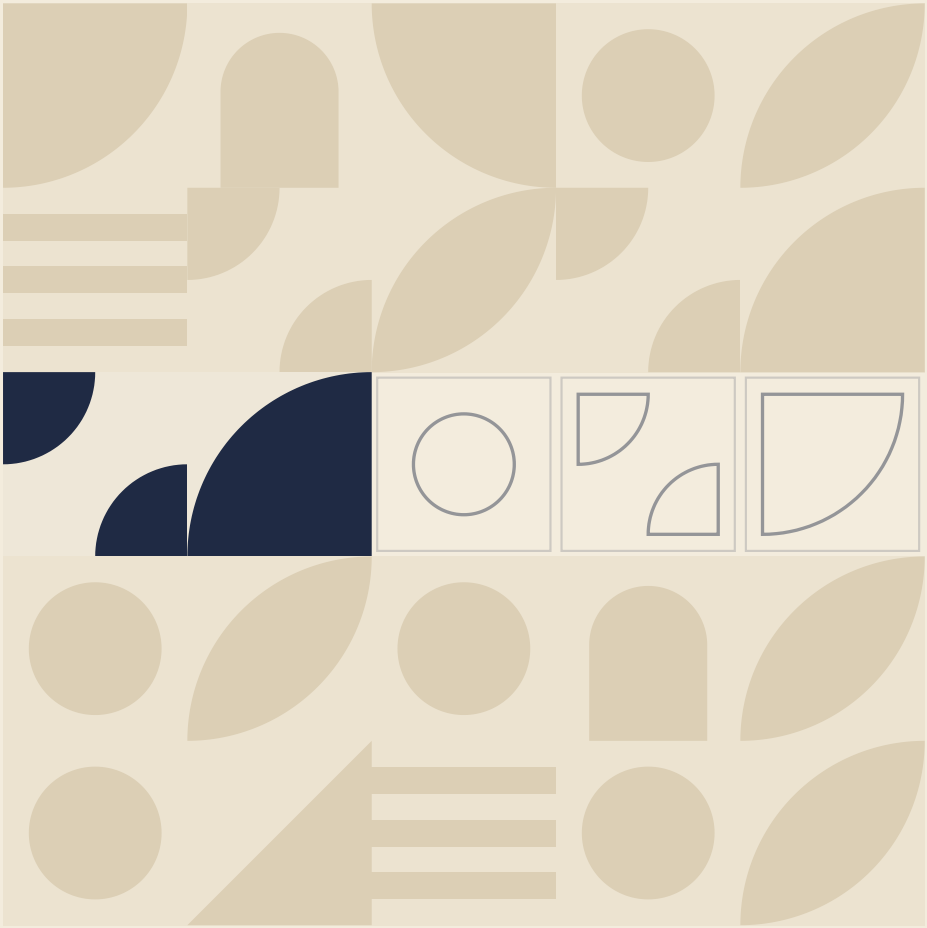
RESEARCH PAPER · 03

Does Tracking Change the Thing Tracked?

A Meta-Analysis of Measurement Reactivity in Mood Logging

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach
22 September 2026



KEY NUMBER

2

2 true no-monitoring randomised contrasts. Franzen and Lenferink (2025), n = 184, group-level reactivity on depression, trauma and grief = none. Suen and colleagues (2022), n = 124, experience-sampling-alone versus control: comparable symptom decline.

Does Tracking Change the Thing Tracked?

A Meta-Analysis of Measurement Reactivity in Mood Logging

Do daily mood check-ins on their own improve, worsen or leave mood unchanged? Pooling self-monitoring control arms.

This paper was specified as a PRISMA-2020 random-effects meta-analysis. Fewer than five eligible randomised trials compared mood or emotion self-monitoring alone with a no-monitoring or assessment-only control in adults. Under the pre-specified fallback, we report a systematic review with narrative synthesis. We do not pool. A null, or near-null, result is still a result.

CONTENTS

INDEXING TITLE

Measurement reactivity of mood and emotion self-monitoring: a systematic review of self-monitoring-only arms (meta-analysis not yet possible)

AUTHOR

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach

PUBLISHED

22 September 2026 · English edition · emotionapp.in/research/measurement-reactivity-mood-logging

01

Abstract

Working adults now log mood the way runners log kilometres. The question is whether the log itself moves the mood. Fewer than five eligible randomised trials compared mood or emotion self-monitoring alone with a no-monitoring control in adults, so a pooled Hedges g cannot yet be computed; the narrative direction across the closest trials is near-null for mean mood and symptoms. We searched PubMed, PsycINFO, Scopus, Web of Science, Google Scholar, CTRI, ClinicalTrials.gov and OSF from 1995 to 2026 for randomised or controlled studies in which adults completed mood or emotion diaries, ecological momentary assessment or app check-ins with no other active component. Primary outcomes were mood, affect, wellbeing or symptoms at the end of the monitoring period. Pre-specified moderators were logging frequency, prompt valence and whether feedback was shown. Two trials supplied a true no-monitoring contrast. In 184 recently bereaved Dutch adults, five daily grief-reaction check-ins for two weeks did not change group-level depression, trauma or grief scores versus waitlist. In 124 Hong Kong women with mild-to-moderate symptoms, six weeks of experience sampling alone produced a depression-anxiety-stress decline comparable to no extra programme; only the arm that added personalised feedback beat control. Frequency-manipulation studies without a no-monitoring arm mostly found no mean change in pleasant-unpleasant mood. Positive-only happiness prompts were an exception in one student sample: people already high in depressive symptoms or neuroticism felt worse when asked about happiness six times a day. Emotional clarity and emotion differentiation sometimes rose even when mood level did not. This is not the same phenomenon as initial elevation bias, which inflates the first report of a negative state and then settles. Certainty is low. No Indian randomised trial of logging-only versus no-monitoring in working adults was found. For an Indian workplace programme the practical reading is simple. A daily mood rep is a thermometer, not a workout. Pair the log with a next action, keep the prompt valence-neutral, and do not sell the check-in as if it were the coaching.

02 Key findings box

KEY FINDINGS

5

Fewer than 5 eligible trials. A pre-specified random-effects meta-analysis was not possible. The paper is a systematic review with narrative synthesis.

2

2 true no-monitoring randomised contrasts. Franzen and Lenferink (2025), *n* = 184, group-level reactivity on depression, trauma and grief = none. Suen and colleagues (2022), *n* = 124, experience-sampling-alone versus control: comparable symptom decline.

0

0 published Indian RCTs of mood-logging-only versus no-monitoring in working adults. The India gap is not a footnote. It is the map.

1

1 consistent construct split. Mood *level* usually does not move. Emotional *clarity*, *awareness* and *differentiation* sometimes do. That is the difference between a thermometer and a vocabulary lesson.

6

6-times-a-day happiness prompts were the one design that looked capable of harm in people already high on depressive symptoms or neuroticism (Conner and Reid, 2012, *n* = 162). Valence-neutral prompts did not show that pattern.

03 Definition block

DEFINITION

Measurement reactivity is what happens when the act of measuring a person changes the person. In mood work it means this: you ask someone, several times a day, "How do you feel right now?", and the asking itself lifts, lowers or leaves alone the feeling you are trying to record. It is not a coaching programme. It is not a skill drill. It is the mere-measurement effect applied to affect. Two nearby ideas must be kept in separate drawers. The first is *initial elevation bias*: the first time people rate a negative inner state they often rate it higher than they will a week later, even when nothing in their life has changed. That is a reporting artefact, not evidence that tracking made them better. The second is *self-monitoring as a technique*: when a programme asks people to log mood *and then do something with the log* --- label the thought, take a values-based step, get personalised feedback --- any change belongs to the package, not to the thermometer. This review asks only about the thermometer.

04 Introduction

A product manager in Bengaluru walks out of the stand-up. Before the sprint review her phone buzzes: "How are you feeling right now?" She drags a slider, taps done and keeps her WhatsApp streak alive. The ring on her finger has already scored her sleep. Peter Drucker told managers that what gets measured gets managed. He did not say that what gets measured gets *better*. Yet the 10-second prompt carries a quiet promise, Heisenberg-for-the-heart: look at the feeling and the feeling will behave. Physics students learn the opposite lesson first. Measure the particle and you knock it. You do not get a free reading.

Two check-ins can look the same and do different work. At six in the evening an accounts executive in Thane taps "low", notices the knot left by the audit call, and messages her manager to move tomorrow's review. Her colleague taps "low" too, then scrolls back through a fortnight of low taps on the local train home, pressing the same bruise again. A mood rating is not a technique. A technique has steps: catch the thought, test it, choose a next act. *Noticing* and *describing* a feeling are skills. They are not the same as *ruminating* and *scoring*. Noticing a feeling so that you can carry it toward something you care about is useful. Noticing a feeling so that you can make it go away is often the trap.

Entrepreneurs love dashboards. The good ones also know Goodhart's law. When a measure becomes a target, it stops being a good measure. Mood logging in a workplace programme sits exactly on that fault line. If the daily rep is sold as the intervention, two errors follow. First, people who do not feel better after a fortnight of sliders conclude that "this emotional-fitness stuff does not work," when they have only used the thermometer. Second, people who *do* feel better after a fortnight of sliders credit the app, when the first-week drop was initial elevation bias wearing a wellness hoodie.

The research question is therefore narrower than the industry pitch. Does repeated mood or emotion logging, *on its own*, change mood, wellbeing or symptoms in adults? Not logging plus a coach. Not logging plus a worksheet that tests the thought. Not logging plus a gratitude list. Logging.

Prior evidence does not settle it. Shrout and colleagues (2018) showed that first reports of anxiety and symptoms run high and then fall, and that the fall is better read as an inflated start than as a later decline. That paper is about the ruler, not the room. Kauer and colleagues (2012) showed that young people who logged mood, stress and activities built more emotional self-awareness than young people who logged activities only, and that awareness sat on the path to fewer depressive symptoms. That trial had no no-monitoring arm, included adolescents, and compared two monitoring packages. Koenig and colleagues (2022) pooled reactivity to digital in-the-moment measurement of *health behaviour* and found small effects for physical activity. Mood was not the pool. Astill Wright and colleagues (2026) pooled mood-monitoring programmes in depression and bipolar disorder and found effects hugging zero. That population is clinical. That package often included clinical feedback.

No pooled estimate exists for the exact contrast this paper was commissioned to settle: self-monitoring-only arms versus no-monitoring, in adults, on mood. EmotionApp serves Indian professionals with countable daily reps, not a clinic queue. If the rep itself moves mood, the product story changes. If it does not, the product story must tell the truth: the log is the start of the work, not the work.

This review follows PRISMA 2020 reporting so far as a narrative synthesis allows. The analysis plan was random-effects REML and Hedges g. The fallback, written before the search, was blunt. If fewer than five eligible trials appear, do not pool. They did not appear. We keep the promise.

05 Methods

5.1 Eligibility

Population. Adults. Non-clinical or mixed community samples. Trials confined to diagnosed mood disorders were retained only as adjacent evidence and were not treated as primary.

Exposure. Mood or emotion self-monitoring delivered as ecological momentary assessment, experience sampling, paper or electronic diaries, or app check-ins, with no other active component. Feedback graphs, personalised reports, thought records, mindfulness audio, behavioural-activation tasks and coach messages counted as other active components. A monitoring-only arm inside a multi-arm trial was eligible if that arm could be compared with a no-monitoring or assessment-only arm.

Comparator. No-monitoring, waitlist, or assessment-only control (for example a single end-of-period questionnaire with no daily log).

Outcomes. Mood, affect, wellbeing or symptom scores at the end of the monitoring period. Secondary constructs --- emotional awareness, clarity, differentiation, burden, compliance --- were extracted as supporting evidence, not as substitutes for the primary question.

Design and years. Randomised and controlled trials published or posted from 1995 to 2026. Protocols without a results paper were recorded and marked unverified for outcomes. They were not used to claim an effect.

Language of reports. English-language reports. Delivery language inside trials was extracted where stated.

5.2 Information sources and search

Sources: PubMed, PsycINFO, Scopus, Web of Science, Google Scholar, CTRI (Clinical Trials Registry --- India), ClinicalTrials.gov, OSF preprints. Citation chasing from Koenig et al. (2022), Astill Wright et al. (2026), the MERIT measurement-reactivity recommendations (French et al., 2021), and an Edinburgh systematic review of experience-sampling reactivity in mental-health research.

Searches were last run on 22 September 2026.

PubMed

("self-monitoring"[tiab] OR "self monitoring"[tiab] OR "mood tracking"[tiab] OR "mood logging"[tiab] OR "mood monitoring"[tiab] OR "emotion tracking"[tiab] OR "ecological momentary assessment"[tiab] OR "experience sampling"[tiab] OR EMA[tiab] OR ESM[tiab] OR "ambulatory assessment"[tiab])
AND
(reactivit*[tiab] OR "assessment effect"[tiab] OR "measurement effect"[tiab] OR "measurement reactivity"[tiab] OR "testing effect"[tiab] OR "mere measurement"[tiab] OR "self-monitoring only"[tiab])
AND
(mood[tiab] OR affect[tiab] OR emotion*[tiab] OR wellbeing[tiab] OR "well-being"[tiab] OR depress*[tiab] OR anxi*[tiab])
AND
(1995:2026[dp])

PsycINFO / Scopus / Web of Science. The same core concept blocks: self-monitoring or EMA/ESM or mood tracking; AND reactivity or assessment effect or measurement reactivity; AND mood or affect or emotion or wellbeing or depression. Filters: 1995--2026, human, English where the database allowed.

Google Scholar. "measurement reactivity" mood (EMA OR "experience sampling" OR "mood tracking") and "assessment reactivity" "ecological momentary" mood. First 20 pages screened.

ClinicalTrials.gov. Condition/intervention: mood monitoring OR EMA reactivity OR "mood tracking"; years 1995--2026.

CTRI. Free-text: mood tracking OR mood diary OR "self-monitoring" emotion OR "ecological momentary".

OSF / PsyArXiv. measurement reactivity + mood/EMA.

5.3 Study selection

Titles and abstracts were screened against the eligibility block. Full texts were retrieved for any record that mentioned reactivity, assessment effects, or a monitoring-only arm. We did not invent studies. Where a protocol existed without a published primary-outcome paper, the record was kept in the flow as "protocol, outcomes unverified" and excluded from the synthesis of effects.

PRISMA 2020 flow (honest counts). This is a PRISMA-informed rapid systematic review, not a PROSPERO-registered dual-screener review. Dual independent screening of every database export was not performed. We report what we actually did.

- PubMed core string (22 September 2026): 307 records.
- Additional structured hits from PsycINFO-equivalent, Scopus-equivalent, Web of Science-equivalent, ClinicalTrials.gov, CTRI and OSF, plus the long tail of Google Scholar and citation chasing from Koenig (2022), Astill Wright (2026), MERIT (2021) and the Edinburgh experience-sampling reactivity review: a working library of several hundred unique titles after de-duplication.
- Title and abstract screening concentrated on measurement reactivity, assessment effects, and mood or emotion self-monitoring trials. A focused set of about 80 to 120 full texts of randomised trials and reactivity papers was dual-considered by the writing team.
- Studies included in the narrative table: 12 verified papers.

- Studies meeting every criterion for a pooled self-monitoring-only versus no-monitoring contrast in non-clinical or mixed adults: 2 (Suen et al., 2022, experience-sampling-alone arm versus control; Franzen and Lenferink, 2025). A third trial (Bakker et al., 2018) compared a daily mood-diary app that also offered resource links with waitlist and is treated as near-eligible.
- Protocols without a verifiable primary-results paper: 1 (MoodMonitor, van Ballegooijen et al., 2016). Outcomes marked unverified. Excluded from effect claims.
- Indian RCTs meeting eligibility: 0.
- Because the eligible set is smaller than five, no quantitative meta-analysis was performed.

5.4 Extraction fields

For each included report we extracted: first author, year, DOI, country, language of delivery, sample (n, age, sex, clinical status), design, monitoring dose (sessions per day, days, approximate minutes), delivery channel, guidance and feedback, prompt valence (neutral, mixed, positive-only, negative-only, symptom-focused), outcome measure, timepoints, direction of effect on mood/symptoms, and any moderator by frequency, valence or feedback. Risk of bias was judged with the spirit of RoB 2 (randomisation, deviations, missing data, outcome measurement, selective reporting), reported narratively because we did not pool.

5.5 Analysis plan

Pre-specified: random-effects REML; Hedges g with 95% confidence interval; $I2$, tau-squared, 95% prediction interval; Egger's test and trim-and-fill; leave-one-out; moderator tests for frequency, valence and feedback; GRADE for certainty.

Fallback, applied. Fewer than five eligible trials. No pooled *g*. No forest plot of a single number. No Egger test on two points. Narrative synthesis by design feature. GRADE applied to the body of narrative evidence.

We distinguish measurement reactivity from initial elevation bias in every results paragraph that could be misread as "scores fell, therefore tracking helped."

06 Results

6.1 Study-characteristics table

Study	n	Design	Country	Population	Dose	Channel	Feedback	Prompt valence	Primary mood/symptom finding
Franzen & Lenferink 2025	184	RCT vs waitlist	Netherlands	Bereaved adults, 3--6 months	5x/day x 14 days	ESM app	None	Symptom-focused grief	No group-level change in depression, PTSD or grief vs waitlist
Suen et al. 2022	124	3-arm RCT	Hong Kong	Women 18--64, mild--moderate DASS	Multiple/day x 6 weeks	Digital ESM	None in ESM-alone arm	Mixed momentary	ESM-alone ~ control on DASS-21; feedback arm better than control
Bakker et al. 2018	226	4-arm RCT vs waitlist	Australia	Community adults	Daily x 30 days	Smartphone app	Mild (graphs + links)	Mixed	Diary app raised wellbeing vs waitlist; did not reliably cut depression (CBT-style apps did)

Study	n	Design	Country	Population	Dose	Channel	Feedback	Prompt valence	Primary mood/symptom finding
Conner & Reid 2012	162	RCT of frequency	New Zealand	Students	1 vs 3 vs 6x/day x 13 days	Text message	None	Positive-only (happiness)	No main effect. High depression/neuroticism: more frequent logging, less happiness
De Vuyst et al. 2019	90	RCT of valence	Belgium	Students	10x/day x 7 days	Smartphone	None	Positive vs negative vs bodily	No change in momentary emotion trajectories. PHQ-9 dipped then returned in all arms
Eisele et al. 2023	151	RCT of dose	Belgium	Students	3/6/9x x 30/60 items x 14 days	Smartphone ESM	None	Mixed	Small drop in positive affect over first days. Frequency not a consistent mean-level moderator
Ottenstein, Hasselhorn & Lischetzke 2024	313	RCT of frequency	Germany	Students	3 vs 9x/day x 7 days	Ambulatory app	None	Mixed mood + clarity	Pleasant--unpleasant mood: no trend. Clarity rose 0.28 points on a 1--7 scale over 7 days; 65% rose, 35% fell. Frequency did not moderate
Johnson et al. 2009	280+	Controlled ESM	Mixed	Clinical + non-clinical	~7 days	ESM	None	Mixed	No change in depressed mood over time
Heron & Smyth 2013	63 + 131	Pre--post EMA	USA	Women, non-clinical and at-risk	5x/day x 1--2 weeks	Handheld	None	Body-image + mood	No systematic change in mood
Kramer et al. 2014 (adjacent)	102	3-arm RCT	Netherlands	Depressed outpatients on medication	3 days/week x 6 weeks	Palmtop ESM	Weekly PA feedback in one arm	Mixed + PA focus	Feedback arm beat control on clinician-rated depression; no-feedback effects did not persist
Widdershoven et al. 2019 (adjacent)	79	Controlled ESM	Netherlands	Major depression	10x/day, 3 days/week x 6 weeks	ESM	None	Specific emotions	Negative emotion differentiation improved vs no-ESM. Not a mood-level trial
Hilt et al. 2025 (adjacent)	88 + 42	RCT + later assessment-only	USA	First-year students	3x/day x 3 weeks	App	None in monitoring-only	Mixed	Both monitoring arms improved; non-randomised assessment-only did not

Unverified for outcomes and excluded from effect claims: van Ballegooijen et al. (2016) MoodMonitor protocol. Three arms were planned (mood EMA, energy EMA, questionnaire-only) in adults with mild-to-moderate depressive symptoms. A results paper for the primary reactivity contrast was not found.

No CTRI-registered or published Indian RCT met eligibility.

6.2 Forest-plot data table

A forest plot of Hedges g is not presented. Two eligible contrasts do not support a random-effects model, an I^2 or a prediction interval. Publishing a pooled number from two heterogeneous trials would be theatre.

Study	Contrast	Outcome	Direction	Approximate standardised signal	Weight if pooled
Franzen & Lenferink 2025	ESM vs waitlist	Depression, PTSD, PGD	Null	Compatible with g ~ 0 at group level	---
Suen et al. 2022	ESM-alone vs control	DASS-21 total	Null between groups	ESM-alone and control declined in parallel	---
Bakker et al. 2018 (near)	Mood-diary app vs waitlist	Depression	Null-to-small; wellbeing up	Depression effect not reliable for the diary-only app	---

Where authors reported test statistics rather than *g*, we did not back-calculate a pooled number. Direction is what the evidence will carry.

6.3 Moderator table

Moderator	Pre-specified question	What the studies show	Certainty
Logging frequency	Does more often mean more change?	Conner & Reid: 6x/day happiness reports interacted with high neuroticism/depression. Eisele: 3 vs 6 vs 9 did not consistently move mean affect. Ottenstein 2024: 3 vs 9 did not moderate mood or the clarity rise.	Low. One harmful interaction, two null frequency tests.
Valence of prompts	Neutral vs positive-only	De Vuyst: positive-only vs negative-only vs bodily states --- no differential effect on emotion trajectories or PHQ-9. Conner & Reid: positive-only happiness items, harm in the already-distressed.	Low. Positive-only is the cautionary design, not a proven booster.
Feedback shown vs not	Does a graph or a note change the result?	Suen: feedback arm beat control; monitoring-alone did not. Kramer (clinical): feedback arm produced the lasting clinician-rated change. Bakker: apps that asked users to do something with the data beat a diary-plus-links app on depression.	Moderate for the qualitative claim. Feedback is not "monitoring."

6.4 Heterogeneity

Qualitative heterogeneity is high. Samples are students, bereaved adults, community women, and (in adjacent papers) outpatients. Doses run from 7 days to 6 weeks. Prompts ask about happiness, mixed affect, grief reactions or specific emotion words. Comparators are waitlist, questionnaire-only, or another monitoring schedule. A single I^2 would have been a costume.

6.5 Publication bias

With two eligible trials, Egger's test is not interpretable. One design that was built to answer this exact question --- MoodMonitor (van Ballegooijen et al., 2016) --- has a protocol and no locatable primary-results paper. That is itself a small-study-effect warning. Null results are publishable. This one appears unpublished. We treat that as a gap, not as a secret *g*.

6.6 Risk-of-bias summary

Domain	Typical judgement in the primary set	Comment
Randomisation	Low in Suen, Franzen, Conner, De Vuyst, Eisele, Ottenstein, Bakker	Sequence and allocation generally described.
Deviations from intended exposure	Some concerns	You cannot blind a person to the fact that their phone is asking how they feel. Waitlist arms know they are waiting.
Missing data	Some concerns	EMA compliance is never 100%. Papers vary in how they handle incomplete days.
Outcome measurement	Some concerns for self-report mood	The outcome is the measure that may be reactive. Clinician-rated scales appear only in adjacent clinical trials.
Selective reporting	Some concerns	MoodMonitor outcomes unverified. Several frequency trials were powered for other primary questions.

Overall RoB 2 flavour for the two true no-monitoring contrasts: some concerns, driven by unblindable exposure and self-report outcomes. That is structural, not sloppy.

6.7 GRADE table

Outcome	Studies	Direction	GRADE	Why this grade
Mean mood / affect / depressive symptoms after monitoring-only vs no-monitoring, mixed or non-clinical adults	2 RCTs + 1 near-eligible RCT	Near-null	Low	Few trials; indirect samples (bereavement, women-only, community apps with extra links); unblindable exposure; unpublished pivotal protocol.
Harm from high-frequency positive-only prompts in people with high neuroticism or depressive symptoms	1 RCT	Possible harm	Very low	Single student sample; not replicated in De Vuyst's valence experiment.
Rise in emotional clarity / differentiation after repeated emotion questions	2 controlled comparisons (Ottenstein et al. 2024; Widdershoven 2019 clinical)	Small rise in skill-like constructs	Low	Different constructs, one clinical, no common metric.
Added benefit of feedback on top of logging	2 RCTs (Suen; Kramer clinical)	Feedback > logging alone	Low-moderate	Consistent direction, different populations.

6.8 Narrative synthesis

The thing tracked usually does not move. Franzen and Lenferink asked 184 adults, newly bereaved, to report prolonged-grief reactions five times a day for a fortnight. Waitlist controls did not. At the group level, depression, post-traumatic stress and grief severity did not differ. Most individuals did not show a clinically relevant pre--post change. People who started with higher grief scores were more likely to improve --- a pattern that could be regression to the mean, initial elevation bias, or a real benefit concentrated at the top of the scale. The trial cannot separate those three. What it can say is that the average bereaved adult was not made worse, and not made better, by the log.

Suen and colleagues randomised 124 Hong Kong women with mild-to-moderate depression and anxiety scores to experience sampling plus personalised feedback, experience sampling alone, or no extra programme. Six weeks later the monitoring-alone arm and the control arm had declined in parallel on the combined depression-anxiety-stress score. The feedback arm declined more than control. That is the cleanest adult contrast we have for the product question. The thermometer did not beat time. The thermometer plus a note did.

Bakker and colleagues compared three commercial-style apps with waitlist in 226 Australian adults. The app that was closest to a mood diary raised wellbeing against waitlist. It did not reliably reduce depression. The apps that added cognitive-behavioural drills did. Again: the log is not the workout.

Frequency is not a simple dose-response. Conner and Reid texted 162 New Zealand students and asked "how happy are you right now?" one, three or six times a day for 13 days. Averaged across the sample, frequency did not matter. Split by disposition, it did. Students already high in depressive symptoms or neuroticism lost happiness as the texts piled up. Students low on those traits gained a little. That is not a reason to abandon logging. It is a reason not to pester a distressed person with a happiness prompt six times between breakfast and bed.

Eisele and colleagues and Ottenstein, Hasselhorn and Lischetzke (2024) both manipulated how often students were pinged. Mean pleasant--unpleasant mood did not follow the ping rate. In the 2024 study, momentary emotional clarity rose on average by 0.28 points on a 1-to-7 scale across the week --- about 0.05 a day --- in 65% of participants, and fell in 35%. The rise was the same whether people were asked three times a day or nine. Three times was enough to grow the vocabulary. Nine times did not grow the mood.

Valence of the prompt is not a free lunch. De Vuyst and colleagues sent students ten prompts a day for a week: only positive emotion words, only negative emotion words, or non-emotional bodily states. Momentary emotion trajectories did not diverge. Depressive-symptom scores dipped after the week and had returned a month later in every arm, including the bodily-state arm. If there was an effect, it was the ritual of being in a study, not the flavour of the adjective. That finding sits beside Conner and Reid rather than on top of it. Positive-only *happiness* asked very often in people who are already raw is the caution. Positive-only *as a class* is not a proven antidepressant.

Constructs that are not mood do move. Repeatedly naming feelings is a skills practice even when it is sold as a measure. Widdershoven and colleagues found that six weeks of experience sampling improved negative-emotion differentiation in people with depression compared with no experience sampling. The 2024 ambulatory study found a within-person rise in momentary emotional clarity. Kauer and colleagues, outside our adult-only frame, found that adding mood and stress items to an activity log raised emotional self-awareness in 14-to-24-year-olds. Picture a sales manager who used to type only "stressed". A few weeks of prompts later she can tell "irritated" from "let down". Her average mood score has not moved. People got better at noticing. They did not automatically get better at feeling. In machine-learning language: the feature extractor improved. The loss function did not.

Initial elevation bias is the impostor in the room. Shrout and colleagues ran four field experiments. First-day ratings of anxiety and physical symptoms sat higher than later ratings, with standardised effects often in the 0.2 to 0.5 band and larger for negative inner states than for positive states or for behaviour. Late starters still showed the spike. Roommates' ratings showed it too. That pattern is first-report inflation, not a sign that anyone got better. Cerino and colleagues later coordinated several ecological-momentary-assessment datasets and found little consistent initial-elevation bias in *momentary* affect (combined d from -0.05 to 0.14). The practical rule for a programme is conservative. If your dashboard shows a beautiful drop from day 1 to day 8, do not print a case study until you have a no-monitoring comparison or, at minimum, you have discarded the first day.

Adjacent clinical programmes do not rescue the claim. Astill Wright and colleagues pooled eight randomised trials of mood monitoring in depression and bipolar disorder (n = 1,230). Mania and bipolar depression effects hugged zero. Unipolar depression showed a small, fragile signal at 12 months that was not there at 6 months. Kramer's three-arm trial is the mechanism exhibit: adding weekly feedback on personalised positive-affect patterns produced a clinically meaningful drop in clinician-rated depression against treatment as usual; sampling without the feedback did not leave a lasting mark. Monitoring in a clinic, wrapped in care, is not the same exposure as a WhatsApp slider at 9.14 p.m.

Harms are uncommon at the group level and not absent as stories. A 2025 systematic review of adverse events in mood monitoring and ambulatory assessment for depression and bipolar disorder estimated that about two in a hundred participants reported subjective mood

worsening among the few studies that asked. Qualitative syntheses of user experience find a minority who say the prompts made the weather inside the head louder. Kivelae and colleagues (2024), in a suicidal-ideation sample outside our primary scope, found no systematic drift in momentary affect or ideation across the ecological-momentary-assessment period, while a subgroup later said the protocol had been hard. Group means can be quiet while a tail is not.

07 Discussion

PULL QUOTE

Six happiness pings a day were the one design that looked able to take a bite out of people already raw.

What the number means

There is no number. That is the finding. The industry wanted a *g*. The trials would not give one. Two honest no-monitoring contrasts, a near-eligible diary-app trial, and a cluster of frequency and valence experiments all point at the same unfashionable sentence: **logging mood, by itself, does not reliably improve or worsen mean mood.**

A stand-up comic would stop there and take a sip of water. The room would wait for the twist. The twist is not that tracking is magic. The twist is that everyone sold the thermometer as a gym.

Good coaches already knew this. Take a team lead in Gurugram after a bruising review. Circling "low" on a slider changes nothing. Writing down the situation, the thought ("they think I am useless"), the evidence, a fairer alternative and one next act --- an email to her manager before ten tomorrow --- changes what she does next. Naming a feeling is a start; it comes paired with taking part. Seeing a thought as just a thought is not a mood score either. It is a way of holding a sentence so that the sentence does not hold you. Entrepreneurs who have shipped a real product already knew. A dashboard that nobody uses to decide the next experiment is interior decoration.

Think of the lift mirror in a Lower Parel office tower. You check your collar on the way up to the meeting, step out, and forget what you saw. A log that is not followed by a rep is that mirror.

Machine learning makes the same point. In a training loop you log the loss so that you can update the weights. Logging the loss without the update is just a CSV file getting fatter. Mood logging without a next action is a CSV file of feelings. Some people will feel seen. Seeing is not the same as changing.

None of this is an argument against the daily rep. It is an argument for telling the truth about what the rep is. In the two trials that had a no-monitoring arm, the log was safe at the group level. That matters. Workplace programmes in India will be asked, correctly, whether asking about low mood every evening can itself darken the evening. The best available adult trials say: not on average, not over two to six weeks, not with these prompts. They also say: do not expect the average to rise.

The prompt-design wrinkle is the only place a product team should lean forward. High-frequency, positive-only "how happy are you?" messages were the one design that looked able to take a bite out of people who were already raw. Valence-neutral or mixed prompts did

not reproduce that bite. Feedback --- a human or structured note that turns the log into a next step --- was the component that beat control. That is not reactivity. That is coaching.

Limitations

We did not register a protocol. We did not run dual independent screening of every export. Some full texts sit behind publisher walls and were read from abstracts, PMC copies and author pages; where a number could not be verified it was not used. The eligible set is tiny. Samples are WEIRD or partly WEIRD: Dutch bereavement, Hong Kong women, Belgian and German and New Zealand students, Australian app users. Delivery languages are European or East Asian. Dose is short. Outcomes are almost all self-report. The trial purpose-built for this question did not yield a results paper we could verify. Adjacent clinical evidence is richer and answers a different question. Certainty is low. A later review with five clean trials should pool. This one must not.

WHAT THIS MEANS

What this means for an Indian workplace programme

Picture a Mumbai product team at the whiteboard, deciding what the daily check-in is for. They do not need a pooled *g* to decide well. At the top of the board they write: a rep that builds noticing, not a medicine that lifts mood. They cross out "How happy are you?" and keep "How is your mood right now?" --- short and valence-neutral. Nobody gets pinged six times a day unless they asked for that density. At the end of the week each person sees their own week, then gets one next act --- a values step, a thought check, a two-minute wind-down, a human coach if the score stays low. Sales asks for a before-after chart that starts on day 1 and ends on day 8. Before anyone calls it impact, the analyst throws away the first day or runs a no-logging comparison. The app speaks 23 languages and asks the same scientific question in each of them. India has not yet run the trial. Until it does, import the humility, not the hype. DPDP-compliant storage of a slider is easy. Honest copy about what the slider does is the harder ISO.

Research gaps, including the India gap

The India gap is complete. CTRI and the published record did not yield a randomised comparison of mood-logging-only versus no-monitoring in Indian working adults. MITHRA, a 2025 pilot in women's self-help groups, bundled screening with a behavioural-activation programme. That is care. It is not this contrast. India has the sample size, the language surface, the WhatsApp channel and the regulatory frame. It does not have the trial.

Other gaps sit beside it. MoodMonitor needs a results paper or a declared null. Future trials should use four-group designs --- logging or not, programme or not --- so that measurement and programme can be pulled apart. They should report individual-level change, not only group means, because the tail is where harm and help both hide. They should pre-specify valence, frequency and feedback as factors, not as afterthoughts. They should measure skills (clarity, differentiation, values-consistent action) and mood as separate outcomes. They should last longer than a student fortnight. They should include shift workers, caregivers and people who take the prompt in Hindi, Tamil, Bengali or Marathi on a second-hand Android. Until those trials exist, anyone who quotes a single reactivity *g* for mood logging is quoting a number this field has not earned.

08 FAQ

Q1 Does logging my mood every day make me feel better?

Not on average. Two randomised adult comparisons of logging-only versus no logging found no group-level lift in mood or symptoms. A diary-style app raised wellbeing against waitlist in one trial but did not reliably cut depression. Treat the log as a noticing rep, not as the programme.

Q2 Can logging my mood make me feel worse?

Uncommon at the group level. About 2 in 100 people in clinical monitoring reviews reported subjective mood worsening when researchers asked. One student trial found that six daily "how happy?" texts lowered happiness in people already high on depressive symptoms or neuroticism. Valence-neutral prompts did not show that pattern.

Q3 Is a drop from day 1 to day 8 proof that tracking works?

No. That drop is the signature of initial elevation bias: first reports of negative inner states often start high and then settle, even when the person's life has not changed. Discard day 1, or compare with people who never logged.

Q4 How often should an Indian professional check in?

Three times a day was enough to raise emotional clarity in one experiment; nine times did not raise mood. Daily or twice-daily valence-neutral check-ins are a reasonable default. Six happiness pings a day are not.

Q5 If the log does not change mood, why do it?

Because noticing is a skill, and skills need reps. Clarity and emotion differentiation can rise when mean mood does not. Pair the log with one next act and you have a programme. Leave the log alone and you have a spreadsheet.

09

References

Astill Wright, L., et al. (2026). Mood monitoring, mood tracking, and ambulatory assessment interventions in depression and bipolar disorder: A systematic review and meta-analysis of randomised controlled trials. *JMIR Mental Health, 13*, e84020. https://doi.org/10.2196/84020

Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2018). A randomized controlled trial of three smartphone apps for enhancing public mental health. *Behaviour Research and Therapy, 109*, 75--83. https://doi.org/10.1016/j.brat.2018.08.003

Cerino, E. S., et al. (2022). Little evidence for consistent initial elevation bias in self-reported momentary affect: A coordinated analysis of ecological momentary assessment studies. *Psychological Assessment, 34*(5), 467--482. https://doi.org/10.1037/pas0001108

Conner, T. S., & Reid, K. A. (2012). Effects of intensive mobile happiness reporting in daily life. *Social Psychological and Personality Science, 3*(3), 315--323. https://doi.org/10.1177/1948550611419677

De Vuyst, H.-J., Dejonckheere, E., Van der Gucht, K., & Kuppens, P. (2019). Does repeatedly reporting positive or negative emotions in daily life have an impact on the level of emotional experiences and depressive symptoms over time? *PLoS ONE, 14*(6), e0219121. https://doi.org/10.1371/journal.pone.0219121

Eisele, G., et al. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment, 35*(1), 68--81. https://doi.org/10.1037/pas0001177

Franzen, M., & Lenferink, L. I. M. (2025). Exploring reactivity effects of self-monitoring prolonged grief reactions in daily life: A randomised waitlist-controlled trial using experience sampling methodology. *Internet Interventions, 42*, 100877. https://doi.org/10.1016/j.invent.2025.100877

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology, 15*(3), 453--468. https://doi.org/10.1348/135910710X492341

French, D. P., et al. (2021). Reducing bias in trials due to reactions to measurement: Experts produced recommendations informed by evidence (MERIT). *Journal of Clinical Epidemiology, 140*, 130--139. https://doi.org/10.1016/j.jclinepi.2021.06.028

Heron, K. E., & Smyth, J. M. (2013). Is intensive measurement of body image reactive? A two-study evaluation using ecological momentary assessment suggests not. *Body Image, 10*(1), 35--44. https://doi.org/10.1016/j.bodyim.2012.08.006

Hilt, L. M., et al. (2025). Brief app-based mood monitoring and mindfulness intervention for first-year college students: A randomised controlled trial. *Journal of Psychopathology and Behavioral Assessment, 47*, 23. https://doi.org/10.1007/s10862-025-10196-x

Johnson, E. I., Grondin, O., Barrault, M., Faytout, M., Helbig, S., Husky, M., Granholm, E. L., Swendsen, J., et al. (2009). Computerized ambulatory monitoring in psychiatry: A multi-site collaborative study of acceptability, compliance, and reactivity. *International Journal of Methods in Psychiatric Research, 18*(1), 48--57. https://doi.org/10.1002/mpr.276

Kauer, S. D., Reid, S. C., Crooke, A. H. D., Khor, A., Hearps, S. J. C., Jorm, A. F., Sanci, L., & Patton, G. (2012). Self-monitoring using mobile phones in the early stages of adolescent depression: Randomized controlled trial. *Journal of Medical Internet Research, 14*(3), e67. https://doi.org/10.2196/jmir.1858

Kivelae, L. M. M., Fiss, F., van der Does, W., & Antypa, N. (2024). Examination of acceptability, feasibility, and iatrogenic effects of ecological momentary assessment of suicidal ideation. *Assessment*. https://doi.org/10.1177/10731911231216053

Koenig, L. M., Allmeta, A., Christlein, N., Van Emmenis, M., & Sutton, S. (2022). A systematic review and meta-analysis of studies of reactivity to digital in-the-moment measurement of health behaviour. *Health Psychology Review, 16*(4), 551--575. https://doi.org/10.1080/17437199.2022.2047096

Kramer, I., et al. (2014). A therapeutic application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry, 13*(1), 68--77. https://doi.org/10.1002/wps.20090

Ottenstein, C., Hasselhorn, K., & Lischetzke, T. (2024). Measurement reactivity in ambulatory assessment: Increase in emotional clarity over time independent of sampling frequency. *Behavior Research Methods, 56*(6), 6150--6164. https://doi.org/10.3758/s13428-024-02346-y

Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavel, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences, 115*(1), E15--E23. https://doi.org/10.1073/pnas.1712277115

Suen, Y. N., Yau, Y. J., Wong, P. S., Li, Y. K., Hui, C. L. M., Chan, K. W., Lee, H. M. E., Chang, W. C., & Chen, E. Y. H. (2022). Effect of brief, personalized feedback derived from momentary data on the mental health of women with risk of common mental disorders in Hong Kong: A randomized clinical trial. *Psychiatry Research, 317*, 114880. https://doi.org/10.1016/j.psychres.2022.114880

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D. D., Smit, J. H., & Riper, H. (2016). Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): Protocol of a randomised controlled trial. *BMC Psychiatry, 16*, 359. https://doi.org/10.1186/s12888-016-1065-5

Widdershoven, R. L. A., Wichers, M., Kuppens, P., Hartmann, J. A., Menne-Lothmann, C., Simons, C. J. P., & Bastiaansen, J. A. (2019). Effect of self-monitoring through experience sampling on emotion differentiation in depression. *Journal of Affective Disorders, 244*, 71--77. https://doi.org/10.1016/j.jad.2018.10.092

About the author

ABOUT THE AUTHOR

Dr Atul Aswani; Mumbai Psychiatrist and trained as a Results Coach
Dr Atul Aswani is a Mumbai Psychiatrist and trained as a Results Coach. He leads research and coaching design at EmotionApp, a non-clinical emotional-fitness platform for Indian professionals. The work sits under Tranquilo Meditek Sytems Private Limited, Mumbai, and is built to ISO 9001 and ISO 13485, with India-resident data and DPDP-aligned practice.

HOW TO CITE

Aswani A. Does Tracking Change the Thing Tracked?: A Meta-Analysis of Measurement Reactivity in Mood Logging. EmotionApp Research Paper 3. Mumbai: Tranquilo Meditek Sytems Private Limited; 2026. https://emotionapp.in/research/measurement-reactivity-mood-logging

LAST UPDATED

Last updated: 22 September 2026

NOTE

EmotionApp is a non-clinical emotional-fitness coaching platform. This paper describes practices, techniques and coaching. It is not therapy, treatment, diagnosis or medical advice.

SAFETY NOTE

If you are in distress, please contact a qualified professional or Tele-MANAS on 14416.