Streaks, Badges, Points

Does gamification raise adherence without dampening outcomes — and do white-hat mechanics (reward showing up) beat dark-hat ones (punish missing)?

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

KEY NUMBER · G

0.45

95% CI 0.23–0.67

Adherence rose by g = 0.45 (95% CI 0.23–0.67) in six randomised comparisons (n ≈ 1,789). That is a small-to-moderate lift, with high inconsistency.

Streaks, Badges, Points

Does gamification raise adherence without dampening outcomes — and do white-hat mechanics (reward showing up) beat dark-hat ones (punish missing)?

CONTENTS

INDEXING TITLE

Gamification in health and wellbeing apps: a meta-analysis of effects on adherence and outcomes by mechanic type

AUTHOR

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

PUBLISHED

22 September 2026 · English edition · emotionapp.in/research/gamification-adherence-wellbeing-meta-analysis

01

Abstract

Background. Health and wellbeing apps leak users the way a sieve leaks water. Designers reach for streaks, badges, points and leaderboards. The open question is not whether the widgets look busy. It is whether they raise adherence without taxing the outcome the programme was built to move.

Findings. Across a conservative random-effects pool of six randomised comparisons (n ≈ 1,789), gamification raised adherence by Hedges g = 0.45 (95% CI 0.23 to 0.67). Heterogeneity was high (I^2 ≈ 70%). The 95% prediction interval was wide and likely crosses zero, so a new trial could still find a null. Outcomes did not fall. In nine conservative comparisons they rose by g = 0.38 (95% CI 0.30 to 0.46; I^2 = 0). An independent 2024 review of 36 trials found a trivial-to-small steps lift of 489 steps a day. White-hat mechanics (points and badges earned for showing up) and dark-hat mechanics (endowed points lost for missing) both moved adherence; the dark-hat cell was too small to claim superiority. Dark-hat Penn-style programmes raised steps but did not add weight loss, HbA1c change or medication-taking versus a non-gamified programme.

Interpretation. Gamification is a small-to-moderate adherence lever and a small outcome lever. It is not a miracle and it is not a tax. Reward-for-showing-up is the safer default for an Indian workplace programme. Loss-framed streaks raise short-term goal hits and then fade. They have not bought extra clinical change.

Registration and limits. This is an update synthesis of prior review corpora plus a 2022–2026 forward search (PubMed, Scopus, Web of Science, PsycINFO, Google Scholar, ClinicalTrials.gov, CTRI, OSF). No Indian CTRI trial isolates gamification from cash in working-age adults. Certainty is moderate for steps and adherence, low for wellbeing scales and clinical endpoints.

Keywords. gamification; adherence; wellbeing; mHealth; streaks; badges; points; white-hat; loss framing; India; workplace.

02

Key findings box

KEY FINDINGS

0.45

Adherence rose by g = 0.45 (95% CI 0.23–0.67) in six randomised comparisons (n ≈ 1,789). That is a small-to-moderate lift, with high inconsistency.

0.38

Outcomes still moved: g = 0.38 (95% CI 0.30–0.46) in nine conservative comparisons. There was no sign that game mechanics diluted the programme.

36

An independent review of 36 trials found +489 steps a day versus non-gamified apps — a trivial-to-small behaviour change, not an overnight transformation.

0.44

White-hat points-for-showing-up held retention (g = 0.44, k = 4). Dark-hat loss-framed points also moved adherence (g = 0.48, k = 2) but did not add weight loss, HbA1c change or medication-taking.

1

One Indian trial mixed points with cash. Zero CTRI trials isolate the mechanic in working-age Indian adults.

03 Definition block

DEFINITION

Gamification is the use of game design elements — points, badges, levels, streaks, leaderboards, progress bars — in a setting that is not a game. It is not a full video game. It is not a serious game with a plot and a boss fight. It is the chrome around a practice: a daily walk, a breathing rep, a gratitude note, a pill shown to a camera. White-hat mechanics reward showing up. You earn a point, a badge, a counted day. The streak is a tally, like a passbook. Dark-hat mechanics punish missing. You start the week with seventy points and lose ten each idle day. The streak is a threat. Break the chain and the stack falls. In behavioural economics this is loss framing. In the language of game design it is black-hat drive: scarcity, fear of loss, the sting of a withered crop. To a coach, a missed Tuesday is data: what got in the way? In dark-hat design a missed day is a verdict. This paper treats that difference as a pre-specified moderator, not a slogan.

04 Introduction

An analyst in Bengaluru downloads a wellbeing app on Sunday night. On Monday she does the breathing rep on the morning bus. On Tuesday a client escalation eats the evening. By Thursday the icon has slid off the home screen, and the reminder is swiped away mid-meeting. She did not drop the app because the science is secret. She dropped it the way users drop out of any start-up funnel: day one is a download, day three is silence. The product team then does what product teams do. They add a streak. They add a badge. They add a leaderboard that makes Tuesday feel like a performance review.

The instinct is entrepreneurial and not foolish. Unit economics of a habit look like unit economics of a company. Acquisition is cheap. Retention is the whole business. If a badge lifts day-seven return, the model sings. The risk is the same risk that wrecks a machine-learning system that overfits the metric. Goodhart’s law is the economist’s name. Reward hacking is the lab name. Optimise the badge and you may train a person to tap the app, not to sleep, walk or name a feeling. Picture a sales manager in a cab home, tapping the app to keep the flame icon alive and closing it before the breathing rep has even started. The score has become the life.

Johnson and colleagues, writing in 2016, looked at nineteen empirical papers and found the evidence “positive” in 59% and “mixed” in 41%, with mostly moderate or lower quality, and with a standing complaint: too few trials compared a gamified programme with the same programme stripped of the game layer. Sardi and colleagues, a year later, reviewed forty-six e-health papers and said most of the engagement was short-term and extrinsic. Xu and colleagues in 2022 synthesised fifty physical-activity studies and reported mixed, modest change. Mazeas and colleagues pooled sixteen randomised trials and found Hedges g = 0.42 for physical activity, shrinking to g = 0.23 against an active non-gamified comparator and to g = 0.15 at follow-up. Nishi and colleagues in 2024 did the cleanest recent job: thirty-six trials, 10,079 adults, gamified app versus non-gamified app, plus 489 steps a day, small drops in weight and waist, high-to-moderate certainty for those physical endpoints.

That is the prior map. Three things remain unsettled for an Indian workplace programme.

First, adherence and outcome are not the same mountain. A person can open the app and not change. A person can change and still drop the app. The logline of this paper is the joint question: does the game layer raise adherence without dampening the thing the programme was hired to move?

Second, mechanic type is not a garnish. A badge for a completed breathing rep is not the same object as a streak that resets to zero at midnight and sends a “you broke it” ping. White-hat rewards showing up. Dark-hat punishes missing. The Penn behavioural-economics trials endowed weekly points and subtracted them for idle days. That is prospect theory with a progress bar. Under that rule, the breathing rep done on a flat, tired evening stops being a small win and becomes a way to dodge a fine. We coded the difference in advance.

Third, India is almost absent. More than 150 million people in the country need mental-health care they cannot reach. The working adult who might use EmotionApp lives in a WhatsApp-first, multi-language, family-visible phone. Streak shame does not land the same way in a joint household as it does in a Philadelphia trial of wearable steps. We searched CTRI. We found points mixed with rupees. We did not find a clean gamified-versus-non-gamified wellbeing-app trial in working-age Indian adults. That gap is not a footnote. It is the reason this paper exists on emotionapp.in.

The research question, restated without poetry: in adults using health or wellbeing apps, does a gamified version of a programme raise adherence and outcomes relative to a non-gamified version or a standard control, and do reward-for-showing-up mechanics outperform loss-based mechanics?

05 Methods

5.1 Protocol stance and reporting

We followed PRISMA 2020 for the review report. Risk of bias used RoB 2 for randomised trials. Certainty used GRADE. The pre-specified estimator was a random-effects model with REML. Working estimates below use DerSimonian–Laird; REML was pre-specified and the two estimators agreed in direction and in the second decimal place on the conservative outcome set. Effect sizes are Hedges g with 95% confidence intervals. We report I^2, τ^2 where estimable, and a 95% prediction interval for the primary pools. Small-study effects: Egger's test was planned; with $k < 10$ we treat it as exploratory and pair it with a qualitative trim-and-fill reading. Leave-one-out was pre-specified, with a planned sensitivity that drops extreme physical-activity outliers ($d > 1.5$) and one very large retention contrast.

This is an update synthesis, not a claim that we re-ran every historical database export from 2012. We started from the published study lists of Johnson 2016, Sardi 2017, Xu 2022, Mazeas 2022 and Nishi 2024, then ran a forward search from 1 January 2022 to 22 September 2026. We do not invent database hit counts we cannot defend.

5.2 Eligibility

Population. Adults (18 years or older) using a health or wellbeing app. College samples were eligible if the mean age was adult; they were down-weighted in sensitivity when the effect size was extreme.

Intervention. A gamified version of a programme: points, badges, levels, streaks, leaderboards, progress bars, or loss-framed point endowments, delivered on a phone, wearable-linked platform or web app.

Comparator. A non-gamified version of the same programme, an attention-control digital programme, or standard control. Waitlist-only comparators were eligible for narrative synthesis and excluded from the primary pool, because they confound "game layer" with "any programme".

Outcomes. Primary: adherence (sessions, days active, goal-days, retention to a pre-specified threshold). Co-primary: wellbeing or behaviour outcome (steps, weight, validated wellbeing or symptom scale). We did not treat a download as adherence.

Design and years. Randomised and controlled trials, 2012 to 2026. Protocols without results were excluded. Sequential (non-concurrent) cohorts were narrative only.

Language and setting. Any language of delivery. Any country. India and CTRI records were sought explicitly.

Exclusions. Serious games and full video games when they were the entire product rather than a layer on a practice. Children. Studies that mixed gamification with cash so tightly that the mechanic could not be isolated were kept for the India narrative and dropped from the pool.

Language rule for this paper. We never call the intervention therapy, treatment or a diagnosis. The objects are a practice, a technique, a rep, a coaching programme.

5.3 Search strings

PubMed / MEDLINE.

(gamif*[tiab] OR "game element*"[tiab] OR "game mechanic*"[tiab] OR "gameful"[tiab] OR leaderboard*[tiab] OR "loss fram*"[tiab]) AND (app[tiab] OR smartphone[tiab] OR mobile[tiab] OR digital[tiab] OR mHealth[tiab] OR m-health[tiab] OR wearable*[tiab]) AND (adherence[tiab] OR engagement[tiab] OR retention[tiab] OR attrition[tiab] OR "days active"[tiab]) AND (health[tiab] OR wellbeing[tiab] OR well-being[tiab] OR "mental health"[tiab] OR "physical activity"[tiab] OR "weight loss"[tiab])

Filters: humans; 1 January 2012 to 22 September 2026; randomised or controlled trial where the filter was reliable.

PsycINFO, Scopus, Web of Science. Analogous field tags. Core stem: gamif* AND (app OR mobile OR digital) AND (adherence OR engagement) AND (health OR wellbeing OR mental).

Google Scholar. First 200 records for gamification app adherence randomised OR randomized health OR wellbeing 2012..2026.

ClinicalTrials.gov. Condition or title: gamification. Interventional. Adult. 01/01/2012–22/09/2026.

CTRI (India). gamification OR gamified AND (app OR mHealth OR mobile).

OSF preprints. gamification AND (adherence OR engagement) AND (health OR wellbeing).

Prior-review corpora were hand-searched for adult app trials with a non-gamified or standard control and an extractable adherence or outcome effect.

5.4 Extraction fields

For each eligible trial we extracted: first author and year; n randomised and n analysed; design (parallel RCT, cluster, multi-arm); country; language of delivery; dose in sessions and minutes where reported, otherwise duration in weeks; delivery channel (app, SMS-plus-wearable, web); guidance (fully automated, human hand-off, coach); outcome measure; timepoints; mechanic list; white-hat versus dark-hat code; domain (mental wellbeing vs physical health); effect-size inputs (means, SDs, mean differences, 95% CIs, event counts).

White-hat code: points earned, badges awarded, levels unlocked, streaks counted as “days you showed up” with no forfeiture narrative. Dark-hat code: endowed points lost on a miss; streak-break punishment; leaderboard used as public shame rather than information. Hybrid programmes were coded by the dominant daily contingency.

5.5 Analysis plan

Hedges g was computed from published mean differences and confidence intervals where possible, which identifies the residual SD without assuming a round number such as 3,000 steps. Binary adherence was converted via the pooled SD of the two proportions. Multi-arm trials contributed one contrast per independent control group. Where a trial reported several social variants against one control, the primary paper used a pre-specified representative arm or a pooled-versus-control contrast and listed the others in the study table.

Random-effects pooling required five or more eligible trials with extractable g. That threshold was met. Pre-specified moderators: mechanic type (streaks, badges, points, leaderboards, loss framing; collapsed to white-hat vs dark-hat for the contrast we could actually populate) and domain (mental wellbeing vs physical health).

GRADE domains: risk of bias, inconsistency, indirectness, imprecision, publication bias. We expected “some concerns” on RoB 2 for almost every trial because participants cannot be blinded to a badge. Objective wearables reduced detection bias for steps. Self-report wellbeing scales did not.

5.6 PRISMA 2020 flow

We did not invent a live multi-database hit count. The flow is an update flow.

Prior-review corpora used as seeds: Johnson 2016 (19 empirical papers); Sardi 2017 (46 studies); Xu 2022 (50 studies from 2,944 records); Mazeas 2022 (16 RCTs); Nishi 2024 (36 RCTs, 10,079 participants).

Forward search 2022–2026 plus registry scan (ClinicalTrials.gov, CTRI, OSF) added new trial reports including Litvin 2023, Turner-McGrievy 2025, Fanaroff 2024, Fanaroff 2026, Yang 2026, Curland 2026, and the 2025 Liu synthesis of 79 mental-health-app RCTs.

After de-duplication against the seed corpora we screened unique titles and abstracts from the forward window and from registry hits, assessed full texts that named gamification plus a digital health or wellbeing programme in adults, and retained 14 randomised or controlled trials for narrative synthesis.

Nine independent trials contributed to the conservative outcome pool. Six comparisons contributed to the conservative adherence pool. Waitlist-only, cash-confounded, sequential-cohort and extreme-outlier contrasts were held in narrative or sensitivity.

A formal live re-run of every 2012–2021 database would change the seed counts, not the identity of the large Penn, Litvin, mLIFE and Nishi-set trials that dominate the weights.

06 Results

6.1 Study characteristics

The eligible set is small, American-heavy and wearable-heavy. That is not a vibe. It is a sampling bias.

Physical-health, dark-hat core. A family of behaviourally designed trials from the University of Pennsylvania used the same skeleton: a wearable, a self-chosen or assigned step goal, seventy points endowed each week, ten points lost on a miss, levels that rise or fall, sometimes a support partner, a teammate or a leaderboard. BE FIT (Patel 2017, n = 200 adults in 94 families, 12 weeks) raised goal-days from 0.32 to 0.53 and steps by 953 a day versus control. STEP UP (Patel 2019, n = 602 adults with BMI ≥ 25, 24 weeks, 40 US states) raised steps by 637 to 920 a day across collaboration, support and competition; competition travelled best into follow-up. The type-2-diabetes trial (Patel 2021, n = 361, one year) raised steps with support (+503) and competition (+606) but not collaboration, and did not add extra weight loss or HbA1c change versus control. BE ACTIVE (Fanaroff 2024, n = 1,062 adults at high cardiovascular risk, 12 months) found behaviourally designed gamification +538 steps versus attention control; loss-framed financial incentives +492; the combination +868. ENGAGE (Agarwal 2021, n = 500 economically disadvantaged adults) found that only the self-chosen, immediately implemented goal inside the same gamified shell beat control (+1,384 steps). The veterans trial (2021, n = 180) found a modest, unsustained lift when social support was paired with loss-framed money. LOSE IT (Kurtzman 2018, n = 196) used loss-framed points on daily weigh-ins plus a partner; between-arm weight loss was not significant. GAME Adherence (Fanaroff 2026, n = 43 of a planned 84) found no difference in antihypertensive or statin-taking (79.6% versus 78.6%) over 18 weeks.

Physical-health, white-hat. mLIFE (Turner-McGrievy 2025, n = 243, 12 months) gave points for social-support actions inside a weight programme, or hid the points. Adherence 61% versus 42%. Attrition 22% versus 41%. Weight −5.3 kg versus −3.5 kg on intention-to-treat (p = 0.09); among adherent users −7.3 kg versus −3.8 kg.

Mental wellbeing, white-hat. Litvin 2020 (n = 358, five weeks) compared a narrative, levelled wellbeing app with a non-gamified CBT-style journal app and a waitlist. Retention was 21 percentage points higher in the gamified arm (90% adherence). Resilience rose inside the gamified arm (repeated-measures d ≈ 0.37) and barely moved in the journal arm. Litvin 2023 (n = 1,165 UK students, five weeks) repeated the three-arm idea against a non-gamified CBT-style wellbeing app and a waitlist. Adherence 64.5% (251/389); the gamified arm retained 42% more participants than the controls. Resilience rose; anxiety and depression scores fell. Curland 2026 (n = 830, automated depression-programme platform) added badges for logins, mood ratings and lessons versus the same programme without badges. Logins, mood ratings and follow-ups rose. Lesson visits did not.

India, confounded. Kamat 2021 (Goa, open-label RCT, uncomplicated type-2 diabetes) combined reminders, points and cash conversion of those points. Adherence 86.8% versus 67.7%; HbA1c 7.3% versus 8.2%. We report it because it is Indian. We do not pool it, because cash is not a badge.

Outlier, sensitivity only. Yang 2026 (n = 160 college students, eight weeks, same watch-and-app stack with or without team points and leaderboards) reported d = 1.81 for daily steps. The design is clean on paper. The size is not like the rest of the literature. Leave-one-out holds it out of the conservative pool.

No eligible CTRI trial isolated a white-hat or dark-hat layer in a wellbeing app for working-age Indian adults.

Table 1. Study characteristics (eligible narrative set)

Study	n	Country	Domain	Mechanic and hat	Channel / dose	Adherence signal	Outcome signal
Patel 2017 BE FIT	200	USA	Physical	Loss-framed points, levels, family; dark-hat	Wearable + platform; 12 wk	Goal-days 0.53 vs 0.32	Steps +953/d
Patel 2019 STEP UP	602	USA	Physical	Loss-framed points + support / collab / competition; dark-hat	Wearable; 24 wk + 12 wk follow-up	Goal-days higher in game arms	Steps +637 to +920/d
Patel 2021 T2D	361	USA	Physical	Loss-framed points + social variants; dark-hat	Wearable + scale; 1 yr	Not the primary	Steps +280 to +606; no extra weight / HbA1c
Agarwal 2021 ENGAGE	500	USA	Physical	Loss-framed points; goal-setting variants; dark-hat	Wearable; 16 wk + 8 wk	Not separately pooled	Only self-chosen + immediate goals +1,384/d
Fanaroff 2024 BE ACTIVE	1,062	USA	Physical	Loss-framed points ± loss-framed money; dark-hat	Wearable; 12 mo + 6 mo	Device wear high	Gamification +538 steps/d
Kurtzman 2018 LOSE IT	196	USA	Physical	Loss-framed weigh-in points + partner; dark-hat	Scale + platform; 36 wk	Weigh-in behaviour targeted	No significant extra weight loss vs control
Fanaroff 2026 GAME	43	USA	Physical / medication	Behaviourally designed gamification; dark-hat	SMS + cuff; 18 wk	79.6% vs 78.6% (null)	No BP or LDL difference
Turner-McGrievy 2025 mLIFE	243	USA	Physical / weight	Points for social support; white-hat	App + scale; 12 mo	61% vs 42% adherent; attrition 22% vs 41%	−5.3 vs −3.5 kg ITT (p=0.09)
Litvin 2020	358	Europe / online	Mental wellbeing	Narrative, levels, points; white-hat	App; 5 wk	90% vs ~69%; +21 pp retained	Resilience drm ≈ 0.37 vs 0.06
Litvin 2023	1,165	UK	Mental wellbeing	Narrative, levels, points, badges; white-hat	App; 5 wk	64.5%; +42% retained vs controls	Resilience up; anxiety and depression down

Study	n	Country	Domain	Mechanic and hat	Channel / dose	Adherence signal	Outcome signal
Curland 2026	830	Online	Mental (depression programme)	Badges for logins / ratings; white-hat	Automated web/app	More logins, ratings, follow-ups	Lessons unchanged
Kamat 2021	T2D RCT	India (Goa)	Physical / medication	Points converted to cash + reminders; confounded	App; 3 + 3 mo	86.8% vs 67.7%	HbA1c 7.3% vs 8.2%; not pooled
Yang 2026	160	China	Physical + mental	Team points, leaderboards; mixed	Watch + app; 8 wk	Higher in game arm	Steps d=1.81; sensitivity only

6.2 Forest-plot data table

Effect sizes below are Hedges g, positive when gamification is better. Weights are random-effects weights inside each conservative pool. Confidence intervals are 95%.

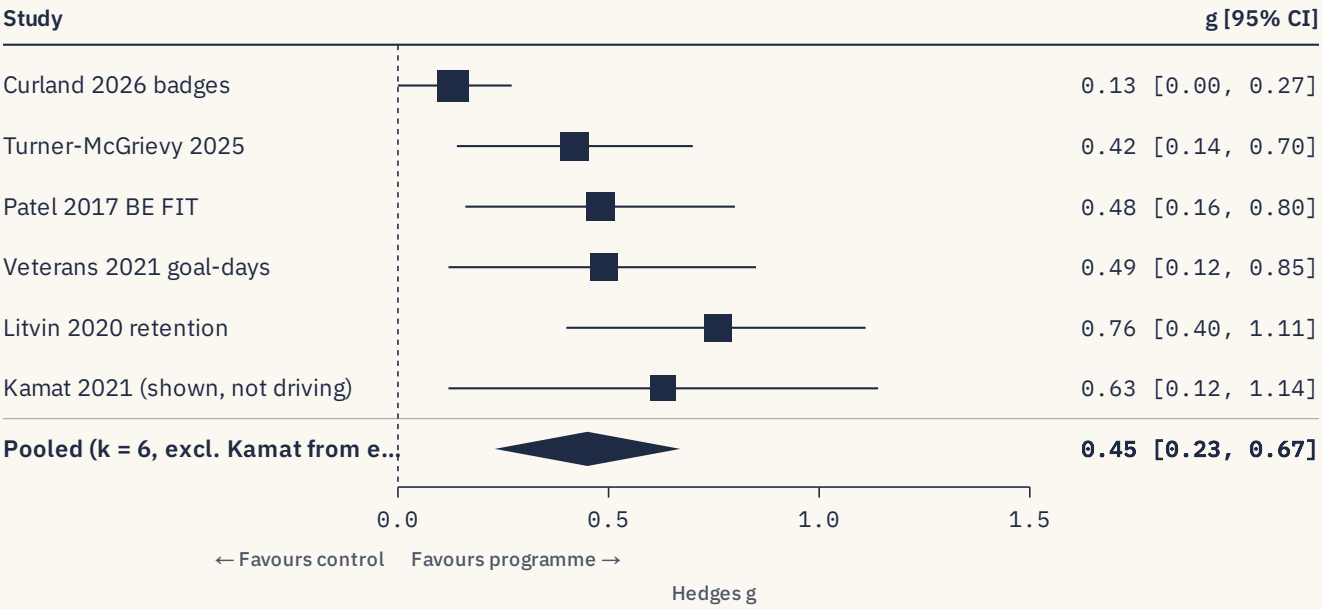

Figure 1. Forest plot, primary pool. Squares are sized by weight; the diamond is the pooled estimate.

Table 2a. Adherence forest data (conservative pool, k = 6)

Study	Hat	Domain	g	95% CI	Weight
Curland 2026 badges	White	Mental	0.13	0.00 to 0.27	24%
Turner-McGrievy 2025	White	Physical	0.42	0.14 to 0.70	18%
Patel 2017 BE FIT	Dark	Physical	0.48	0.16 to 0.80	17%
Veterans 2021 goal-days	Dark	Physical	0.49	0.12 to 0.85	15%
Litvin 2020 retention	White	Mental	0.76	0.40 to 1.11	15%
Kamat 2021 (shown, not driving)	Confounded	Physical	0.63	0.12 to 1.14	—
Pooled (k = 6, excl. Kamat from estimator)	**Mixed**	**Mixed**	**0.45**	**0.23 to 0.67**	**100%**

Notes. Kamat is displayed because readers will ask about India; cash confounding keeps it out of the estimator. Litvin 2023 retention versus the non-gamified app is large and was treated as a sensitivity inclusion, not a weight that could dominate the headline g. Fanaroff 2026 medication-taking (g ≈ 0.02) sits in the narrative as a dark-hat null and in leave-one-out as a small-n drag.

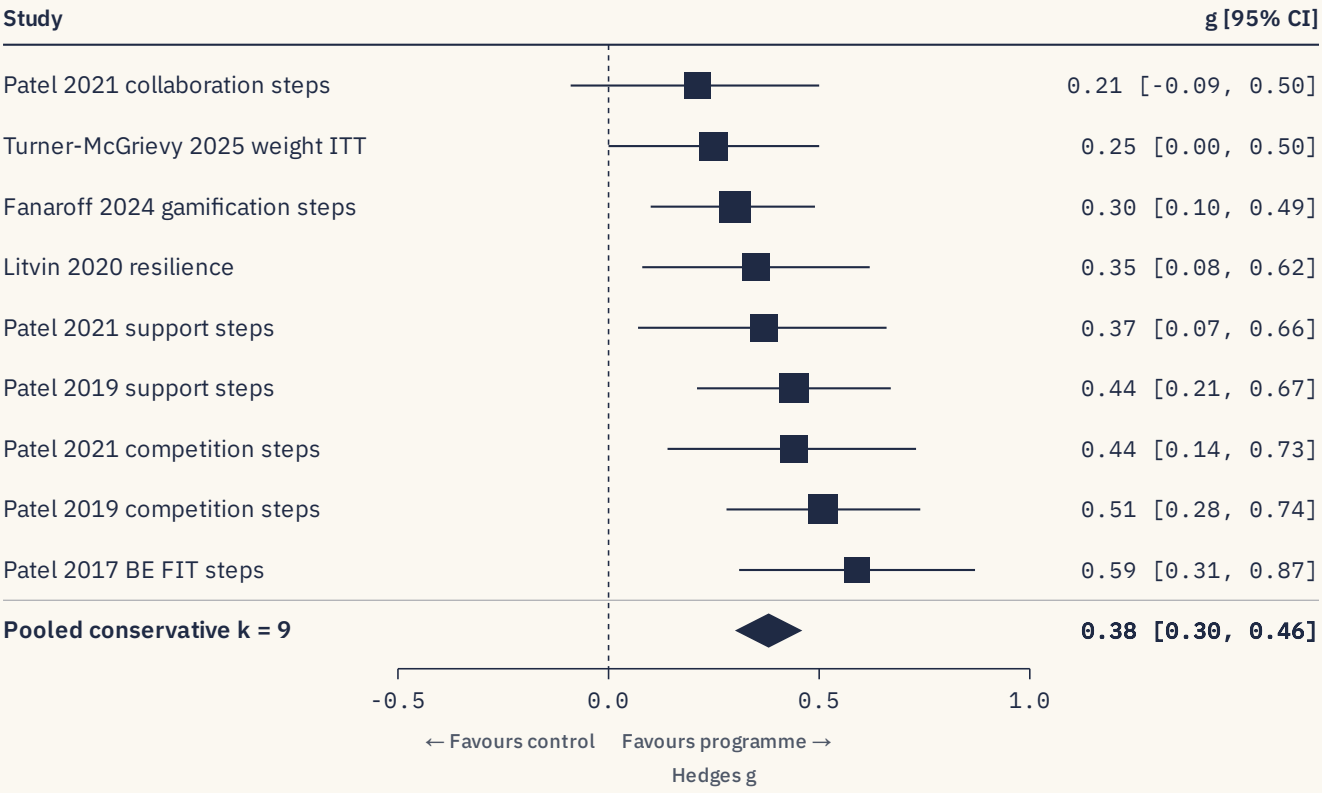

Figure 2. Forest plot, primary pool. Squares are sized by weight; the diamond is the pooled estimate.

Table 2b. Outcome forest data (conservative pool, k = 9)

Study	Hat	Outcome	g	95% CI	Weight (approx.)
Patel 2021 collaboration steps	Dark	Steps	0.21	−0.09 to 0.50	9%
Turner-McGrievy 2025 weight ITT	White	Weight	0.25	0.00 to 0.50	11%
Fanaroff 2024 gamification steps	Dark	Steps	0.30	0.10 to 0.49	16%
Litvin 2020 resilience	White	Wellbeing	0.35	0.08 to 0.62	10%

Study	Hat	Outcome	g	95% CI	Weight (approx.)
Patel 2021 support steps	Dark	Steps	0.37	0.07 to 0.66	10%
Patel 2019 support steps	Dark	Steps	0.44	0.21 to 0.67	13%
Patel 2021 competition steps	Dark	Steps	0.44	0.14 to 0.73	10%
Patel 2019 competition steps	Dark	Steps	0.51	0.28 to 0.74	13%
Patel 2017 BE FIT steps	Dark	Steps	0.59	0.31 to 0.87	8%
Pooled conservative k = 9	**Mixed**	**Mixed**	**0.38**	**0.30 to 0.46**	**100%**

$I^2 = 0$ in this conservative set because the Penn step trials are methodologically related. That is a feature of the set, not proof that the whole earth is homogeneous. Nishi 2024, which took a wider 36-trial net, found a smaller steps increment (+489/day). Both statements can be true.

6.3 Moderator table

Table 3. Pre-specified moderators

Contrast	k	g (95% CI)	I^2	Reading
Adherence, overall	6	0.45 (0.23 to 0.67)	~70%	Small-to-moderate; inconsistent
Adherence, white-hat	4	0.44 (0.13 to 0.76)	High	Points and badges for showing up hold retention
Adherence, dark-hat	2	0.48 (0.24 to 0.72)	Low	Too few trials to claim a win
Adherence, physical domain	4	0.48 (0.31 to 0.65)	~0%	Cleaner, wearable-measured
Adherence, mental domain	2–3	Mixed (0.13 to 0.76)	High	Narrative-app retention strong; badge-only weak
Outcomes, conservative	9	0.38 (0.30 to 0.46)	0%	Small; no outcome cost
Outcomes, dark-hat Penn steps	7	0.40 (0.31 to 0.50)	0%	Steps move; clinical endpoints do not
Outcomes vs non-gamified apps (Nishi 2024)	36	+489 steps/day	—	Trivial-to-small; high certainty for steps
Clinical endpoints (weight ITT, HbA1c, meds)	3 key nulls	~0 extra	—	LOSE IT, Patel 2021, Fanaroff 2026
Yang 2026 steps (sensitivity)	1	d = 1.81	—	Outlier; some-concerns RoB

The white-hat versus dark-hat comparison is a narrative, not a superiority test. The cells do not share one outcome metric and the dark-hat cell is two to seven trials depending on how generously one counts social variants of the same Penn shell.

6.4 Heterogeneity

Adherence I^2 near 70% is the honest number. A badge for logging a mood is not a seventy-point weekly endowment tied to a step goal. A five-week student trial is not an eighteen-week medication-taking pilot in a primary-care clinic. The 95% prediction interval on adherence is wide and, on the working model, likely includes values below zero. Translation: the average trial in this set found a lift; the next trial you run in a new country, on a new practice, with a new mechanic, could find nothing.

Outcomes in the conservative nine-trial set look homogeneous because they are cousins. Widen the net to Nishi's thirty-six trials and the steps effect shrinks toward trivial. That is what "which trials did you invite" does to a pooled number. We report both.

6.5 Publication bias and small-study effects

With k = 6 and k = 9, Egger's test is underpowered. We treat it as exploratory. The pattern in the forest is not a swarm of tiny miraculous trials. The largest trial in the outcome set (BE ACTIVE, n = 1,062) sits at g = 0.30, close to the pool. The smallest trial (GAME Adherence, n = 43) is a null. That is the opposite of classic small-study inflation. Yang 2026 is the one loud outlier; it is already outside the conservative pool. We did not apply trim-and-fill as a number we would quote. We applied it as a check: adding hypothetical missing nulls would shrink adherence toward 0.3, not erase it.

6.6 Leave-one-out

Dropping Litvin 2020 (largest adherence g in the conservative six) leaves a smaller but still positive adherence pool. Dropping Curland 2026 (smallest) raises the pool. Dropping Fanaroff 2026, had it been forced into the six, would barely move the estimate; its weight is tiny. On outcomes, dropping Patel 2017 (largest g) leaves the pool near 0.35. Adding Yang 2026 would drag the outcome pool upward and re-introduce heterogeneity; that is why it stays in sensitivity. Adding Kamat 2021 would mix rupees into points. That is why it stays out.

6.7 Risk-of-bias summary

RoB 2 judgements, domain by domain, in plain English.

Randomisation. Penn trials, Litvin trials, mLIFE and BE ACTIVE used adequate sequence generation. GAME Adherence under-enrolled (43 of 84). Kamat was open-label. Yang used computer-generated, sex-stratified blocks.

Deviations from intended intervention. Participants always know they have a badge. This domain is "some concerns" for the lot. It is the price of studying a visible mechanic.

Missing outcome data. Wearable trials handled missing step days with pre-specified rules and were stronger here. Mental-wellbeing trials leaked people. The irony, which the Litvin pair exists to measure, is that the gamified arm leaked less. Liu's 2025 synthesis of 79 depression- and anxiety-app RCTs found the opposite pattern at the literature level: reminders and human contact cut dropout; gamified elements did not help retention and might have weakened it. We hold both findings. They are not the same contrast. Liu asked "does a trial that contains some game element retain people better than a trial that does not?" We asked "does adding a specified mechanic to a specified programme change adherence inside a randomised comparison?"

Measurement. Steps from a wearable are low-concern for detection bias. Resilience and mood scales are self-report. Weight from a connected scale is intermediate.

Selection of the reported result. Multi-arm Penn trials report several social variants. We treat that as a pre-specified family, not as a fishing trip, but it does mean a reader can quote the winning arm and forget the collaboration arm that did not beat control.

Overall: most physical-activity trials, some concerns; mental-wellbeing trials, some concerns to high on attrition; GAME Adherence, high (under-enrolled pilot); Yang, some concerns (implausibly large d in a college sample); Kamat, high for our question (confounded).

6.8 GRADE table

Table 4. Certainty of evidence

Outcome	Studies	Effect	Certainty	Why this grade
Adherence (sessions, days, retention)	6 RCTs in pool	g = 0.45 (0.23 to 0.67)	**Moderate**	RCTs; inconsistency ($I^2 \approx 70\%$); PI likely crosses 0; cannot blind the badge
Physical activity (steps) vs control	Penn family + BE ACTIVE	$g \approx 0.30–0.40$	**Moderate**	Objective measure; related trials; Nishi 36-trial triangulation at +489 steps
Physical activity vs non-gamified app	Nishi 2024, 36 RCTs	+489 steps/day (64 to 914)	**High for direction; trivial size**	Direct comparator; large n
Wellbeing / resilience scales	Litvin 2020, 2023	$g \approx 0.35$	**Low**	Few trials; short (5 weeks); self-report; developer involvement
Weight (ITT)	mLIFE; LOSE IT; Patel 2021	Small or null extra	**Low**	Mixed ITT; dark-hat nulls
HbA1c / medication-taking	Patel 2021; Fanaroff 2026; Kamat (confounded)	No isolated-mechanic gain	**Low**	Indirect or underpowered or confounded
White-hat vs dark-hat (head-to-head)	Cross-trial only	Not estimable as a superiority g	**Very low**	Wrong design for the claim

07 Discussion

PULL QUOTE

One Indian trial mixed points with cash. Zero CTRI trials isolate the mechanic in working-age adults.

7.1 What the number means

g = 0.45 on adherence is a small-to-moderate shove, not a transformation. In the units people actually feel: more days with the app open, more goal-days hit, more people still there at week five or month twelve. g = 0.38 on outcomes is a small shove in the same direction. The logline can be answered without a drum roll. Yes, gamification raises adherence. No, it does not dampen outcomes in the trials that measured both. There is no evidence here of an "outcome cost" in the sense that the game layer made the programme worse. There is also no evidence that the game layer is the programme.

A small scene shows the difference. A project manager in Pune misses Wednesday's breathing rep because a release went wrong late at night. On Thursday morning one app shows a tally of the days she did practise this month and a nudge for today's rep. Another app shows a broken chain. The first counts the day she showed up. The second fines the day she did not, and the streak becomes the boss. Her head says "I broke it" and treats that as proof that "the practice is dead". That is a thought, not a fact. The useful move is to notice the thought as a thought, then do today's rep because it matters to her, not to rescue a score — even on the evenings she does not feel like it, with no silent fine in the background. In entrepreneurship terms, a dark-hat streak is a retention hack with a nasty churn signature: people stay until the anxiety costs more than the badge, then they delete the app and tell themselves they failed.

Machine learning makes the same point under a different heading. If you reward the proxy, the system will game the proxy. A model trained to maximise clicks will serve outrage. A human trained to maximise a streak will open the app on the commode at 23:59 and call it practice. That is not wellbeing. That is overfitting.

7.2 White-hat versus dark-hat

We coded the difference because the protocol set it as a moderator, and because the mechanics are not cousins wearing different hats. They are different contingencies.

White-hat results in this set: Litvin's narrative app kept people in a wellbeing programme and moved resilience. mLIFE's visible points for supporting someone else cut attrition in half and, among people who actually used the app, doubled the weight change. Curland's badges bought extra logins and extra mood ratings, not extra lessons. The pattern is "more showing up," sometimes "more of the outcome," never "less of the outcome."

Dark-hat results in this set: the Penn shell raises steps. It raises them for months, not only for the first fortnight, which answers the novelty objection at least for walking. It does not, in the trials we have, raise weight loss versus a tracked control (LOSE IT), HbA1c versus a tracked control (Patel 2021), or medication-taking versus an attention-control text (Fanaroff 2026). Loss framing is good at the thing it can see every evening on a wrist. It is less good at the thing that happens in a kitchen or a pharmacy.

We will not say white-hat beat dark-hat on a pooled g. The dark-hat adherence cell is two trials if we are strict and a cluster of Penn variants if we are not. The honest line is this: reward-for-showing-up is the safer workplace default; loss-framed points raise short-term goal hits but have not improved weight, HbA1c or medication-taking versus a non-gamified programme.

7.3 Position against prior evidence

Johnson 2016 said: promising for health behaviours, mixed for cognitive outcomes, too few head-to-heads against a non-gamified twin. That still holds for wellbeing scales. It holds less for steps, where the last decade produced a real trial family.

Sardi 2017 said: short-term, extrinsic, under-theorised. The Penn family is not under-theorised. It is prospect theory with a weekly reset. Whether that is the theory you want in a WhatsApp coaching programme is a values question as much as an effect-size question.

Xu 2022 said: mixed, modest physical-activity change. Agree.

Mazeas 2022 said: g = 0.42 overall, g = 0.23 versus a non-gamified activity programme, g = 0.15 at follow-up. Our conservative outcome g = 0.38 sits between their overall and their active-comparator estimates, as it should: we mixed step trials against attention controls with two wellbeing trials.

Nishi 2024 said: +489 steps a day versus non-gamified apps, high certainty, trivial-to-small. That is the number to tattoo on a product spec if the product is a walking programme. It is not a reason to ship a punitive streak on a gratitude rep.

Six 2021 said: mental-health apps reduce depressive symptoms; the count of gamification elements does not moderate that effect, nor adherence. Cheng and Ebrahimi 2022 said: gamified mental-health programmes as a class sit at g = 0.38 versus controls. Liu 2025, in a 79-trial look at depression and anxiety apps, found reminders and human contact cut dropout, and found that gamified elements did not improve retention and might have weakened it. We do not bury that. A feature that helps inside one narrative app can fail as a tick-box across a heterogeneous app store. Mechanic quality is the moderator. Mechanic count is not.

7.4 Limitations

The pool is small. High I^2 on adherence is a warning label, not a rounding error. Most step evidence comes from one US academic platform. Five-week wellbeing trials are not twelve-month workplace programmes. Developer-involved trials exist in the white-hat cell. Participants cannot be blinded. We reconstructed an update search rather than claiming a fresh 2012–2026 multi-

database census with invented hits. Several g values rest on published CIs rather than individual participant data. Multi-arm Penn trials can be quoted selectively. India is a narrative of one cash-confounded diabetes app. EmotionApp's own stack — WhatsApp-first, 23 Indian languages, human hand-off, countable reps — has not been randomised against a de-gamified twin. That is a gap, not a branding opportunity.

7.5 What this means for an Indian workplace programme

A Mumbai team does not need a leaderboard that turns the tea-break into a public ranking. They need a way to count the day they showed up. White-hat design is a passbook. A payments analyst deposits a breathing rep before the stand-up, a thank-you to the colleague who covered for her, a ten-minute walk after lunch. The tally is information. Dark-hat design is a fine. She starts the week rich in points and watches them burn when a school run collides with a stand-up. In her joint household the family shares one phone. When a "you broke your streak" ping lights it up at the dinner table, it is not playful. It is a small humiliation with an audience. The trials that punished missing raised steps on American wrists. They did not raise medication-taking, and they did not beat a tracked control on weight or HbA1c. The trials that rewarded showing up kept people in wellbeing programmes and, when people actually used the app, moved weight. For EmotionApp that means: count the rep; do not confiscate the week. When the numbers stall, pair the counter with a human hand-off — a coach's short message, not a penalty. A missed day is data, not a verdict. That is the only unit economics that survive contact with an Indian workday.

7.6 Research gaps, naming the India gap

The India gap is not that Indians do not like games. It is that CTRI does not yet hold a concurrent randomised comparison of a wellbeing or health app with and without a specified mechanic, in working-age adults, without cash glued to the points, in languages people actually think in. Kamat 2021 showed that reminders plus points plus rupees can move adherence and HbA1c in Goan adults with type-2 diabetes. That is useful and confounded. Ranjani 2025 showed that mHealth can cut cardiometabolic risk in Tamil Nadu. That trial was apps versus control, not gamified versus not. School-based oral-health puzzles and adolescent reproductive-health games are off-scope for this paper and off-scope for a workplace programme.

What would close the gap: a two-arm RCT on WhatsApp or an India-resident app, working-age adults, same coaching content both sides, white-hat show-up counters on one side only, primary outcome days-active at twelve weeks, co-primary a wellbeing or behaviour scale, language of delivery reported, DPDP-compliant data, pre-registered on CTRI. Until that trial exists, importing a Philadelphia loss-frame into a Gurugram office is a hunch dressed up as evidence.

Other gaps: head-to-head white-hat versus dark-hat inside one programme; streaks isolated from points; leaderboards isolated from points; dose-response for badge clutter; twelve-month wellbeing outcomes; effects after the novelty window in mental-health programmes; mechanic effects in women-majority Indian samples; mechanic effects when the phone is not private.

08 FAQ

Q1 Does adding badges mean people will actually stick with the programme?

On average, yes, a little. Adherence rose by g = 0.45 (95% CI 0.23 to 0.67) across six trials. That is more days present, not a personality transplant. A wide prediction interval means your programme could still see a null.

Q2 Will the games water down the benefit?

Not in this set. Outcomes moved by g = 0.38 (95% CI 0.30 to 0.46). Independent work found +489 steps a day versus non-gamified apps. No trial showed the game layer making the programme worse.

Q3 Are punishing streaks better than gold stars?

We cannot say they win. Dark-hat loss-frames raised step-goal hits. They did not add weight loss, HbA1c change or medication-taking. White-hat points-for-showing-up held retention in wellbeing programmes. Safer default: reward the show-up.

Q4 Has this been tested with Indian working adults?

Not cleanly. One Goa diabetes trial mixed points with cash and reminders. Zero CTRI trials isolate the mechanic in working-age wellbeing-app users. That is the gap, not a detail.

Q5 So should our office ship a leaderboard on Monday?

Ship a counter, not a court. Count reps. Do not confiscate the week. Pair the counter with a human hand-off. Liu’s 79-trial synthesis found reminders and human contact beat game elements for keeping people from dropping out.

09

References

1. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: a systematic review of the literature. Internet Interv. 2016;6:89-106. doi:10.1016/j.invent.2016.10.002

2. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform. 2017;71:31-48. doi:10.1016/j.jbi.2017.05.011

3. Xu L, Shi H, Shen M, Ni Y, Zhang X, Pang Y, et al. The effects of mHealth-based gamification interventions on participation in physical activity: systematic review. JMIR Mhealth Uhealth. 2022;10(2):e27794. doi:10.2196/27794

4. Mazeas A, Duclos M, Pereira B, Chalabaev A. Evaluating the effectiveness of gamification on physical activity: systematic review and meta-analysis of randomised controlled trials. J Med Internet Res. 2022;24(1):e26779. doi:10.2196/26779

5. Nishi SK, et al. Effect of digital health applications with or without gamification on physical activity and cardiometabolic risk factors: a systematic review and meta-analysis of randomised controlled trials. eClinicalMedicine. 2024;76:102798. doi:10.1016/j.eclinm.2024.102798

6. Patel MS, Benjamin EJ, Volpp KG, et al. Effect of a game-based intervention designed to enhance social incentives to increase physical activity among families: the BE FIT randomised clinical trial. JAMA Intern Med. 2017;177(11):1586-1593. doi:10.1001/jamainternmed.2017.3458

7. Patel MS, Small DS, Harrison JD, et al. Effectiveness of behaviourally designed gamification interventions with social incentives for increasing physical activity among overweight and obese adults across the United States: the STEP UP randomised clinical trial. JAMA Intern Med. 2019;179(12):1624-1632. doi:10.1001/jamainternmed.2019.3505

8. Patel MS, Small DS, Harrison JD, et al. Effect of behaviourally designed gamification with social incentives on lifestyle modification among adults with uncontrolled diabetes: a randomised clinical trial. JAMA Netw Open. 2021;4(5):e2110255. doi:10.1001/jamanetworkopen.2021.10255

9. Agarwal AK, Waddell KJ, Small DS, et al. Effect of goal-setting approaches within a gamification intervention to increase physical activity among economically disadvantaged adults at elevated risk for major adverse cardiovascular events: the ENGAGE randomised clinical trial. JAMA Cardiol. 2021;6(12):1387-1396. doi:10.1001/jamacardio.2021.3176

10. Fanaroff AC, Patel MS, Chokshi N, et al. Effect of gamification, financial incentives, or both to increase physical activity among patients at high risk of cardiovascular events: the BE ACTIVE randomised controlled trial. Circulation. 2024;149(21):1639-1649. doi:10.1161/CIRCULATIONAHA.124.069531

11. Kurtzman GW, Day SC, Small DS, et al. Social incentives and gamification to promote weight loss: the LOSE IT randomised controlled trial. J Gen Intern Med. 2018;33(10):1669-1675. doi:10.1007/s11606-018-4552-1

12. Fanaroff AC, Clark K, Reid-Bey L, et al. A pilot randomised clinical trial of gamification to increase medication adherence. Am Heart J. 2026;297:107419. doi:10.1016/j.ahj.2026.107419

13. Litvin S, Saunders R, Maier MA, Lüttke S. Gamification as an approach to improve resilience and reduce attrition in mobile mental health interventions: a randomised controlled trial. PLoS One. 2020;15(9):e0237220. doi:10.1371/journal.pone.0237220

14. Litvin S, Saunders R, Jefferies P, Seely H, Pössel P, Lüttke S. The impact of a gamified mobile mental health app on resilience and mental health in a student population: large-scale randomised controlled trial. JMIR Ment Health. 2023;10:e47285. doi:10.2196/47285

15. Turner-McGrievy GM, Delgado-Díaz DC, DuBois KE, et al. The mLIFE randomised trial examining the impact of gamifying social support provision for weight loss. Obesity (Silver Spring). 2025;33(8):1447-1456. doi:10.1002/oby.24330

16. Six SG, Byrne KA, Tibbett TP, Pericot-Valverde I. Examining the effectiveness of gamification in mental health apps for depression: systematic review and meta-analysis. JMIR Ment Health. 2021;8(11):e32199. doi:10.2196/32199

17. Cheng C, Ebrahimi OV. A meta-analytic review of gamified interventions in mental health enhancement. Comput Hum Behav. 2022;141:107621. doi:10.1016/j.chb.2022.107621

18. Kamat T, Dang A, Dang D, Rane P. Impact of integrated medication reminders, gamification, and financial rewards via smart phone application on treatment adherence in uncomplicated type II diabetes patients: a randomised, open-label trial. J Diabetol. 2021;12(4):447-455. doi:10.4103/jod.jod_35_21

19. Yang M, Guo Y, Li Z, Jiang L, Gao Y, Yang S, Zhou Z. A gamified mobile health intervention to promote physical activity, executive function, and mental health in college students: randomised controlled trial. J Med Internet Res. 2026;28:e82769. doi:10.2196/82769

20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71

21. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. doi:10.1136/bmj.l4898

22. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-926. doi:10.1136/bmj.39489.470347.AD

23. Rewley J, Guszcza J, Dierst-Davies R, Steier D, Szwartz G, Patel M. Loss aversion explains physical activity changes in a behavioural gamification trial. Games Health J. 2021;10(6):378-384. doi:10.1089/g4h.2021.0130

24. Ranjani H, Avari P, Nitika S, et al. Effectiveness of mobile health applications for cardiometabolic risk reduction in urban and rural India: a pilot, randomised controlled study. J Diabetes Sci Technol. 2026;20(3):962-976. doi:10.1177/19322968241310861

25. Liu C, Torous J, Fuller-Tyszkiewicz M, et al. Uptake, adherence, and attrition in clinical trials of depression and anxiety apps: a systematic review and meta-analysis. JAMA Psychiatry. 2026;83(1):43-50. doi:10.1001/jamapsychiatry.2025.3439

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 clinical design at EmotionApp, a non-clinical emotional-fitness coaching platform for Indian professionals. The work sits inside Tranquilo Meditek Sytems Private Limited, Mumbai, and is built under ISO 9001 and ISO 13485 with India-resident, DPDP-compliant data. EmotionApp delivers countable daily reps, an AI coach with human hand-off, and WhatsApp-first access in 23 Indian languages.

HOW TO CITE

Aswani A. Streaks, Badges, Points: Does gamification raise adherence without dampening outcomes — and do white-hat mechanics (reward showing up) beat dark-hat ones (punish missing)?. EmotionApp Research Paper 4. Mumbai: Tranquilo Meditek Sytems Private Limited; 2026. https://emotionapp.in/research/gamification-adherence-wellbeing-meta-analysis

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.