelve-month retention (the5krunner, from the S-1). If Whoop's number were better than this, I'd expect to have seen it.

Would a validated score be worth more to Whoop? I think less, for three reasons.

First, strict validation shrinks the score. Most mornings, the gap between today and your baseline is within the noise of a single night. An honest number would say so, and 'no real change' is a hard thing to build a daily verdict on. Nobody pays hundreds for a device to "Tell you only when something's off," a feature which Apple Vitals already does for free. The range is what makes the number feel like a daily verdict, and the daily verdict is what Whoop charges for.opens an app eight times a day for "usual".

The second reason is regulation, which WHOOP has already had issues with. In July 2025, the FDA sent WHOOP a warning letter regarding its Blood Pressure Insights feature (FDA warning letter). WHOOP argued that the feature was only a wellness tool, but the FDA rejected this due to the fact that blood pressure is "inherently associated with the diagnosis of hypo- and hypertension." WHOOP did not pull the feature, and the letter was eventually closed in June 2026. However, this was only after WHOOP changed the dial boundaries so that the feature would not be confused with classifying blood pressure clinically (MedTech Dive). The important detail is that the FDA was looking at how the number was shown to users, and not at the sensor. Recovery is not a clinical measure in the way that blood pressure is, so the two cases are not exactly the same. A validated score that tells athletes when to train, though, would be a much more specific claim than what WHOOP says about Recovery now. In turn, validating the score may bring more of this kind of attention to WHOOP, giving the company another reason to leave the score as it is.

The third reason is that the market has recently voted. Apple spent two years refusing to ship a composite score, where its Vitals app shows five overnight metrics against a seven-night typical range and only notifies you "if multiple metrics are outside of your typical range," with the note that it is "not intended for medical use" (Apple). Then recently, in September 2026, the company shipped a 0–10 Readiness score with four bands, ranging from "Recover" to "Go for it," on the Series 12 and Ultra 4. They have provided, however, the caveat that it "should not be solely relied on to determine whether it is safe to exercise" (Apple). Here, we see that even the most careful company in consumer health decided a number to look at and think about beats being honest about what the number explicitly means.

Regardless, it is evident that the ambiguity behind the score is a strategic choice The tell is what WHOOP has chosen to validate. Its ECG feature is FDA-cleared (K243236). Its Blood Pressure Insights feature was fought

to a draw with the regulator. Its blood testing products Advanced Labs, are clinician-reviewed through Quest (Whoop Advanced Labs). Recovery on a day-to-day basis, however, the thing 2.5 million people wake up to, has effectively not been tested against its own definition. The way that the score is constructed is also not disclosed, keeping the single number as the product in place of raw biomarkers.

The best case for Whoop

There are three points to consider for WHOOP's side, and I accept all of them.

First of all, no real criterion exists, and therefore every other company is also constructing data when they sell a recovery score. Oura, Garmin, Fitbit and now Apple all sell a composite, but none of the companies disclose a formula (Doherty et al. 2025). Therefore, WHOOP is in no way uniquely 'guilty.' Ambiguity is the category, and this is how the business operates. WHOOP simply does it with an efficient HRV sensor and a loud definition for what recovery means.

Second, this memo is not stating that HRV does nothing – even though the effects might not be linked effectively to an athlete's recovery status. In three different analyses of HRV-guided versus pre-planned training, each one with around 200 independent athletes, there was no significant gain in performance found. The studies reported sizes of 0.08 (Düking et al. 2021), 0.20 (Manresa-Rocamora et al. 2021) and 0.14 (Ramos-Campo et al. 2020). However, what these reviews did find is that HRV-guided training produced fewer non-responders, meaning fewer people completed the program and saw no improvement at all, and that individuals experienced "fewer high-intensity sessions".

Third, the behaviour change is of real value. Whoop's own analysis links each drink to a 4.2% lower next-day Recovery (Whoop), and its 2026 studies, company-authored, report a 25.2% relative decline in the daily probability of drinking across 30,000 new members over 72 weeks (Whoop, Apr 2026). If a constructed number gets people to drink less and sleep more, the construction did something true, whatever the number measured.

These make the score a great behavioural product that fits the consumer wearables industry. The problem is that none of them makes it a measurement, and the company sells it explicitly as one.

The position, and my two calls you can check

Position. Whoop constructs recovery. The score is a composite with undisclosed weights, built on a clean signal, diluted with a poor one, scored one noisy night at a time, and tuned to its own main input. Its commercial value comes from being believed rather than from being right, which is why the company has validated everything around it and never the day-to-day score itself. That makes Whoop a very good behavioural product and a $10 billion one. It does not make the number a measurement.

Call 1, science. By 21 September 2027, no independent, peer-reviewed study will show that Whoop's Recovery score predicts next-day performance better than its own raw HRV input does, within the same person. If one does, I am wrong that the composite destroys useful information, and that section of this memo fails. Note the bar, that a significant correlation with performance is not enough, because HRV alone already has one. The score has to beat its ingredient.

Call 2, business. By the same date, Whoop will not have published Recovery's weights or a validation of Recovery against performance, while continuing to add validated clinical features around it. If Whoop validates the score, I am wrong about what the score is for, and I would expect that validation to come with a smaller, flatter number.

What I'd ask Whoop. Publish the day-N Recovery versus day-N performance test. In ten years, and with a million days of paired data and 2.5 million active members, your company has never published a peer-reviewed test of whether this morning's score predicts this day's performance. The closest you came was the 2016 basketball case study, never peer-reviewed, which you have since removed from its site. If the score works objectively, it is the cheapest marketing you will ever run. If it doesn't, you already know, and this memo is only saying it out loud.

The wider point: In the chain I'm tracking, from sense, infer, to act, Whoop has demonstrated that the "act" layer can be as thin as a simple colour, a number, and a piece of advice. And even within all this, it can still carry a $10 billion business, as long as the inference underneath is believed and trusted. The next question is what happens when the act gets real. The advice is already there, but when the number stops being a colour and actually provides strict advice, or mirrors the language of a clinician, that is where being right will start to matter, and that is where I'll look next.

Sources

Every figure above links to its page inline. Flags: (W) Whoop-authored, funded or from a Whoop-supported lab; (I) independent; (S) secondary report, check the primary before quoting.

Whoop's own pages (W)

1. Whoop 101, developer docs — inputs and score bands
2. Recovery, API docs — RMSSD field
3. Recovery explainer, 15 Jul 2026 — alternative input list, HRV "greatest influence," Strain Coach
4. Recovery 101 — definition
5. Whoop podcast #84 — Capodilupo on the design target
6. Strain Coach — Recovery to strain range
7. Training zones — "nearly a million days… next-day HRV"
8. Research FAQ (PDF, via U. Arizona) — HRV window
9. Series G press release, 31 Mar 2026 — valuation, members, bookings, app opens
10. Membership pricing
11. Advanced Labs, 30 Sep 2025
12. Alcohol and biometric data — per-drink Recovery effect
13. Alcohol studies press release, 29 Apr 2026
14. NBA performance coach on Whoop

Independent research (I)

15. Dial et al. 2025, Physiological Reports — Whoop 4.0 and Oura vs ECG, 536 nights
16. Schyvens et al. 2025, SLEEP Advances — sleep staging vs PSG, 62 adults
17. Lundstrom et al. 2023, Int J Sports Sci Coaching — 23 NCAA swimmers, no funding
18. Ungaro et al. 2026, Sensors — within-person HRV vs wellbeing; PepsiCo-funded
19. Doherty et al. 2025, Translational Exercise Biomedicine — 14 composite scores reviewed
20. Kellmann et al. 2018, IJSPP consensus statement
21. Saw, Main & Gastin 2016, BJSM — 56-study review
22. Buchheit 2014, Frontiers in Physiology — HRV day-to-day variation
23. Plews et al. 2014, IJSPP — three days vs seven
24. Le Meur et al. 2013, MSSE — overreached triathletes
25. Draganich & Erdal 2014, J Exp Psychol — placebo sleep
26. Gavriloff et al. 2018, J Sleep Res — sham feedback
27. Ibrahim et al. 2024, Appl Psychophysiol Biofeedback — 17 wearable users
28. Düking et al. 2021, J Sci Med Sport; Manresa-Rocamora et al. 2021; Ramos-Campo et al. 2020 — HRV-guided training meta-analyses
29. Taylor et al. 2012, J Aust Strength Cond — practitioner survey
30. Neupert et al. 2022, J Sports Sci — UK monitoring practice
31. Akenhead & Nassis 2016, IJSPP — 41 football clubs
32. Rabbani et al. 2019, Frontiers in Physiology — Hooper index vs HRV
33. Iizuka et al. 2020, Sports — Japan national badminton team
34. Schneider et al. 2020, PLOS ONE; Barth et al. 2019, Frontiers in Physiology — German national badminton monitoring
35. McCall et al. 2023, Sports Medicine — athletes gaming wellness forms

Companies, regulators, reporting

36. FDA warning letter to Whoop, 14 Jul 2025
37. MedTech Dive, Jun 2026 — warning letter closed (S)
38. FDA 510(k) K243236 — Whoop ECG clearance (S)
39. TechCrunch, 31 Mar 2026 — Series G (S)
40. the5krunner, 4 Sep 2026 — Oura S-1 figures (S; verify against the SEC filing)
41. Apple Vitals support page; Apple Readiness support page
42. Polar Nightly Recharge
43. NFLPA–Whoop, 2017; Notre Dame Athletics, 2024