Understanding Measures — Validity & Reliability

Companion to the Week 5 episode. Your hypothesis is only as good as your ruler.

1. The one idea

Ask two **different** questions of every measure: **Validity** — does it measure the construct I care about? **Reliability** — does it give the same value twice? A measure can be reliable yet invalid (consistently measuring the wrong thing). Measurement is upstream of everything — a bad measure = rigorously analysed nonsense.

2. Validity vs reliability — the four quadrants

Reliable + Valid ✓
Tight group, on the bullseye. Consistent AND correct — what you want.

Reliable, NOT valid ⚠
Tight group, off-target. Consistently measures the *wrong* thing — the seductive trap (steadiness fools you). A rock-steady "recovery score" that's never been validated.

Valid, NOT reliable
Scattered around the bullseye. Right on average, too noisy to trust any single value.

Neither
Scattered and off. Useless.

The trap: reliability *feels* like rigour, so "reliable but not valid" is the most dangerous quadrant — check the two properties *separately*. A device always gives a number; the question is whether it's a *valid* index of your construct, in athletes like yours (Impellizzeri & Marcora, 2009; Flake & Fried, 2020).

3. Reliability & the noise floor

Concept	Meaning
Typical error (Hopkins, 2000)	The random noise in a measure — SD of an individual's repeats, often a CV %. Every measure has one.
Systematic bias vs random error (Atkinson & Nevill, 1998)	Bias = learning/fatigue shifting *everyone* on retest; random = trial-to-trial scatter. Know both.

Why it matters — the noise floor: a signal *smaller* than your measurement noise is **invisible**, however real. If your test's typical error is 3% and you're chasing a 1% effect, you cannot see it. So: a

change must *exceed* the typical error to be believed (smallest worthwhile change vs error), and reliability **directly drives the sample size you need** (Week 7). A noisy measure can make a good study *incapable* of answering its question.

4. Match the design to the causal claim

Design	What it does	Strongest for
Observational	Watches; no intervention	Descriptive & hypothesis-generating; rarely causal (confounding, reverse causation)
Quasi-experimental	Intervention, no full randomisation (e.g. before/after; team vs team)	Causal-ish, but vulnerable to what you didn't control
Randomised controlled trial	Intervenes + randomises	Causation — randomisation balances measured *and* unmeasured factors

> **Rule:** the strength of causal claim you can make is *capped by your design*. Don't let the conclusion outrun the method (→ Week 11).

5. Do this — this week

Earn your outcome measure

1. Validity: write why it's a valid index of your construct for *your* athletes — cite validation evidence, or flag honestly where it's missing.
2. Reliability: find or **pilot-test** its **typical error** (test a few participants twice on *your* kit) and compare to the **smallest change you'd care about**.
3. Protocol: draft it in enough detail that someone else could reproduce it exactly (warm-up, timing, device settings, calibration, instructions).
If the noise floor > the effect you're chasing, you've learned something vital before wasting a season.

References

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine, 26*(4), 217–238.
Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465.
Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1–15.
Impellizzeri, F. M., & Marcora, S. M. (2009). Test validation in sport physiology: lessons learned from clinimetrics. *International Journal of Sports Physiology and Performance, 4*(2), 269–277.