VITALITY TRACKER

Sleep Scoring Methodology

Open Specification · Version 2.0.1

Post-review release · Public edition

Every weight. Every formula. Every limitation. Every correction.

What changed, and why it matters

Version 1.1.0 of this specification was sent for external review by an independent data scientist. The reviewer identified a self-contradiction, a measurement-chain error, and a component whose inclusion violated our own first design principle.

We accepted almost all of it. Version 2.0.0 removes sleep-stage data from the score entirely, replaces the hard regularity gate with a continuous function, introduces measurement provenance so that a user is not scored differently because of which device they bought, and corrects two overreaching causal claims.

The full audit trail — every criticism, our response, and the three points we pushed back on — is in the Changelog overleaf. We publish the criticism as well as the method. A specification that has never been shown to be wrong in public has not been tested.

Novak Consulting Ltd · Auckland, New Zealand · 13 July 2026
Public edition v2.0.1, 8 October 2026. This edition specifies the Sleep Health Index and the nightly score; the questionnaires available in the app are outside its scope.

DOI 10.5281/zenodo.23223863 · https://doi.org/10.5281/zenodo.23223863
Cite as: Novak C. Vitality Tracker Sleep Scoring Methodology, Open Specification v2.0.1, public edition. Novak Consulting Ltd, 8 October 2026. doi:10.5281/zenodo.23223863
Text © Novak Consulting Ltd, licensed under CC BY 4.0: reuse with attribution.

Changelog

v2.0.1, public edition, 8 October 2026 — editorial only, no change to any output: US spelling; the external reviewer described as an independent data scientist; the deep-sleep night-to-night variability figure corrected to its source (§4.1); the device bias table re-derived from Schyvens et al. (2025) after a spot-check found one column misaligned, with the two Fitbit devices added and the Oura Ring 4 and Ring 5 note (§7.4); the shift-work, Ohayon (2004) and first-generation Oura statements aligned to their sources (§4.1, §5.3, §6.2, §9.1); the data-quality exclusion rule stated under Coverage (§8). The validated questionnaires offered in the app, and what the app shows around them, are outside the scope of this specification.

v2.0.0, 13 July 2026 — the post-review release. Removing a scoring component is a MAJOR version bump under the policy in §9. Historical scores are recomputed under v2.0.0 and both values retained. Every change below is attributed to the review point that prompted it.

Accepted — changes made

Change	Review finding	Why we accepted
Sleep-stage share removed from the score entirely. N3+REM now displayed as descriptive, low-confidence information at 0% weight.	Assigning 10% "because users expect stage data to count" is an availability-and-market rationale, which our own Principle 1 explicitly forbids. Cross-device agreement could not rescue it: two devices can agree and both be wrong, and agreement establishes neither outcome relevance nor a valid target.	The objection is correct and we should have caught it. The compromise weight was a commercial hedge dressed as a scientific position. Removing it also produces a stronger claim than the hedge did: **we are the only platform that refuses to score sleep stages, and we explain why.**
Weights rescaled to Regularity 35 / Duration 30 / Continuity 15 / Circadian 10 / Restoration 10.	Consequential to the removal above.	Preserves the relative ordering of the retained components.
Hard regularity gate replaced with a continuous multiplier G(SRI), applied only to persistent patterns across two consecutive windows.	A cap of 50 at SRI 71.65 converts an observational quintile boundary from one cohort into a deterministic product rule. SRI 71.6 is not categorically different from SRI 71.7. There is no biological discontinuity there.	Correct. The quintile boundary is an artifact of how the cohort was divided, not a threshold in nature. The soft multiplier preserves the entire function of the gate while removing the false cliff (see §5.3 and the worked example in §10.2).
Measurement provenance and source-class calibration introduced (§7). Continuity is no longer scored against PSG norms using uncalibrated wearable values.	The age-banded continuity anchors derive from level-1 PSG, but the inputs are wearable estimates, which systematically over-estimate sleep efficiency and under-estimate WASO, and differ by manufacturer and firmware. A user could score differently because of their device rather than their sleep.	**The single most important finding in the review.** The error compounds with age-banding: a device reporting a mean WASO of 4.0 min against a PSG mean of 42.4 would score 100 against a 64-minute age anchor. Continuity was becoming free points, and the size of the gift depended on which device the user bought.
Missing-data redistribution abolished. Partial scores are now reported with an explicit coverage percentage (§8).	Redistributing Restoration's weight implies the objective metrics can substitute for it — which contradicts the stated reason for including it, namely that subjective sleep is a non-redundant dimension. A score of 88 with Restoration is not the same construct as 88 without it.	Correct, and it was an unforced error: we justified the component on non-redundancy grounds and then treated it as redundant when absent.
CPAP / OSA causal language rewritten (§2.2).	The inference from three null CPAP trials to "OSA is a marker, not a cause" is not licensed. A null treatment trial admits several explanations: insufficient adherence, late intervention, a partially-corrected pathway, a selected population, or an insensitive endpoint.	Accepted (with one qualification we retained — see below).
"No RCT of any sleep intervention has used cardiovascular events or mortality as a primary endpoint" — corrected.	Flatly contradicted by our own discussion of SAVE, which was explicitly a cardiovascular-endpoint trial.	Indefensible, and embarrassing. The claim we meant — that no *behavioral* sleep-duration, timing or regularity intervention has been tested against hard outcomes — is both true and more interesting. Corrected throughout.
"The only sleep variable with Mendelian-randomization support" — corrected to "the only SHI component with MR support."	Our own causal-evidence table lists MR findings for insomnia symptoms, chronotype and daytime napping.	Correct. Sloppy drafting.

Change	Review finding	Why we accepted
Principle 1 restated as "weight by evidence AND effect size, both disclosed" (§1).	Raised internally during review response.	Regularity is Grade B and carries the largest weight; Duration is Grade A — the only component with causal support — and carries less. Grade B outranking Grade A is defensible on effect size and modifiability, but the principle as previously worded did not license it. Now stated openly rather than left for a hostile reader to find.
Age-banding claims moderated (§6.2, §10).	"A 57-year-old with 64 minutes of WASO is not a deteriorating sleeper — he is a normal one" overreaches. A cross-sectional pooled mean cannot establish that an individual is healthy or free of pathology. And the 80/100 anchor mapping is a design convention, not an empirical result.	Both correct. The 80/100 mapping is now explicitly labeled as an unvalidated value judgment.
Timing reformulated as circadian alignment with an explicit evidence hierarchy (§5.5).	Users who have not completed a chronotype questionnaire were receiving an absolute clock-time judgment, which may be less valid than no judgment at all.	Accepted. Alignment, not clock hour, is the construct our own science review endorses.

Contested — where we pushed back

Review recommendation	Our position	Basis
"Exempt or separately classify rotating-shift schedules" from the regularity penalty.	**Rejected.**	Irregularity that is involuntary is still irregularity — the biology does not care why the schedule moved. Shift work carries MI RR 1.23 ($I^2 = 0$) across ~2 million people (Vyas et al., BMJ 2012) and IARC classifies night shift work Group 2A. Exempting shift workers would tell the highest-risk group in our user base that their principal risk factor does not apply to them. We **contextualize rather than exempt**: the state is named "schedule instability," acknowledged as often non-negotiable, and the guidance pivots to the mitigations the user can control (§5.3).
Score social jetlag as part of the circadian component.	**Rejected as a scored input; accepted as descriptive.**	We accept the reviewer's circadian hierarchy in full, but not this specific metric. Our own literature review found no epidemiological evidence linking social jetlag to any hard outcome — the literature is cross-sectional, small-n, surrogate-endpoint-only, and confounded by shift work, socioeconomic status and depression. Importing it would smuggle a Grade C/D exposure into the component we had just finished cleaning up.
Moderate the categorical prose throughout.	**Partially accepted.**	The three specific overreaches identified (long sleep, the 60-year-old deep-sleep example, "strongest single sleep variable") are conceded and corrected. We decline a blanket hedging pass: the value of this document lies in its willingness to reject weak claims plainly. Softening every statement to "may be associated with" would produce exactly the unfalsifiable prose the category already suffers from. We are aiming for precision, not timidity.
On CPAP: the trials say nothing about whether OSA is causal.	**Accepted, with one qualification retained.**	The reviewer is right that a null treatment trial cannot establish that an exposure is merely a marker. But the ISAACC evidence we relied on was not the treatment null — it was the *observational* comparison within the trial: untreated OSA versus no OSA, HR 1.01 (0.76–1.35). The exposed group genuinely did not experience more events. That is more probative than "null trial, therefore no causation." We retain it as evidence, while accepting that it cannot be generalized beyond non-sleepy post-acute-coronary-syndrome patients.

Deferred — acknowledged, scheduled

The reviewer noted that several operational rules are not yet fully reconstructable from the document, and is right that this falls short of the promise on our cover. The following are now tracked as a completeness backlog for v2.1.0: exact Restoration conversion and KSS/Epworth blending; interpolation continuity at all curve boundaries; rounding policy; timezone and daylight-saving handling; naps and multiple sleep episodes; cross-midnight and travel; minimum valid epochs per night; duplicate and overlapping device-record resolution; source-priority rules; night-attribution (sleep-start date versus wake date); and recomputation

policy when an upstream device algorithm changes without a discoverable firmware version. A published reference implementation will accompany v2.1.0, so that the specification and the shipping code cannot drift apart.

1. Design Principles

Seven rules govern the score. Where a rule and a commercial instinct conflict, the rule wins, and the conflict is disclosed.

#	Principle	What it means in practice
1	**Weight by evidence AND effect size — both disclosed**	A component earns weight from two things: the strength of the evidence linking it to health outcomes, and the size of the association. Neither alone is sufficient, and the two can point in different directions. **Regularity (Grade B) outweighs Duration (Grade A) in this score.** That is a deliberate choice: duration has better causal support, regularity has the larger observed effect and is more modifiable. We state this openly rather than let the tension sit unexamined. What we do NOT permit is weight earned by availability, visual appeal, or user expectation.
2	**The habitual pattern is the signal; the single night is mostly noise**	Almost all outcome evidence in sleep science concerns habitual patterns over weeks. The flagship number is a rolling 28-night index. The nightly score exists, is labeled low-confidence, and is not the score we ask users to optimize.
3	**Never score what you cannot measure**	If the instrument in the user's hand cannot measure a quantity to a useful tolerance, that quantity does not enter the score — however much the user expects it to. This is why sleep stages are displayed but not scored (§4).
4	**Score against the right reference system**	A measurement is only comparable to a norm derived by the same method. Wearable estimates are not polysomnography, and must be calibrated or scored against a personal baseline instead (§7).
5	**Age-band only where the evidence shows an age effect**	Continuity changes with age and is age-banded. Sleep-stage proportions do not change with age. Applying an age adjustment where none is warranted would let the score explain away a genuine problem.
6	**Never rescore history silently**	Algorithm changes are versioned, published with a changelog and rationale, and historical scores are recomputed with both values retained.
7	**Publish the criticism, not just the method**	This specification carries its own external review, including the points we lost. A method that has never been shown to be wrong in public has not been tested.

1.1 Evidence grading scheme

Grade	Definition	Max weight
A	Causal support: randomized controlled trial or Mendelian randomization, against a hard outcome.	Unrestricted
B	Strong observational: large prospective cohorts, consistent direction, survives adjustment. No causal test exists.	Unrestricted
C	Weak or contested: small cohorts, inconsistent replication, or a serious reverse-causation threat.	≤ 15%
D	Descriptive only: a normative population statistic with no demonstrated link to any outcome.	**0% — display only**
E	Expert consensus: an appropriateness vote, not derived from data.	≤ 20%

Changed in v2.0.0. Grade D previously carried a 10% ceiling. It is now zero. A population statistic with no demonstrated link to any outcome is information, not a target, and it cannot earn points. This single change removes sleep-stage data from the score.

2. What the Evidence Does and Does Not Support

Two claims in v1.1.0 outran the evidence. Both are corrected here, and both corrections are worth stating plainly because they cut against the direction our own product would prefer.

2.1 On sleep interventions and hard outcomes

Corrected claim

PREVIOUSLY (v1.1.0): "No randomized controlled trial of any sleep intervention has ever used cardiovascular events or mortality as a primary endpoint. Not one."

This was false, and was contradicted by our own analysis. SAVE (n = 2,717), ISAACC (n = 1,264) and RICCADSA (n = 244) were all randomized trials of CPAP with cardiovascular primary endpoints.

CORRECTED (v2.0.0): "No randomized trial of a BEHAVIORAL sleep intervention — extending sleep duration, shifting sleep timing, or improving sleep regularity — has ever demonstrated a reduction in cardiovascular events or mortality. The randomized evidence that does exist for hard outcomes concerns CPAP for obstructive sleep apnea, and it is null."

This matters for what we may claim. Every behavioral recommendation this product makes — sleep longer, sleep more regularly, sleep at a consistent hour — rests on observational and Mendelian-randomization evidence, not on trial evidence that acting on it improves hard outcomes. We say so.

2.2 On obstructive sleep apnea

Corrected claim

PREVIOUSLY (v1.1.0): "This directly supports the view that OSA in this population is a marker of visceral adiposity and metabolic syndrome rather than an independent cause."

CORRECTED (v2.0.0): "These trials do not support prescribing CPAP to non-sleepy patients with established cardiovascular disease for the primary purpose of reducing cardiovascular events. They do not resolve whether untreated OSA contributes causally to cardiovascular disease earlier in its natural history, or in more adherent and symptomatic populations. A null treatment trial admits several explanations besides absence of causation: insufficient adherence (mean 3.3 h/night in SAVE), intervention begun too late, a causal pathway only partially corrected, a selected population, or an endpoint insensitive to the relevant benefit."

One qualification we retain. The evidence we leaned on from ISAACC was not the treatment null but the observational comparison inside the trial: untreated OSA versus no OSA gave HR 1.01 (0.76–1.35), p = 0.93. The exposed group did not experience more events. That is a stronger form of evidence than "the treatment failed, therefore the exposure is innocent" — but it is confined to non-sleepy post-acute-coronary-syndrome patients and cannot be generalized. We hold it at that scope.

Separately, and importantly for the product: CPAP demonstrably improves snoring, daytime sleepiness, mood and quality of life. That is why the app offers the questionnaires a clinician uses to decide whether a sleep study is warranted, for you to complete and take to your doctor, and that offer does not depend on the cardiovascular question being resolved.

3. Architecture

	Sleep Health Index (SHI)	Nightly Sleep Score
Role	**The flagship.** The number that represents sleep health and the number we ask users to improve.	A descriptive daily read-out. Present because users expect one. Not the target.
Window	Rolling 28 nights (minimum 14 valid nights).	Last night only.
Confidence	Moderate, and reported as a coverage percentage (§8).	**Low, and labeled as such.** Most sleep metrics are dominated by night-to-night variance at this resolution.
Regularity	Centrepiece — 35% weight plus a continuous multiplier (§5.3).	Only as schedule adherence against the user's own 28-night baseline. A true SRI cannot be computed from one night.
Sleep stages	**Not scored.** Displayed as descriptive information.	**Not scored.** Displayed as descriptive information.

The product hierarchy a user sees, in order:

- **1. What matters over time** — the Sleep Health Index.
- **2. What changed last night** — the nightly score, labeled low-confidence.
- **3. What we measured confidently** — coverage, source device, measurement confidence.
- **4. What we cannot measure confidently** — estimated sleep stages, explicitly excluded from scoring.
- **5. What you can take to your doctor — the validated questionnaires, outside the score.**

3.1 SHI banding

Band	Range	Interpretation
Optimal	85–100	Consistent with the best-supported evidence across all dimensions. **Requires coverage ≥ 90% (§8).**
Good	70–84	Broadly healthy. One or two dimensions have room to improve.
Fair	55–69	A meaningful dimension is underperforming. Actionable.
Needs attention	40–54	Multiple dimensions underperforming.
Red	< 40	Consider a conversation with your doctor.

4. Why We Do Not Score Sleep Stages

This is the most visible difference between Vitality Tracker and every competing product, and it is the change we are most confident about.

4.1 The reasoning

- **There is no target to hit.** No randomized trial has ever assigned anyone to a sleep-stage percentage and measured a health outcome. The "20–25% deep sleep" figure that saturates the category is a population mean (N3 20.4%, REM 19.0%; Boulos et al. 2019) that has been reified into a goal. And more is not better: Lucey et al. (Brain 2021) found cognitive decline at BOTH low and high slow-wave activity — an inverted U that is fatal to any maximization framing.
- **The instrument cannot measure it.** Consumer devices classify four stages at roughly 50–65% epoch-by-epoch accuracy ($\kappa \approx 0.2–0.5$). Per-stage errors run 30–60 minutes a night, with limits of agreement that reach ±150–200 minutes on some devices. Independent clinical replication puts Oura's four-stage accuracy at 53%, against 92% in the manufacturer-adjacent study.
- **Even a perfect instrument would not help at nightly resolution.** In 94 healthy men recorded over 470 home nights with laboratory-grade EEG, the night-to-night coefficient of variation of the deep-sleep share was 0.22, and of sleep latencies and wake time 0.44–0.58 (Chouraki et al., Sleep 2023). A single night's figure carries limited information even when measured flawlessly, and consumer devices do not measure it flawlessly.
- **Merging N3 and REM does not fix it.** We considered this seriously and tested it against the literature. It only works if devices confuse deep sleep specifically with REM, so the errors cancel. The data show this happens for some devices and not others — it cancels for the first-generation Oura ring in adolescents and young adults (−20 min N3, +17 min REM; de Zambotti et al., 2019) but compounds for WHOOP (+31.5, +15.3) and Apple (−25.2, −13.4). Worse, the dominant confusion in the literature is light ↔ deep, not deep ↔ REM, which merging does nothing to correct.
- **Cross-device agreement could not have rescued it.** Our v1.1.0 plan was to validate the merged block against our own multi-device cohort. The reviewer pointed out the flaw: two devices can agree and both be wrong. Agreement demonstrates neither outcome relevance nor a valid target. We abandoned the plan rather than run a study that could not answer the question.

> **What the user sees instead**
>
> Estimated sleep-stage distribution — descriptive, low confidence, excluded from your Sleep Health Index.
>
> The hypnogram remains in the app. The data remain exportable. What changes is that the number does not move the score, and we tell the user exactly why. A sustained extreme deviation produces a neutral informational note — not a points deduction, and not an implication of pathology from a single consumer algorithm's stage estimate.

We are aware this will read, to some users, as a missing feature. That is a cost we accept. The alternative is to assign points to a quantity we have said in print cannot be measured or targeted — and a specification that contradicts itself in public is worth less than a missing dial.

5. Components, Weights and Curves

```
SHI_raw = 0.35·Regularity + 0.30·Duration + 0.15·Continuity
        + 0.10·Circadian + 0.10·Restoration

SHI     = SHI_raw × G(SRI)          # continuous multiplier, §5.3
```

Component	Weight	Grade	Effect size / basis
Regularity (SRI)	**35%**	B	Largest observed effect among objectively measured sleep variables. Most regular quintile vs least: all-cause mortality HR 0.70 (0.59–0.83), fully adjusted (Windred et al., Sleep 2024; UK Biobank, n = 60,977). Adding duration to a regularity model did not improve fit (LR p = 0.14–0.20). Also the most behaviorally modifiable variable in the score.
Duration	**30%**	A	The only SHI component with Mendelian-randomization support for a hard outcome: short sleep → myocardial infarction, HR 1.19 (1.09–1.29) (Daghlas et al., JACC 2019). Better evidence than regularity; smaller effect. Hence second, not first — see Principle 1.
Continuity	**15%**	E	NSF consensus targets (Ohayon 2017), age-banded to PSG norms (Boulos 2019) and **source-calibrated** (§7). Expert appropriateness votes, not outcome-derived.
Circadian alignment	**10%**	C	Alignment between biology and schedule. Sleep onset 22:00–23:00 associated with lowest CVD incidence, but the curve is U-shaped and reverse causation is unresolved. Low weight for this reason.
Restoration	**10%**	B	Subjective satisfaction and daytime alertness. Correlates poorly with objective sleep (Grandner 2006) — which is exactly why it is not redundant. Never substituted for by objective metrics (§8).
Sleep stages	**0%**	D	**Display only.** See §4.
	100%		

5.1 Regularity — 35%, Grade B

SRI computed per Phillips et al. (2017) over the trailing 28 nights at 1-minute epoch resolution: the probability of being in the same state (asleep/awake) at any two timepoints 24 hours apart, scaled 0–100.

SRI	≤ 40	55	71.65	78.56	83.16	87.31	≥ 92
Score	0	25	50	70	85	95	100

Anchored to UK Biobank quintile boundaries (Q1 < 71.65; Q2 71.65–78.56; Q3 78.56–83.16; Q4 83.16–87.31; Q5 > 87.31). Median 81.0, IQR 73.8–86.3, negatively skewed.

Language moderated in v2.0.0. We previously called regularity "the strongest single sleep variable" and "the one thing that matters most." The defensible claim is narrower: regularity was the strongest predictor among the objectively derived sleep measures compared in the largest cohort to date. That is still remarkable, and it is much harder to attack.

5.2 Duration — 30%, Grade A, age-banded

Age band	Full score (100)	Long-sleep note
18–25	7.0 – 9.0 h	> 11 h
26–64	7.0 – 9.0 h	> 10 h

Age band	Full score (100)	Long-sleep note
65+	7.0 – 8.0 h	> 9 h

Short arm: 6.5–7.0 h → 70–100 · 6.0–6.5 h → 40–70 · 5.0–6.0 h → 10–40 · < 5.0 h → 0–10.

Long arm: taper to 90 at +1 h beyond the band, then floor at 85. Not penalized further.

Why long sleep is not punished. The causal status of long sleep remains unestablished, and reverse causation probably explains a substantial proportion of the observational association — cancer, heart failure, depression and frailty all cause long sleep. Mendelian randomization corroborates short sleep as causal for myocardial infarction; it does not carry the long-sleep arm across. Deducting points would risk penalizing a user for a symptom. The app shows a neutral note suggesting a conversation with your doctor instead.

5.3 The regularity multiplier — replacing the hard gate

Version 1.1.0 capped the SHI at 50 for any user below SRI 71.65. The reviewer was right that this converts an observational quintile boundary from a single cohort into a deterministic product rule, and that SRI 71.6 is not categorically different from SRI 71.7. The cliff has been replaced with a continuous function.

```
G(SRI):   SRI ≤ 60    → 0.70
          SRI = 71.65 → 0.80
          SRI = 78.56 → 0.90
          SRI = 83.16 → 0.97
          SRI ≥ 85    → 1.00        # no penalty

# Persistence rule — the penalty is not applied at full
# strength on a single window:
if not low_SRI_in_two_consecutive_windows:
    G = 1 − (1 − G) / 2
```

The multiplier requires at least 28 valid nights, and reaches full strength only when low regularity persists across two consecutive calculation windows. A single disrupted month — a holiday, a newborn, a bereavement — is treated as an episode, not a verdict.

Where we disagreed with external review: shift workers are not exempt

The reviewer suggested exempting or separately classifying rotating-shift schedules. We decline, and the disagreement is substantive rather than defensive.

Irregularity that is involuntary is still irregularity. The biology does not care why the schedule moved. Shift work carries a myocardial-infarction relative risk of 1.23 with $I^2 = 0$ across two million people, with every schedule except evening shifts raising coronary risk (Vyas et al., BMJ 2012), and IARC classifies night shift work as Group 2A, probably carcinogenic. Exempting shift workers from the regularity penalty would tell the highest-risk group in our user base that their principal risk factor does not apply to them.

What we do instead is contextualize. The state is named "schedule instability" rather than implying globally poor sleep health; it is acknowledged as frequently non-negotiable; and the guidance pivots to what the user can control — anchor sleep periods, light exposure timing, and consistency of the non-work days. Naming a real risk honestly, while helping with what is actionable, is more respectful than pretending the risk is absent.

5.4 Continuity — 15%, Grade E, age-banded and source-calibrated

Three equally weighted sub-components (5% each of the SHI): sleep efficiency, sleep onset latency, and wake after sleep onset. The age-banded PSG anchors are set out in §6; the source-calibration rules that determine whether those anchors may be used at all are in §7.

5.5 Circadian alignment — 10%, Grade C

Reformulated in v2.0.0. Previously this component judged users against an absolute clock window (mid-sleep 02:00–04:00). The reviewer correctly observed that a user who has not declared a chronotype was receiving an absolute clock-time judgment that may be less valid than no judgment at all. Scoring now follows a strict evidence hierarchy:

Tier	Available data	What is scored	Confidence
1	MCTQ completed, or sufficient behavioral history to estimate chronotype.	**Alignment with the user's own corrected mid-sleep on free days (MSF_sc).** Deviation between habitual sleep timing and estimated circadian phase.	Highest available
2	No chronotype, but ≥ 28 nights of history.	Internal consistency with the user's own habitual sleep timing. Are they sleeping when they normally sleep?	Moderate
3	Neither.	**Component suppressed.** The absolute clock-time association is shown as weak contextual information only, and is not scored.	Not scored

Social jetlag is displayed but not scored. We accept the reviewer's hierarchy but not this specific metric. Our literature review found no epidemiological evidence linking social jetlag to any hard outcome — the evidence base is cross-sectional, small-n, surrogate-endpoint-only, and confounded by shift work, socioeconomic status and depression. Scoring it would import a Grade C/D exposure into the component we had just finished cleaning up. It appears as descriptive information.

5.6 Restoration — 10%, Grade B

Input	Weight within component	Instrument
Morning sleep satisfaction	60%	Consensus Sleep Diary item, 1–5, prompted on wake.
Daytime alertness	40%	Karolinska Sleepiness Scale (state, mid-afternoon) blended with Epworth (trait, quarterly).

If the user responds on fewer than 4 of the 14 most recent days, Restoration is NOT redistributed — it is reported as missing, and the SHI is reported as a partial score with a coverage percentage (§8). This is a change from v1.1.0 and is explained there.

6. Age-Banded Normative References

All age-banded anchors derive from Boulos et al., Lancet Respir Med 2019;7(6):533–543 — 169 studies, 5,273 healthy adults, level-1 in-lab polysomnography, AASM scoring.

6.1 The reference data

Age band	TST	Efficiency	WASO	SOL	Arousal index	N3+REM
18–34	411 min	89.0%	32.1 min	14.3 min	9.6 /h	41.2%
35–49	387 min	85.4%	51.1 min	14.4 min	12.5 /h	39.7%
50–64	372 min	83.2%	64.0 min	15.7 min	16.5 /h	36.8%
65–79	346 min	77.5%	77.1 min	19.5 min	18.8 /h	37.6%
Per-decade change	**−10.1 min**	**−2.1 pts**	**+9.7 min**	**+1.1 min**	**+2.1 /h**	**not significant**

The final row is the meta-regression result. N3+REM shows no significant age effect — which is one reason sleep stages are neither age-banded nor scored (§4).

6.2 Dual-reference scoring — and its status as a design convention

Age-banding creates a hazard: if the age-band mean scores 100, the score is telling an older user that age-typical sleep is optimal sleep. It may not be. Ohayon et al. (2004) found that the age effects on sleep parameters shrank, and in some cases disappeared, once studies screened participants rigorously for mental, organic and sleep disorders — meaning some of the "normal" age-related decline in the Boulos norms may be undiagnosed pathology in inadequately screened samples.

We therefore anchor each age-banded curve at two points: the age-band mean scores 80, and the young-adult reference scores 100. Both are always visible in the breakdown.

Disclosure required by the reviewer

The 80/100 mapping is a design convention, not an empirically derived anchor. It is intended to distinguish age-typical continuity from aspirational continuity. It has not been validated against health outcomes, and no such validation exists for any sleep score.

Relatedly, we have removed the following sentence from v1.1.0: "A 57-year-old with 64 minutes of WASO is not a deteriorating sleeper — he is a normal one." A cross-sectional pooled mean cannot establish that an individual is healthy, is not deteriorating, or is free of sleep pathology. The correct statement is: "This value is typical of the published PSG distribution for adults aged 50–64. Typical does not necessarily mean optimal, and does not exclude an individual sleep problem."

6.3 WASO anchor points by age band (PSG reference system)

Age band	100	90	80 (band mean)	55	25	0
18–34	≤ 16 min	24 min	**32.1 min**	48 min	64 min	96 min
35–49	≤ 20 min	38 min	**51.1 min**	77 min	102 min	153 min
50–64	≤ 20 min	48 min	**64.0 min**	96 min	128 min	192 min
65–79	≤ 20 min	58 min	**77.1 min**	116 min	154 min	231 min

General form: 100 at min(20 min, 0.5 × mean) · 90 at 0.75 × mean · 80 at the mean · 55 at 1.5 × mean · 25 at 2 × mean · 0 at 3 × mean.

6.4 Sleep efficiency anchor points by age band (PSG reference system)

Age band	0	25	55	80 (band mean)	100
18–34	69.0%	77.0%	83.0%	**89.0%**	94.0%
35–49	65.4%	73.4%	79.4%	**85.4%**	90.4%
50–64	63.2%	71.2%	77.2%	**83.2%**	88.2%
65–79	57.5%	65.5%	71.5%	**77.5%**	82.5%

General form: 80 at the band mean · 100 at mean + 5 points · 55 at mean − 6 · 25 at mean − 12 · 0 at mean − 20.

Critical: these anchors are in the PSG reference system. They may not be applied to raw consumer-wearable values. See §7.

7. Measurement Provenance and Source Calibration

The error this section fixes

Version 1.1.0 scored continuity by comparing wearable estimates directly against polysomnography-derived age-band norms. Those two things are not the same measurement.

Consumer wearables systematically over-estimate sleep efficiency (by 2–10 percentage points) and under-estimate wake after sleep onset (by 12–48 minutes), and the size of the error differs by manufacturer and by firmware version. One tested device reported a mean WASO of 4.0 minutes against a polysomnographic mean of 42.4.

Under v1.1.0, that device's user would have scored 100 against a 64-minute age anchor. Continuity was becoming free points, and the size of the gift depended on which device the user had bought — not on how they had slept. This was the most serious defect external review identified, and it outranks any individual threshold.

7.1 Every input carries provenance

No value enters the score without a provenance record. This is a schema requirement, not a nicety.

Field	Description
metric	The quantity (TST, WASO, SE, SOL, sleep onset, wake time, …).
value	The measurement.
source_device	Manufacturer and model (e.g. Oura Ring Gen 3).
source_class	One of: PSG · research actigraphy · consumer wearable · sleep diary · manual entry.
algorithm_version	Firmware or algorithm version where discoverable. Recorded as UNKNOWN where not — and its absence is itself surfaced, because it limits what we may claim.
reference_system	The norm set against which this value may legitimately be scored.
measurement_confidence	High · Moderate · Low, derived from the calibration tier below.

7.2 The three calibration tiers

Tier	Condition	How continuity is scored	Confidence
1 — Calibrated	The device has a published PSG validation with usable bias estimates for the relevant metric.	Apply the published bias offset, then score against the age-banded PSG anchors in §6.	High
2 — Uncalibrated wearable	A consumer device with no published validation, or an unknown algorithm version.	**Continuity is scored as a within-device personal trend** against the user's own 90-day baseline — not against PSG anchors. The PSG norms are displayed as context only, clearly marked as not directly comparable.	Low
3 — Diary	Consensus Sleep Diary entries.	Scored against diary-appropriate anchors. Note that for sleep efficiency and onset latency the diary is arguably the better instrument, since wearable SE is inflated and wearable SOL is unreliable.	Moderate

7.3 The statement that appears in the product

> **Displayed wherever continuity is scored from an uncalibrated source**
>
> PSG-derived normative anchors are not directly equivalent to consumer-wearable estimates. Until source-specific calibration is available for your device, your continuity score reflects change relative to your own baseline, and should not be compared with other users or across devices.

7.4 Known bias estimates (v2.0.1 calibration table)

Device	TST bias	WASO bias	SE bias	Tier
Apple Watch Series 8	+19.6 min	−21.2 min	+4.4 pts	1 (partial)
WHOOP 4.0	+24.5 min	−19.2 min	+4.1 pts	1 (partial)
Garmin Vivosmart 4	+38.4 min	**−38.3 min (reported mean 4.0 vs PSG 42.4)**	+8.2 pts	1 (partial)
Withings ScanWatch	+39.9 min	−47.9 min	+10.2 pts	1 (partial)
Fitbit Sense	+6.3 min (not significant)	−12.4 min	+2.2 pts	1 (partial)
Fitbit Charge 5	−5.7 min (not significant)	−16.9 min	+3.4 pts	1 (partial)
Oura Gen 3	Varies by algorithm (OSSA 1.0 vs 2.0)	Not reliably established	Not reliably established	**2 — uncalibrated**
Oura Ring 4 and Ring 5	**No published polysomnography validation of sleep staging found (PubMed, 7 Oct 2026); Ring 5 announced May 2026**	—	—	**2 — uncalibrated**

Bias values are from Schyvens et al. (2025): one laboratory night, 20 to 41 participants per device, so they are indicative rather than calibration constants, and they are re-derived whenever a larger validation appears. Context from de Zambotti et al. (2019), Chinoy et al. (2021) and Searles et al. (2026). Ring 4's overnight heart rate and heart-rate variability were validated against ECG (Dial et al., Physiological Reports 2025), which says nothing about its sleep staging. Bias estimates are cohort-specific and are wider in older adults — one study found total-sleep-time bias of −75 min in adults aged 56–80 versus healthy young adults. All tier-1 classifications are marked "partial" because no published validation covers every metric, and because a validation is a validation of a firmware snapshot, not of the device a user owns today.

8. Coverage — No Redistribution

Version 1.1.0 redistributed a missing component's weight pro-rata across the others. The reviewer identified the flaw: this implies the remaining components can substitute for the missing one. For Restoration, that directly contradicts the reason we include it — that subjective sleep is a non-redundant dimension carrying independent information. A score of 88 with Restoration is not the same construct as 88 without it.

```
SHI_partial = ( Σ wᵢ·sᵢ ) / ( Σ wᵢ )   for available components i

coverage    = ( Σ wᵢ ) / 1.00           expressed as a percentage
```

The arithmetic is similar. The presentation is not, and the presentation is the point.

Displayed	Example
Sleep Health Index	84
Coverage	90%
Missing	Restoration — no diary entries in the last 14 days
Confidence	Moderate
Source	Oura Ring Gen 3 · continuity scored against personal baseline (uncalibrated)

- **A partial score can never be classified "Optimal."** The Optimal band requires coverage ≥ 90%. A user cannot reach the top classification by simply not answering the questions.
- **Coverage is always displayed alongside the score.** Never in a submenu, never on hover.
- **The specific missing component is named.** "Restoration unavailable" — not "some data missing."
- **Nights with less than 4 hours of wear, under 80% inter-beat-interval validity, or a failed regularity computation are excluded from the window and from coverage.**

9. Versioning and Governance

Change	Bump	Obligation
Bug fix, no change to any output	PATCH	Changelog entry.
Change to a weight, curve, age band, calibration or threshold	**MINOR**	Published changelog with rationale and citation. Historical scores recomputed; both values retained and viewable. Users notified.
Addition or removal of a component	**MAJOR**	As above, plus a published rationale and 30-day notice. **This release is an instance: sleep stages were removed, so v1.1.0 → v2.0.0.**

9.1 Standing removal conditions

- **Age bands (§6):** revised or removed if better-screened normative data become available. We are explicitly aware that the Boulos norms may encode undiagnosed pathology, and that Ohayon et al. (2004) found the age effects on sleep parameters shrinking or disappearing with rigorous screening. If a well-screened normative dataset contradicts our bands, our bands lose.
- **Calibration tiers (§7):** any device moves to Tier 2 (uncalibrated) the moment its firmware changes without a discoverable version, or if its published validation is superseded.
- **Any component:** de-weighted if its evidence grade is downgraded by new primary literature. Grades reviewed annually against the published record.
- **The whole score:** if we cannot demonstrate that SHI has better test-retest reliability than its best single component, we will say so publicly and reconsider the composite. No competitor has ever published this statistic. We will.

9.2 What this score is not

- **Not a medical device, and not diagnostic.** It cannot rule out sleep apnea and cannot replace polysomnography.
- **Not validated against health outcomes.** No sleep score has ever been. Our components are weighted by evidence and effect size; the composite itself remains an untested construct, and we will not claim otherwise.
- **Not a prediction of today's performance.** That claim is routine across the category and has no support.
- **Not comparable across devices** where continuity is scored from an uncalibrated source (§7).

> **Why we published the review as well as the method**
>
> A user who wants to know why their score moved can read this document and reconstruct the arithmetic. A clinician or a statistician can inspect the reasoning and tell us where we are wrong; an independent data scientist just did, in ways that materially changed the product. A competitor can copy the weights, and is welcome to.
>
> The weights were never the moat. The moat is the willingness to publish the criticism, delete a feature that users expect, and record in the changelog exactly which argument we lost and why.

10. Worked Examples

10.1 A healthy 57-year-old (band 50–64, SRI 86)

Component	Weight	Score	Contribution
Regularity (SRI 86)	35%	91.8	32.1
Duration (7.2 h)	30%	100	30.0
Continuity (SE 83%, SOL 15 min, WASO 64 min)	15%	86.4	13.0
Circadian alignment (MCTQ completed, well aligned)	10%	100	10.0
Restoration	10%	85	8.5
Sleep stages (N3+REM 39%)	**0%**	displayed	**0.0**
SHI (raw)			**93.6**
G(SRI 86) = 1.00 — no regularity penalty			
SHI (final)			**93.6 — "Optimal"**

10.2 The shift worker (age 40, SRI 66) — soft gate versus hard gate

Sleeping 7.8 h on average, 88% efficiency, WASO 32 min, SOL 14 min, good subjective ratings — but sleep and wake times swing across a 3-hour window because of rotating shifts.

Component	Weight	Score	Contribution
Regularity (SRI 66)	35%	41.5	14.5
Duration (7.8 h)	30%	100	30.0
Continuity (band 35–49: SE 88%, SOL 14, WASO 32)	15%	94.6	14.2
Circadian alignment	10%	80	8.0
Restoration	10%	75	7.5
SHI (raw)			**74.2**

Regularity treatment	Result	Band
v1.1.0 — hard gate (SRI < 71.65 → cap 50)	**50.0**	Needs attention
v2.0.0 — soft multiplier, persistent across two windows: G = 0.752	**55.8**	Fair
v2.0.0 — soft multiplier, first window only: G = 0.876	**65.0**	Fair

The soft multiplier preserves everything the gate was for. This user is still pulled down decisively from 74 — the score still says that regularity is the dominant problem, and the interface still names it. What has gone is the false discontinuity: a user at SRI 71.6 is no longer treated as categorically different from one at 71.7, and a single disrupted month no longer produces the same verdict as a chronic pattern.

The state is labeled "schedule instability," not "poor sleep health," and the guidance addresses what this user can actually control. It does not tell them their sleep is fine, because it is not — and it does not exempt them from the finding because their schedule is not their fault.

Vitality Sleep Score v2.0.1 · Novak Consulting Ltd · Published in full, including its external review, and maintained under the versioning policy in §9.