Measuring the Inner Critic: How SMST, PHQ-9, BDI-II, Zung, CES-D and WHO-5 Compare for Tracking Change

Ernesto Manolo Beelke

Leadership Evolution Newsletter 15.11.2025

https://www.linkedin.com/pulse/measuring-inner-critic-how-smst-phq-9-bdi-ii-zung-beelke--hckoe

Abstract

The “inner critic” is not a diagnosis but a pattern of harsh self-evaluation, perfectionism and shame that erodes performance and well-being. High self-criticism predicts poorer psychotherapy outcomes and is a explicit target in several intervention models (Löw et al., 2020; Wakelin et al., 2022). Yet most routine instruments were designed to detect depression, not to monitor change in self-management or resilience. This article compares the Self-Management Self-Test (SMST) with classical depression and well-being scales: PHQ-9 and PHQ-2, Beck Depression Inventory-II (BDI-II), Zung Self-Rating Depression Scale (Zung-SDS), Center for Epidemiologic Studies Depression Scale (CES-D) and the WHO-5 Well-Being Index. We review constructs, psychometrics and practical advantages, then propose specific measurement packages for coaching, psychotherapy, digital interventions and clinical trials focused on inner-critic work. A three-axis dashboard combining self-management (SMST), symptom burden (PHQ-9 or BDI-II) and well-being (WHO-5) offers a more precise view of change than any single instrument.

---

1. Introduction: when the inner critic becomes a KPI

The inner critic is the part of mental life that says “not good enough” at every turn. It drives overwork, perfectionism and avoidance. High trait self-criticism is consistently associated with greater depressive symptom severity, broader psychopathology and poorer outcomes in psychotherapy across diagnostic groups (Löw et al., 2020; Papa et al., 2024; Zaccari et al., 2024).

If we want to reduce the impact of the inner critic, we need to measure change. However, almost all mainstream scales in routine use were built to answer a different question: “Does this person have depression?” They do that job well. They do not necessarily tell us whether an intervention has changed how people manage themselves, make decisions or experience their own life.

For inner-critic work, at least three dimensions matter:

1. **Symptoms and distress** Depressed mood, anhedonia, guilt, fatigue, cognitive slowing.

2. **Positive well-being** Vitality, interest, calmness, sense of engagement in life.

3. **Functioning and self-management** Capacity to organise daily life, maintain relationships, make and implement decisions under stress.

Classical depression scales such as PHQ-9, BDI-II, Zung-SDS and CES-D largely live on axis 1 (Negeri et al., 2021; Vilagut et al., 2016; Wang & Gorenstein, 2013). WHO-5 primarily captures axis 2 (Topp et al., 2015). The Self-Management Self-Test (SMST) was explicitly built for axis 3 (Wehmeier et al., 2020).

The aim of this article is not to crown a single "best" scale, but to answer a practical question: *Which instruments, alone or in combination, are most useful for tracking change in inner-critic work and individual improvement over time?*

2. Concepts and psychometric priorities

2.1 Self-criticism as a transdiagnostic process

Self-criticism is best understood as a transdiagnostic vulnerability factor. Meta-analytic data show that higher pre-treatment self-criticism predicts poorer psychotherapy outcome across disorders (Löw et al., 2020). Conversely, self-compassion-based interventions produce medium-sized reductions in self-criticism and parallel improvements in mood (Wakelin et al., 2022).
This aligns with dimensional views of depression: self-criticism is not identical with depressive symptoms, but is closely linked to them and interacts with other processes such as perfectionism and shame (Zaccari et al., 2024).

2.2 What makes a tool useful for inner-critic work?

Four psychometric and practical aspects matter:

1. **Construct:** Does the scale measure symptoms, well-being, functioning, or some mix?

2. **Reliability and validity:** Internal consistency (Cronbach's alpha around .80 or higher), test–retest reliability, convergent and criterion validity.

3. **Accuracy for depression, where relevant:** Sensitivity, specificity and area under the receiver operating characteristic (ROC) curve (AUC) for major depression diagnoses, using structured interviews as reference (Negeri et al., 2021; Levis et al., 2019; Vilagut et al., 2016).

4. **Responsiveness and burden:** Is it sensitive to change but still brief enough to use repeatedly without exhausting people?

With that frame, we start with the one instrument that was not built as a depression scale at all: the SMST.

3. The Self-Management Self-Test (SMST)

3.1 Construct and structure

The SMST is a 5-item self-rating scale designed to measure self-management competence in individuals with or without psychiatric disorders (Wehmeier et al., 2020). Items cover key domains:

- awareness of internal and external states
- maintaining relationships and social contact
- planning and future orientation
- decision making
- taking and sustaining action in everyday life

Each item is rated from 0 to 4, yielding a total score from 0 to 20, with higher scores indicating better self-management (Wehmeier et al., 2020; Wehmeier, 2016).

3.2 Validation study

Wehmeier and colleagues (2020) evaluated the SMST in:

- **87 inpatients with major depression**, and
- **595 adults from the general population**, with a matched non-clinical subsample of 87 individuals screened as free of psychiatric disorder using the PHQ.

Participants completed the SMST and five additional stress-related instruments, including the PHQ depression scale and the Multidimensional Fatigue Inventory. Key findings were:

- **Convergent validity:** Correlations between SMST and stress-related measures ranged from r = −.40 to −.64, with the strongest associations for fatigue and depressive symptoms, as expected for overlapping but distinct constructs (Wehmeier et al., 2020).

- **Internal consistency:** Cronbach's alpha was .80, indicating good internal reliability.

- **Test–retest reliability:** Over 4 to 6 weeks, test–retest reliability was r = .71 in the population sample. The authors interpret this as sufficient stability combined with sensitivity to change.

- **Group discrimination:** Depressed inpatients had a mean SMST score of 9.36 (SD 3.39) versus 12.94 (SD 2.47) in the matched non-clinical subsample, corresponding to a very large effect size (d = 1.3).

- **Diagnostic discrimination:** ROC analysis for distinguishing depressed inpatients from non-depressed controls yielded an AUC of 0.81, which is in the "good" range and suggests useful discrimination, although SMST is not intended as a diagnostic depression scale.

Overall, the SMST behaves like a short functional outcome measure that is strongly, but not exclusively, related to depression.

3.3 Why SMST is relevant for the inner critic

Self-criticism does not just hurt feelings. It changes behaviour: people hesitate, avoid decisions, withdraw from relationships and fail to act on plans. Meta-analytic work confirms that higher self-criticism is associated with poorer psychotherapy outcomes (Löw et al., 2020) and that self-compassion-based interventions reduce self-criticism in parallel with improvements in psychological well-being (Wakelin et al., 2022).
The SMST does not ask about self-criticism directly. It asks about what self-criticism does to life. For interventions targeting the inner critic, an ideal response profile is:

- SMST increasing (improved self-management)

- WHO-5 increasing (better well-being)

- PHQ-9 or BDI-II decreasing or remaining in the non-clinical range

This makes SMST a particularly attractive primary performance measure in non-diagnostic contexts such as coaching, leadership development and digital resilience programs.

4. Classical depression scales

4.1 PHQ-9 and PHQ-2

The Patient Health Questionnaire-9 is currently the most widely used depression screener in primary care and many research settings. An individual participant data meta-analysis of 58 studies (n > 17,000) reported that a cut-off of 10 or more maximised combined sensitivity and specificity for major depression, with pooled sensitivity and specificity around 0.85 and an AUC close to 0.88 to 0.89 (Levis et al., 2019; Negeri et al., 2021).
The PHQ-2 comprises the first two PHQ-9 items (anhedonia and depressed mood). It is commonly used as a very brief screener. A two-step strategy with PHQ-2 followed by PHQ-9 (PHQ-2 ≥ 2 → PHQ-9) yields an AUC of about 0.90 versus structured interview diagnoses (Levis et al., 2019; He et al., 2019).
For inner-critic work, PHQ-9 and PHQ-2 have three main roles:

1. **Safety net.** They detect clinically relevant depression that requires proper diagnosis and treatment rather than pure coaching.

2. **Symptom trajectory.** They provide a simple measure of depressive symptom change over time.

3. **Regulatory alignment.** PHQ-9 is familiar to clinicians, regulators and payers.

But they mostly cover axis 1: distress and symptoms. They say little about self-management or positive well-being.

4.2 Beck Depression Inventory-II (BDI-II)

The BDI-II is a 21-item self-report scale aligned with DSM criteria, with a strong emphasis on cognitive and affective content such as self-dislike, guilt and pessimism (Beck et al., 1996). A systematic review of 70 validation studies in various medical settings found high internal consistency (typical alpha > .89), good test–retest reliability and strong convergent validity with other depression and anxiety measures (Wang & Gorenstein, 2013).
Across studies, AUC values for detecting major depression often approach 0.90, with sensitivity and specificity in the 0.80 to 0.90 range at appropriate cut-offs (Wang & Gorenstein, 2013). Psychometrically, BDI-II is excellent.
Its main disadvantages are:

- proprietary status and licensing cost
- longer administration time than PHQ-9
- content that may feel overly clinical in workplace or coaching contexts

However, because BDI-II explicitly addresses self-critical cognitions, it can be particularly useful where inner-critic work is embedded in formal psychotherapy or research in clinical populations.

4.3 Zung Self-Rating Depression Scale (Zung-SDS)

The Zung-SDS is a 20-item scale developed in the 1960s to measure depressive affect, with items covering psychological and somatic symptoms (Zung, 1965). Internal consistency is typically acceptable (alpha around .80), and convergent validity with other depression scales is good.
Recent work suggests that Zung-SDS has adequate screening accuracy but, in comparative meta-analyses, offers no clear advantage over newer tools such as PHQ-9 or WHO-5 (Vilagut et al., 2016). Sensitivity and specificity in various studies often lie in the 0.70 to 0.80 range, depending on population and cut-offs.
Zung-SDS remains useful in legacy research or where it is institutional standard, but for new programs focused on inner-critic change it is rarely the optimal first choice.

4.4 CES-D

The Center for Epidemiologic Studies Depression Scale (CES-D) is a 20-item instrument designed for epidemiological use. It assesses frequency of depressive symptoms over the past week.
A systematic review with meta-analysis of 28 studies showed that CES-D has high sensitivity (around 0.85 to 0.87) but more modest specificity (approximately 0.70) for major depression at common cut-offs (Vilagut et al., 2016). It is therefore excellent for screening in large populations.
However, its length and lower specificity, combined with a focus on symptom distress rather than functioning or self-management, make it less attractive for repeated outcome measurement in individual change programs.

5. WHO-5: the well-being axis

The WHO-5 Well-Being Index is a 5-item scale that measures positive subjective well-being: cheerfulness, calmness, energy, interest and feeling refreshed after sleep (World Health Organization, 1998). Items are rated over the past two weeks on a 0–5 scale and usually converted to a percentage score (0–100).
A systematic review of 213 studies concluded that the WHO-5 has high clinimetric validity, good internal consistency (typically alpha around .80), and functions both as a sensitive outcome measure and as a screening tool for depression with good

sensitivity and specificity (Topp et al., 2015). A percentage score below 50 has been suggested as indicative of poor well-being, warranting further assessment (World Health Organization, 1998).
More recent work confirms that WHO-5 behaves as a unidimensional scale with robust psychometric properties across medical and psychiatric populations and across languages (Domenech et al., 2025).
For inner-critic work, WHO-5 has several advantages:

- It frames change in positive terms (more vitality, more interest), which fits coaching and resilience narratives better than "depression scores".
- It is extremely brief and therefore suitable for repeated monitoring.
- It is transdiagnostic and applicable in mixed populations.

The limitation is non-specificity: WHO-5 improves with better sleep, improved physical health, relational changes and financial relief, not only with a quieter inner critic. Interpretation requires context and ideally a functional measure such as SMST.

6. Comparative overview

Table 1 summarises the key instruments and their practical value for inner-critic and improvement work.

Table 1. Overview of instruments relevant for inner-critic work

Instrument	Primary construct	Items & scale	Typical cut-offs for depression	Accuracy snapshot (approx.)	Licensing	Core strengths for inner-critic / improvement work	Main constraints
SMST	Self-management competence (awareness, relationships, planning, decision & action)	5 items, 0–4 (total 0–20)	No official diagnostic cut-off; AUC 0.81 distinguishing PHQ-defined MDD vs controls(PubMed)	Sensitivity/specificity each ~0.75–0.80 at optimal threshold	Author permission; freeware for many uses	Directly reflects behavioural impact of the inner critic; very brief; sensitive to change	Not a DSM-based depression tool; fewer norms & translations; not suitable as sole safety screen
PHQ-9	Depressive symptoms (DSM)	9 items, 0–3 (0–27)	≥10 for possible MDD	Large IPD meta-analyses: sens. ≈0.85, spec. ≈0.85, AUC ≈0.88–0.89(PubMed)	Free, widely translated	Standard clinical metric; strong evidence base; works as severity scale and safety net	Focus on pathology; not ideal as sole measure in coaching or leadership settings
PHQ-2	Core depressive symptoms (mood, anhedonia)	2 items, 0–3 (0–6)	≥2 or ≥3 as screener	Combination PHQ-2≥2 → PHQ-9≥10: sens. 0.82, spec. 0.87, AUC 0.90(JAMA Network)	Free	Ultra-brief front-door screen; good gateway before full assessment	Too coarse for outcome tracking; high false-positive rate at low cut-offs
BDI-II	Depressive symptom severity with strong cognitive/self-evaluative emphasis	21 items, 0–3 (0–63)	≥14–19 mild; ≥20 moderate	Meta-analysis: α ≈ .89, test–retest ≈ .75; AUC often around 0.90 depending on sample(ResearchGate)	Proprietary	Rich data on guilt, self-dislike and pessimism; excellent severity tracking	Licensing cost; higher burden; may not fit non-clinical or workplace contexts
Zung-SDS	General depressive affect (psychological + somatic)	20 items, 1–4 (raw 20–80)	Often ≥39–50	In elderly: sens. 79%, spec. 72% at cut-off 39; some studies report AUC ≈0.90(PMC)	Generally free	Long history, acceptable accuracy, good for research continuity	Longer, less modern; mixed somatic/cognitive items; weaker evidence than PHQ-9/WHO-5
CES-D (20-item)	Depressive symptom frequency	20 items, 0–3 (0–60)	≥16 or ≥20	Meta-analysis: sens. ≈0.86–0.87, spec. ≈0.70; acceptable screening accuracy but not diagnostic(PubMed)	Public domain	Very sensitive; standard in epidemiology	Less specific; relatively long; content more about distress than functioning
WHO-5	Positive well-being (cheerfulness, calmness, energy, interest, refreshed sleep)	5 items, 0–5 (0–25, usually ×4 for 0–100)	≤50% for impaired well-being; ≤28% often used for probable depression	Systematic review: sens. ≈0.86, spec. ≈0.81 for depression; high clinimetric validity(Karger Publishers)	Free, many translations	Positive framing; excellent as outcome metric; works across mental and somatic disorders	Indirect for inner critic; changes can reflect many factors beyond self-management

7.1 Coaching, leadership development and workplace programs

Goals: reduce destructive self-criticism, improve resilience and performance in people who are mostly not seeking clinical care.
Suggested core battery:

- SMST as primary performance outcome (self-management)
- WHO-5 as primary well-being outcome
- PHQ-2 at baseline and key milestones as a safety screen; PHQ-9 only if PHQ-2 is elevated or if SMST/WHO-5 deteriorate unexpectedly

This combination creates a three-axis dashboard:

- "Can I organise my life?" (SMST)
- "How does my life feel?" (WHO-5)
- "Is there clinically relevant depression risk?" (PHQ-2/9)

Measurement every 4–6 weeks is usually sufficient. Weekly measurement of WHO-5 is feasible in digital formats, but SMST should not be administered so frequently that responses become mechanical.

7.2 Psychotherapy and clinical mental-health services

Goals: treat depression, anxiety and related disorders while explicitly addressing self-criticism as a maintaining process.
Suggested core battery:

- PHQ-9 or BDI-II as main symptom severity measure, depending on local standards and licensing
- SMST as functional/process outcome
- WHO-5 as well-being outcome

Where self-criticism is a central target (e.g. compassion-focused therapy, emotion-focused therapy), it is useful to add a dedicated self-criticism or self-compassion

measure, such as the Forms of Self-Criticizing/Attacking and Self-Reassuring Scale or the Self-Compassion Scale (Wakelin et al., 2022; Zaccari et al., 2024).
Measurement at intake, every 4–6 weeks and discharge usually balances information and burden. In intensive day-programs or inpatient settings, more frequent WHO-5 measurement can provide fast feedback without overwhelming patients.

7.3 Digital mental-health and self-help applications

Goals: scalable low-burden monitoring in heterogeneous populations, with early detection of deterioration.
Suggested strategy:

- Onboarding: PHQ-2, WHO-5, SMST.
- Ongoing: WHO-5 weekly or bi-weekly; SMST every 4–8 weeks; PHQ-9 if WHO-5 worsens or if PHQ-2 flags risk.

This tiered approach respects psychometric quality while keeping the questionnaire load tolerable for users.

7.4 Clinical trials targeting resilience, burnout or self-management

For trials that explicitly target self-management, resilience or inner-critic processes, SMST is a plausible candidate for a primary patient-reported outcome, supported by PHQ-9 or BDI-II and WHO-5 as key secondary outcomes (Wehmeier et al., 2020; Topp et al., 2015). Regulators and HTA bodies are increasingly open to multi-dimensional PRO strategies, as long as instruments are validated and the hierarchy of endpoints is prespecified.

8. Implementation issues: from cut-offs to change indices

8.1 Cut-offs versus within-person change

Cut-offs are useful for triage (e.g. "PHQ-9 ≥ 10: consider major depression"), but for inner-critic work, within-person change is more informative. Common rules of thumb include:

- PHQ-9: a reduction of 5 or more points is often considered clinically meaningful (Manea et al., 2015).
- WHO-5: an increase of 10–20 percentage points is typically interpreted as relevant change (Topp et al., 2015; World Health Organization, 1998).

- SMST: change should be anchored in behavioural milestones (e.g. new decisions taken, boundaries set), because normative thresholds are less established.

8.2 Reliable change and measurement error

Formal reliable change indices require standard deviations and test–retest reliability coefficients. For PHQ-9 and BDI-II, these are well documented (Negeri et al., 2021; Wang & Gorenstein, 2013). For SMST, initial data (SD around 2.5–3.4 and test–retest r = .71) allow calculation of reliable change thresholds in specific samples (Wehmeier et al., 2020).

8.3 Floor and ceiling effects

Highly functional samples (for example senior executives) may show near-maximum SMST and WHO-5 values at baseline. This can produce ceiling effects where further improvement is not captured by total scores. In such cases it is useful to:

- pay attention to item-level changes, and
- complement questionnaires with qualitative indicators or targeted self-criticism scales.

9. Beyond questionnaires: AI and language-based measures

There is growing interest in using language, voice and digital behaviour to infer mood and cognitive style. Models trained on text, speech and smartphone data can sometimes achieve diagnostic accuracy comparable to traditional scales for depression (Zaccari et al., 2024). For the inner critic, language models that track self-attacking phrases could in principle offer high-frequency monitoring.
However, current systems face major issues:

- limited generalisability beyond training datasets
- opaque decision processes
- unresolved privacy and regulatory questions

For the foreseeable future, questionnaires such as SMST, PHQ-9 and WHO-5 will remain the primary standards. AI-derived metrics can complement them, not replace them.

10. Conclusion: a three-axis dashboard for the inner critic

Treating the inner critic as "just depression" is too narrow. Self-criticism is a process that affects symptoms, functioning and well-being in different ways. The evidence suggests a simple but robust measurement strategy:

1. **Self-management axis:** SMST
2. **Symptom axis:** PHQ-9 or BDI-II, with PHQ-2 as gatekeeper
3. **Well-being axis:** WHO-5

Zung-SDS and CES-D remain acceptable in legacy contexts but bring no special advantages for inner-critic work that are not better covered by the trio above.
Used together, these instruments produce a more honest picture: whether the inner critic is still running the show, or whether it has finally been pushed out of the driver's seat and relegated to a grumpy passenger in the back.

References

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Psychological Corporation.

Domenech, A., et al. (2025). Systematic review of the use of the WHO-5 Well-Being Index across conditions. *Advances in Therapy*.

He, C., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., … Thombs, B. D. (2019). The accuracy of the Patient Health Questionnaire-9 algorithm for screening to detect major depression: An individual participant data meta-analysis. *Psychotherapy and Psychosomatics, 88*(4), 221–234.

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ, 365*, l1476.

Löw, C. A., Schauenburg, H., & Dinger, U. (2020). Self-criticism and psychotherapy outcome: A systematic review and meta-analysis. *Clinical Psychology Review, 75*, 101808.

Manea, L., Gilbody, S., & McMillan, D. (2015). A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm for major depressive disorder. *Journal of Affective Disorders, 178*, 67–74.

Negeri, Z. F., Levis, B., Sun, Y., He, C., Krishnan, A., Wu, Y., … Thombs, B. D. (2021). Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: Updated systematic review and individual participant data meta-analysis. *BMJ, 375*, n2183.

Papa, C., et al. (2024). “You’re ugly and bad!”: A path analysis of the interplay between self-criticism, self-compassion, and psychopathology. *Current Psychology*.
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.
Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): A systematic review with meta-analysis. *PLoS ONE, 11*(5), e0155431.
Wakelin, K. E., Perman, G., & Simonds, L. M. (2022). Effectiveness of self-compassion-related interventions for reducing self-criticism: A systematic review and meta-analysis. *Clinical Psychology & Psychotherapy, 29*(3), 824–842.
Wang, Y.-P., & Gorenstein, C. (2013). Assessment of depression in medical patients: A systematic review of the utility of the Beck Depression Inventory-II. *Clinics, 68*(9), 1274–1287.
Wehmeier, P. M. (2016). *Erfolg ist, wenn es mir gut geht! Burnout vermeiden durch Selbstmanagement* (2nd ed.). Vandenhoeck & Ruprecht.
Wehmeier, P. M., Fox, T., Doerr, J. M., Schnierer, N., Bender, M., & Nater, U. M. (2020). Development and validation of a brief measure of self-management competence: The Self-Management Self-Test (SMST). *Therapeutic Innovation & Regulatory Science, 54*(3), 534–543.
World Health Organization. (1998). *Wellbeing measures in primary health care: The DepCare project*. WHO Regional Office for Europe.
Zaccari, V., et al. (2024). State of the art of the literature on definitions of self-criticism. *Frontiers in Psychiatry, 15*, 1239696.
Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*(1), 63–70.