TECHNICAL APPENDIX

# The evidence behind How to Change

## Claims, source checks, scoring, and disclosures

Katy Milkman · Evidence assessed through 13 September 2026

66 min read

This appendix is the supporting record for the reader review. Start with the claim register to locate an issue, then follow the link to its detailed assessment, sources, and qualifications. The tables use a consistent layout across the series, but each book retains its own audit method and original identifiers.

Overall 72% · Scientific Accuracy 75% · Reference Accuracy 67% · Practical Value 75%.

### About This Review

This review combines factual reporting with editorial judgments about the identified book and its evidence. Ratings apply our published rubric; they are not percentages of truth or judgments about an author's honesty. Criticism of support does not itself show that a claim is false. Evidence is assessed through 13 September 2026, with source-access limits disclosed. Preparation used AI, and independent human fact-checking or legal clearance is not documented. Read the methodology, book-specific disclosures, and full editorial notice with the review. Corrections and substantive responses are welcome.

## Find the supporting evidence



EVIDENCE MAP

# Claim register

Use this register to locate a specific claim or a question about how the book uses a source. Each row gives a verdict and links to the detailed assessment. The verdict applies only to the claim or source relationship in that row—not to the whole book or its author. Follow the linked exhibit for the full wording, sources, qualifications, original audit identifiers, and recommended treatment.

> **Verdicts:** Supported; Partly supported; Overstated; Incorrect; Unresolved; Practical judgment. "Incorrect" requires an identified factual or measurement problem; "unresolved" is an access or verification limit, not a failed claim. "Practical judgment" identifies advice or a safeguard rather than a tested effect. These routing labels add no numerical score. Definitions.

| ID | Claim or focal issue | Verdict | Treatment and supporting detail |
|---|---|---|---|
| C01 | Fresh-start savings invitations | **Supported** | Relative framing, not an arithmetic error. Full assessment and sources |
| C02 | Calorie labelling does essentially nothing | **Incorrect** | Direct mismatch; purchasing is not consumption. Full assessment and sources |
| C03 | "Carrot-eater" labels change our actions | **Incorrect** | Wrong outcome used as support. Full assessment and sources |
| C04 | Initial 55% exercise increase with bundling | **Supported** | Defensible rounding; first-week effect only. Full assessment and sources |
| C05 | Teaching bundling yields lasting behavior change | **Partly supported** | Package/design compression. Full assessment and sources |

| ID | Claim or focal issue | Verdict | Treatment and supporting detail |
|---|---|---|---|
| C06 | Locked savings accounts produce 80% more saving | **Overstated** | Denominator and bank-versus-household scope.<br>Full assessment and sources |
| C07 | Self-set deadlines yield 50% fewer errors | **Incorrect** | Endpoint and percentage error; underlying paper retracted on 2 September 2026. No supporting credit.<br>Full assessment and sources |
| C08 | Cash commitments help smokers quit | **Supported** | Strong support for the bounded direction of effect.<br>Full assessment and sources |
| C09 | Public clinician pledges cut inappropriate prescribing | **Partly supported** | Small positive US trial; larger English trial had a null primary outcome under a different implementation.<br>Full assessment and sources |
| C10 | Planning prompts increase flu vaccinations by 13% | **Supported** | Accurate relative effect.<br>Full assessment and sources |
| C11 | Surgical checklists cut complications/mortality by 35–45% | **Overstated** | Large original association is not a universal effect; Ontario null and favorable Norwegian randomized results need separate interpretation.<br>Full assessment and sources |
| C12 | Generic prescribing rises from 75% to 98% | **Partly supported** | Accurate numbers; observational setting.<br>Full assessment and sources |
| C13 | Opioid prescriptions cut in half by a ten-pill default | **Incorrect** | Direct conflict with the linked study.<br>Full assessment and sources |
| C14 | Flexible incentives create more persistent exercise | **Partly supported** | Good policy evidence; broader habit language needs qualification.<br>Full assessment and sources |
| C15 | Giving advice improves students' grades | **Supported** | Accurate bounded trial description.<br>Full assessment and sources |
| C16 | Housekeeper expectations improve health without changed routines | **Overstated** | Tentative evidence narrated too confidently. Include the mixed 2011 replication and retain the mechanism and generalizability limits.<br>Full assessment and sources |
| C17 | Milkshake expectations alter ghrelin response | **Partly supported** | Acute biomarker finding; minor reference correction, no retraction identified, and direct independent replication remains unestablished.<br>Full assessment and sources |
| C18 | Growth mind-set produces broadly much greater achievement | **Overstated** | Narrow result extended beyond its demonstrated scope.<br>Full assessment and sources |
| C19 | Engineered peer groups can backfire | **Supported** | Strongly supported cautionary example.<br>Full assessment and sources |
| C20 | Energy-report effects decay after discontinuation | **Supported** | Accurate durability caveat; not 10–20% growth in household energy use.<br>Full assessment and sources |

## 01 The verdict in context

**A useful book about changing the conditions of behavior—not a validated system for diagnosing every reader's obstacles.** Its strongest material concerns concrete interventions: plans, defaults, commitments, incentives, and continuing support. Its weakest passages turn narrower findings into larger conclusions or misstate what an illustrative study measured.

The distinction matters in practice. Readers can use the book to generate inexpensive, measurable experiments without accepting the seven-obstacle framework as a scientifically calibrated personalization tool. Its unusually candid treatment of maintenance deserves credit: an arrangement can remain useful even when its effects depend on keeping it in place.

Several corrections are consequential rather than cosmetic. A study of labels and children's judgments was not an eating-behavior experiment; the cited opioid-default study does not establish a halving of overall prescribing; and purchase-calorie changes are not direct measures of eating or weight. These do not make the whole toolkit ineffective. They do weaken confidence in some of its scientific storytelling. The source audit gives the exact claims, locations, and limits.

**Best use:** select a specific behavior, identify the actual obstacle, choose a low-burden intervention, and evaluate results and costs. Treat that sequence as a testable working method, not a guarantee supplied by the book.

> **Recommendation:** Recommended as a practical toolkit, with important corrections.

**Confidence:** Moderate confidence in the overall appraisal; high confidence in the specific outcome and arithmetic corrections identified below.

**Ease of application:** Fairly easy to begin; some strategies need continuing support.

**Best use:** Practical tools with real field evidence; strongest when used selectively and measured.

**Rating profile:** Overall 72% · Scientific Accuracy 75% · Reference Accuracy 67% · Practical Value 75%. Full rationale.

## 02 Scope and method

**Question:** How well do the reviewed book's important scientific claims hold up, how faithfully does it use sources, and what can readers responsibly do with its advice? The intended audience is an interested general reader, with enough detail for researchers and editors to audit the reasoning.

**Edition and scope.** 2021 Portfolio / Penguin ebook; ISBN 9780593083765. Coverage comprises 20 purposively selected claim–source checks and all eight chapters. Book citations identify chapters and named sections of the reviewed edition. The source copies have not been authenticated against publisher-controlled masters.

**Preparation.** This appendix documents the claim-by-claim evidence assessment behind the reader review. It includes the source-specific replication, correction, and retraction checks described in Sections 08–12. The source-access record identifies what was examined; it is not a claim that every original analysis was independently rerun. See also the shared methodology.

**Claim selection.** Three central propositions summarize the book's organizing argument, its proposed mechanism or intervention, and important practical extensions. Selection is purposive and retrospective, not random or preregistered. All chapters remain covered to expose peripheral but consequential claims. The three proposition grades are not a sample-based estimate of the truth of the whole book. Narrow existence claims are easier to support than universal superiority claims; the stated proposition must always accompany the grade.

**Evidence rules.** The analysis distinguishes source identity, what the cited study actually found, the strength of its design, independent corroboration, and transfer to the book's practical claim. A controlled component study does not validate a branded package; an association does not identify an intervention effect; failure to locate evidence is not proof of its absence. Null findings, attrition, selection, selective reporting, measurement limitations, and competing explanations matter where they change an inference. Evidence available when the book was published and research published later are identified separately.

**Ratings.** The three categories are Scientific Accuracy, Reference Accuracy and Practical Value. Each uses three explicitly anchored judgments. Category percentages are calculated from the original inputs; the overall rating gives each category equal weight. Final displays use whole-number percentages, with all rounding performed after calculation. Ease of application and consequential cautions are reported separately. Section 09 and the series methodology provide the exact inputs, rules and limits. These are editorial judgments, not probabilities, estimates of the percentage of true content, or validated psychometric measurements.

**Reference checks.** Targeted source audits are stress tests, not prevalence estimates. The number of discrepancies in a purposive sample cannot estimate the proportion of a book's references that are wrong. The Grit report additionally preserves an explicitly restricted-frame seeded selection, with its missing-source cases and inherited inclusion coding disclosed. It is not the same audit design as the other books.

**Relationship to Red Pen Reviews.** The series adapts the separation of central scientific claims, reference support and practical consequences from the public Red Pen Reviews process and method, version 2.2. It uses a behavioral-science Practical Value category instead of nutrition-specific Healthfulness. Reference Accuracy is a common three-criterion assessment, not Red Pen Reviews' ten-reference random-sample score. The three category percentages are averaged, but there is no claim of an independent second expert reviewer. These departures mean the scores are not directly interchangeable with official Red Pen Reviews scores. This review is not affiliated with or endorsed by Red Pen Reviews.

**Research cutoff:** 13 September 2026. Searches covered the studies and claims discussed in this review and appendix, including named studies without separate reference rows. They do not establish exhaustive coverage of all endnotes in the original book or every study inside a cited synthesis. Source access and unresolved identities remain explicit in Section 10.

## Search record

The research record below identifies the documented searches, source reads, and assessments. Targeted repeat checks are distinguished from the initial research; they do not mean that every search or analysis was repeated. Source access is recorded in Section 10.

### Evidence gathering and decision rules

Research was conducted on 8 September 2026. The search began with the book's linked studies, then used their titles, authors, DOIs, and named interventions to seek primary publications, later trials, methodological critiques, and quantitative syntheses. Priority went to the central recommendations, large or memorable statistics, causal-mechanism claims, and possible disagreements between the narrative and a cited source. Publisher sites, PubMed/PMC, scholarly societies, author manuscripts, and university repositories supplied the evidence. The bibliography records whether the relevant material was an abstract, selected primary-text sections, or a full document.

This is a **critical narrative review with a structured source audit**, not a systematic review of the entire behavior-change literature. The search date is not a guarantee that all research published by that date was captured. A recently published meta-analysis may also synthesize older searches; publication year is not the endpoint of its underlying evidence collection. In particular, the 2025 calorie-labelling review and the recent planning meta-analysis disclose search-update limitations. [4, 18]

Five distinctions govern the analysis. First, a study can support the existence of an effect without establishing a useful effect in every setting. Second, intervention packages do not automatically identify their active ingredient. Third, observed attendance, purchasing, and enrollment are not interchangeable with automatic habit, consumption, and net wealth. Fourth, short-run differences do not establish permanent change. Fifth, a claim must be judged at the strength at which the book makes it: "might help" and "proven to transform" deserve different evidentiary burdens.

## 03 The strongest fair reading

The book's constructive thesis is that difficulty changing behavior need not be explained by weak character or insufficient desire. Different obstacles call for different interventions, and altering the situation can be more productive than demanding more willpower. Its chapters organize this message around starting, impulsivity, procrastination, forgetting, inertia, confidence, and social influence, before addressing maintenance.

A fair review therefore must not reduce the book to a claim that a single trick creates permanent habits. The closing chapter explicitly discusses disappointing persistence after short interventions and the possible need for continuing support. Equally, a fair reading cannot treat every piece of advice as equally established merely because some cited field experiments are convincing. The assessment below keeps the broad practical idea, each individual technique, and the proposed matching framework separate. Reviewed book; see especially Introduction; ch. 8.

## 04 Three central scientific claims

These are the three rated propositions, stated at the level assessed—not three verbatim quotations or an exhaustive list of the book's claims.

**CLAIM 1 · 50%**

**Matching a strategy to the obstacle beats an all-purpose approach.**

A useful design heuristic, but the seven-obstacle taxonomy is not itself a validated diagnostic or treatment-matching instrument. The relevant comparison is matched treatment against a credible alternative, not merely intervention against no intervention.

**CLAIM 2 · 100% · Bounded existence claim**

**Behavioral tools can improve concrete action.**

Randomized field interventions support this carefully bounded existence claim. The grade does not extend to every technique, every population, or a promise of lasting individual transformation.

**CLAIM 3 · 75%**

**Maintaining change often requires continuing support and adaptation.**

Maintenance is an important corrective to one-shot self-help. Persistence after intervention and the need for continued support vary by behavior; neither permanent support nor permanent automaticity is established universally.

## The three central claims

### Claim 1: Matching a strategy to the obstacle is better than relying on an all-purpose strategy

> **Rating: 50% — sensible and partly supported, but not validated at the level promised.**

Milkman repeatedly argues that progress improves when people identify their particular obstacle and tailor their approach. The introduction presents this as the book's organizing advantage over generic self-help. It is a good engineering heuristic: a reminder is unlikely to solve a lack of access, and a reward is unlikely to repair a missing skill. But the book's appealing taxonomy is not itself a tested diagnostic instrument. [1, Introduction]

Some experiments support the narrower idea that circumstances matter. Planning prompts performed differently at one-day and multiday vaccine clinics; flexible versus time-restricted exercise incentives produced different attendance patterns; and deliberately engineered peer groups harmed the students they were intended to help. These findings argue against indiscriminate deployment. They do not establish that readers can reliably classify themselves into the book's seven categories and thereby select a superior intervention. [17, 23, 30]

A direct test would measure obstacles before intervention, use a prespecified matching rule, and compare its outcomes with credible alternatives: a strong general program, the best single intervention, participant choice, or an equally resourced but differently matched program. It would also test whether the measured obstacle predicts **which treatment works better**, not merely who has poorer outcomes. This distinction is crucial. Being forgetful may predict missing appointments without proving that a reminder is the best incremental intervention for that person.

The reviewed evidence does not establish that complete chain for Milkman's framework. Nor does its absence prove matching is ineffective. **The defensible claim is that obstacle diagnosis is a promising design principle requiring outcome-based testing, not an already validated personalization algorithm.** For an individual reader, its value may lie in generating better experiments rather than in delivering a correct diagnosis on the first attempt.

### Claim 2: Behavioral tools can help people translate goals into concrete action

> **Rating: 100% — strongly supported for the bounded claim that such tools can improve concrete action.**

This is the book's empirical center of gravity. Changing an interface default, prompting a private vaccination plan, offering a commitment account, and changing exercise incentives are actual interventions, not merely correlations between successful people and favorable traits. The supporting research is sufficient to reject the blanket claim that behavioral science supplies only attractive stories. [21, 17, 14, 23]

The strong rating for this existence claim is not an endorsement of every item on the menu. The book packages very different evidence bases together. An administrative outcome in a randomized field trial, an acute hormone response in a small laboratory study, an observational relationship, and a story about an accomplished person do not deserve the same confidence. Successful implementation also frequently includes infrastructure the reader does not acquire simply by learning a concept: convenient clinics, payroll systems, repeated messages, tracking devices, or funded incentives. [17, 5, 26, 11]

The practical consequence is not to discard small effects. A four-percentage-point improvement can matter across thousands of people. It is to distinguish **population leverage** from **individual certainty**. An intervention that improves the average probability of an action may be worthwhile while leaving most individual outcomes unchanged. A vivid relative percentage should never be read as a promise that a reader's entire life will improve by that amount.

#### Claim 3: Maintaining change often requires continuing support and adaptation

> **Rating: 75% — a strong corrective, with limits on the universal formulation.**

The final chapter acknowledges that most brief gym interventions did not yield large durable effects. It argues for ongoing management and reconfiguration when obstacles change, rather than treating temporary intervention as a cure. The subsequent publication of the gym megastudy is consistent with that account, and the energy-conservation evidence directly tests what happens when recurring reports stop. [1, ch. 8] [36, 37]

This is one of the book's most important strengths. It changes the standard of success from "Did I become a permanently different person?" to "Does the arrangement keep producing the desired behavior at an acceptable cost?" Continuing a reminder or a well-chosen default need not represent failure. It can be the intervention working as designed.

The limitation is the breadth of the chronic-disease analogy. Some skills, one-time actions, capital purchases, and established routines can persist with little ongoing intervention; different behaviors have different maintenance requirements. An effect after withdrawal also need not be habit automaticity—it could reflect learning, a changed environment, or another persistent consequence. The best-supported version is therefore **to plan for maintenance and test its necessity**, not to assume that every useful technique must be applied permanently. [37, 23]

# 05 Chapter-by-chapter assessment

## Chapter 1: Getting Started

### What the chapter contributes

Milkman's "fresh start" framework treats timing as part of intervention design. Calendar landmarks, birthdays, moves, and resets in performance metrics can make a new attempt feel more psychologically available. She distinguishes real environmental change from symbolic temporal boundaries and warns that disruption can also damage an already successful routine. These qualifications are integral to the chapter, not additions supplied by this review. [1, ch. 1]

Its strongest practical lesson is not "wait for New Year." It is to notice moments when a person is already reconsidering their behavior and make the desired next step accessible then. The chapter also does a useful piece of scientific teaching: it acknowledges that early studies of intentions and student samples were insufficient to answer whether a fresh-start invitation changes meaningful real-world behavior. [1, ch. 1]

### Evidence and its proper scale

The retirement-saving field study provides that stronger test. Among 6,082 employees at four universities, labelling a future starting date as a fresh start increased take-up by approximately 2.2–2.3 percentage points against a 7.6% comparison baseline. That is a substantial relative increase but a modest absolute one. The book's 20–30% framing is broadly defensible; it would be more informative alongside the baseline. The outcome is especially relevant because choices were implemented through payroll arrangements, not merely recorded as intentions. [5]

The same infrastructure limits the lesson. Once someone elects a payroll deduction, repeated saving can continue without repeated psychological recommitment. That supports the value of combining a timely invitation with a low-friction system; it does not demonstrate that a birthday generates months of extra motivation by itself. Neither does a success at one type of future-date invitation establish that any personally meaningful landmark is optimal for any goal.

The original savings article also reports that not every landmark framing worked. A different randomized medication-reminder trial did not detect a statistically significant adherence advantage from birthday or New Year timing or framing during its 90-day follow-up. Some reminders arrived late, other reminders were used, and there was no reminder-free arm. This is a boundary on transfer to medication adherence, not a replication failure of payroll saving. The original savings result remains credited at its own population, comparison, and endpoint. [5, 58]

### Two supporting passages need correction

The chapter contrasts fresh-start opportunities with purportedly ineffective calorie information. Yet its cited Starbucks study found lower calories purchased per transaction, not essentially no effect. Later Cochrane evidence also supports a **small** reduction in calories selected or purchased, with less certain evidence on actual consumption. The correction should preserve both halves: the intervention is not a dramatic solution to eating behavior, but "essentially ineffective" is not what the linked evidence establishes. [1, ch. 1] [3, 4]

The "carrot-eater" example is a different kind of problem. The study concerns how labels affect children's inferences about another person's characteristic, not whether a label makes the child eat more carrots. That is an outcome mismatch, not merely a dispute over effect size. The adjacent voter-identity example also deserved a qualification in 2021: an independent 2016 field study failed to find the original noun-over-verb advantage. Its delivery method and election setting differed, so it is a failure of generalization rather than a perfect replication. [1, ch. 1] [6, 7]

Further voter tests include a larger 2018 field replication, a 2019 reanalysis by Bryan and colleagues, and a preregistered presidential-election test conducted in 2016 and published online in 2020. The reanalysis reports a positive effect under restricted specifications and disputes the earlier replication's choices. It is not a new independent experiment. The closer presidential-election test, with 2,219 participants and voter-record outcomes, did not establish a noun advantage. Approximate two-sided 95% normal intervals reconstructed here from the pooled, rounded published estimates and standard errors exclude the original 11–14-percentage-point advantage in this sample; the paper's printed one-sided bounds do not themselves support that exclusion. A 2024 synthesis reports a small significant average across exact and conceptual noun-language tests, but unstable diagnostic evidence. That does not establish reliable noun superiority and does not retest the carrot study's inference outcome. [44, 46, 48, 49]

The 2018 replication has a later sign correction: its corrected full-sample raw noun-minus-verb table contrast is positive but imprecise, not evidence of a reliable large benefit. The notice's prose contains a conflicting decimal, so no new exact effect estimate is certified here. The 2019 reanalysis's 2021 correction is a separate printed-reference repair, not a changed voter result. [45, 47]

### Arithmetic, anecdotes, and causation

The baseball illustration says Orlando Cabrera's batting average rose from .246 to .294, a 29% increase. Using the book's own numbers, the increase is **19.5%**. This is an internal arithmetic error, independent of whether the underlying baseball research is sound. More importantly, a single player's improvement is illustrative, not evidence identifying a psychological reset as the cause. [1, ch. 1]

The same restraint applies to compelling life-change stories. A successful person's account can illuminate how a fresh start felt. It cannot establish what would have happened without that landmark. Selection into memorable success stories makes such accounts poor estimators of the probability that a similar attempt will work for someone else.

**Revised takeaway:** use a meaningful transition to launch an immediately executable action; do not postpone a useful action merely to obtain a more impressive starting date. Protect routines that already work. Treat the timing effect as an aid to initiation, not as an engine of guaranteed maintenance.

## Chapter 2: Impulsivity

#### The core idea survives scrutiny

This chapter proposes changing the immediate experience of a worthwhile action instead of repeatedly demanding self-denial. Its two principal tools are temptation bundling—pairing a desired indulgence with an avoided activity—and gamification. Milkman acknowledges incompatibilities between demanding tasks and warns that imposed "fun" can backfire. These limitations make the recommendations more usable than a blanket instruction to reward everything. [1, ch. 2]

The conceptual distinction between **immediate reward** and **distant benefit** is useful. But "impulsivity" and "present bias" are not complete synonyms for every case of preferring a pleasant activity. The book uses them broadly. A person may choose leisure because it is genuinely valuable, because they lack energy, or because an institutional goal is not their own. Calling a choice impulsive is not a substitute for identifying the relevant trade-off.

#### Temptation bundling: a genuine effect, correctly time-limited

The original experiment randomized 226 participants to gym-restricted audiobooks, a self-restriction suggestion, or a control. The headline that initially looks inflated—55% more exercise—is defensible: Table 2 reports **1.16 versus 0.75 gym visits in the first week**, a raw difference of about 54.7%. The article's often-quoted 51% is a different, regression-adjusted estimate. The two numbers need not contradict each other. [8]

Milkman explicitly reports that benefits faded around the Thanksgiving interruption. She should receive credit for that disclosure. The important constraint is the outcome: gym entries are not total physical activity, fitness, or proof of an automatic habit. Nor was the early percentage an effect sustained at the same magnitude over the entire experiment. [1, ch. 2] [8]

The larger follow-up is encouraging but methodologically more complicated than the narrative suggests. Teaching was tested against audiobook provision within a broader exercise program. Random assignment also included the no-audiobook comparison, using unequal probabilities that changed over time. The authors describe the unpreregistered, weighted pooled comparison as quasi-experimental. Participants also received planning support, reminders, and small rewards. The incremental randomized effect of the teaching component was modest, especially for the number of visits rather than whether any visit occurred. This supports bundling as part of an implemented program—not a claim that a suggestion alone reproduces all of the reported benefit. [9]

#### Gamification: do not confuse an intervention label with a mechanism

The Wikipedia awards study supports an effect of symbolic recognition on volunteer retention. But recognition can operate through status, identity, community belonging, or attention as well as enjoyment. Calling the intervention "gamification" does not establish that converting work into a game was its causal ingredient. A leaderboard, a badge, and a thoughtfully designed social program should not be assumed equivalent. [10]

The book's discussion of an unsuccessful sales game is a useful counterweight. However, employees' measured buy-in was not randomly assigned, so differences between enthusiasts and nonenthusiasts do not by themselves prove that consent caused the different effects. People who like a program may already differ in motivation or circumstances. The ebook footnote also clarifies that the reported sales decline was only marginally significant. [1, ch. 2; ch. 2, footnotes]

Later evidence strengthens a bounded claim for some exercise programs. In the 2024 BE ACTIVE trial, gamification increased daily steps during a 12-month intervention; the combined gamification-and-incentive condition had the clearest sustained effect during follow-up. This was a structured program in selected higher-risk adults, not a test of arbitrary workplace games. It measured activity rather than cardiovascular events. [11]

**Revised takeaway:** make a repeated action more pleasant when the pairing does not interfere with performance or create another problem. Test whether the person actually uses and enjoys it. Evaluate a game by behavior, welfare, and continued participation—not by whether it contains points or badges.

## Chapter 3: Procrastination

#### What works about commitment

A commitment device changes future options or consequences before temptation arrives. The book distinguishes hard restrictions and financial stakes from softer pledges, discusses low uptake, and acknowledges that an emergency can make a penalty unfairly costly. It also cautions against excessive paternalism by managers. These are substantive safeguards. [1, ch. 3]

The evidence supports the existence of useful commitments. A Philippine smoking study found that offering a deposit-based commitment increased the probability of passing a cessation test by about three percentage points, with benefits also observed at twelve months; only 11% took up the offer. This is stronger than a testimonial because both the offer and the objective outcome were studied. It does not imply that every person who declines a deposit is mistaken. [14]

#### Savings: the numerator is not the whole result

The Green Bank study supports increased savings held at the participating bank. However, its roughly 411-peso treatment effect is scaled to **preintervention savings balances** in the paper. The book's illustration—if a control customer saved $100, the comparable treatment

customer saved $180—suggests a different denominator. It also risks turning a bank-account outcome into a statement about total saving or financial welfare. Those are not interchangeable. [1, ch. 3] [12]

A locked account could raise measured balances while reducing liquidity or shifting funds from somewhere else. The study's positive result remains meaningful, but evaluating welfare requires more than an increase in one account. Importantly, the book's footnote does acknowledge a marketing-only comparison group. It would be unfair to criticize the author for omitting that comparison altogether. [1, ch. 3, footnotes]

#### Deadlines: a retracted source, not merely an outcome error

The passage describing "about 50 percent fewer errors" misstates the task and outcome of the paper. Its reported task was to **find planted errors in supplied texts**, not write assignments with fewer errors. Even if its reported contrast were reliable, "50% more detected" would not imply "50% fewer left." The paper was retracted on 2 September 2026 and is retained here only to document the historical source relationship. Its figure is no longer treated as reliable affirmative evidence. [1, ch. 3] [13, 42]

A close preregistered replication of Study 2, published on 15 July 2026, did not reproduce the original deadline-performance results. The main study randomized 124 participants; original instructions were unavailable, and computerization, incentives, and the participant setting differed. The replication therefore tests a close implementation rather than an identical protocol. Its null performance findings are separate from the journal's decision to retract the original paper over data reliability. Neither finding invalidates independent commitment studies or proves that every deadline is ineffective. [43]

The publisher's full notice states that the authenticity and completeness of a dataset believed to belong to the original study could not be confirmed, and that the editor could no longer attest to the findings' reliability. Both authors agreed to retraction. This review gives the 2002 paper no affirmative evidentiary credit. The notice is dated 2 September 2026, long after the book's 2021 publication. It is not evidence that Milkman anticipated the retraction or acted deceptively. [42]

#### Harder penalties are not automatically the better intervention

The chapter's strongest overreach is its general ranking of soft commitments below hard ones. The relevant decision has at least two stages: whether a person adopts the arrangement and what happens after adoption. In a large smoking-cessation trial, reward programs were accepted by 90% of those offered them versus 13.7% for deposit programs. Six-month abstinence was higher under reward assignment, despite stronger conditional estimates for deposits among people who would accept either. [15]

This trial is **not** a direct comparison of public pledges with cash commitments. It nevertheless demonstrates why greater conditional pressure need not produce a better population intervention. Selection, affordability, liquidity, aversion to loss, monitoring reliability, and the possibility of circumstances changing all matter. The book's "naïf" versus "sophisticate" distinction should be treated as a modeling device, not a moral ranking of people who decline financial penalties.

The clinician-poster experiment offers qualified support for softer commitment. It randomized fourteen clinicians and found a large reduction in inappropriate prescribing. Yet a public poster also communicates with patients and can serve as a reminder; the experiment does not isolate guilt as the mechanism. It is evidence for a particular package, not a general law that making a pledge improves any behavior. [16]

A larger English cluster trial randomized 196 practice units and did not improve the primary overall antibiotic-dispensing outcome. Posters were delivered remotely, the content and engagement differed, attrition mattered, and total dispensing was not the original inappropriate-prescribing endpoint. The authors therefore warn against treating it as an exact replication. Favorable post-hoc phone-message analyses do not establish a primary poster benefit. The US result remains bounded positive evidence; the English trial limits confident scale-up. [50]

**Revised takeaway:** first remove avoidable friction and clarify the action. Add a voluntary commitment when a predictable temptation remains. Prefer a limited, reversible consequence tied to a controllable behavior; do not assume that bigger stakes are wiser or that nonadoption proves self-ignorance.

## Chapter 4: Forgetfulness

### One of the book's stronger chapters

Milkman distinguishes wanting to act from remembering to act and recommends timely reminders, cue-based plans, and checklists for complex sequences. She also notes that plans are less useful when the person has no interest in the goal and that attempting too many plans can become discouraging. These are important restrictions on what otherwise could become a simplistic "just schedule it" message. [1, ch. 4]

The chapter's broad label, "forgetfulness," is less precise than the behavioral phenomena it covers. Failing to complete an intended action can reflect memory, distraction, competing priorities, practical access, or a change of mind. A planning intervention may help through several of those routes. Its success does not identify forgetting as the original cause in every participant.

### Vaccination planning: the relative statistic checks out

In the workplace vaccination experiment, the control vaccination rate was 33.1%. A date-and-time planning prompt produced a 37.1% raw rate, a **4.0-percentage-point** difference; the adjusted estimate was **4.2 points**, with a 95% confidence interval of 0.5–7.8 points. Dividing 4.2 by 33.1 gives approximately 12.7%, which supports the book's rounded 13% relative improvement. The date-only prompt had a smaller, statistically uncertain adjusted effect. [17]

The useful inference is that a small addition to a reminder can improve follow-through when a convenient service already exists. It is not evidence that writing a plan solves vaccine hesitancy or lack of access. Milkman correctly explains that participants were not booking appointments and that researchers did not receive the written plans. [1, ch. 4]

### A later synthesis supports planning, but moderates the sales pitch

A meta-analysis of 642 tests reports benefits across several outcomes. Its average behavioral effect was modest, and its overall bias-adjusted model estimate was smaller still. The authors found stronger effects for if–then formats, motivated participants, and rehearsed plans. That supports cue-based planning as a useful low-cost technique, not a uniformly large effect on goal achievement. The paper also notes limits on updating its literature search; it is not a census of studies through its 2025 issue date. [18]

The chapter's historical forgetting percentages should also be read cautiously. Results from a specific memory paradigm do not establish that every adult loses a fixed fraction of all meaningful information after twenty minutes or a day. Prospective memory for an intended action is not interchangeable with retention of newly learned syllables. The useful recommendation—externalize important reminders—does not depend on presenting one historical curve as a universal daily-life rate. [1, ch. 4]

### Checklists require implementation, not just a list

The surgical-checklist example needs more qualification. The original eight-hospital study compared outcomes before and after a checklist implementation program: complications fell from 11.0% to 7.0%, and mortality from 1.5% to 0.8%. These are important associations, but the study was not a randomized isolation of a paper checklist. Training, communication, and practice changes accompanied implementation. [19]

An independent evaluation across 101 Ontario hospitals did not find statistically significant reductions in mortality or complications after adoption. That result was published in 2014, before this book. Differences in baseline care and implementation could matter; the Ontario result does not establish that checklists are useless. It does undermine a portable, context-free claim of 35–45% reductions whenever professionals use one. [1, ch. 4] [20]

A later stepped-wedge randomized rollout in five surgical clusters across two Norwegian hospitals found fewer complications and shorter stays. The overall mortality reduction was not statistically significant. This favorable result belongs beside the Ontario null. It supports checklist implementation in that setting, not a universal 35–45% benefit or an isolated effect of a list without changes to care. Access for this addition is the primary structured abstract, not a full independent statistical reproduction. [60]

**Revised takeaway:** link a desired action to an observable cue and remove practical obstacles to responding. Use a short checklist when a sequence genuinely strains memory. Measure omissions and completion quality; do not mistake the existence of a checklist for reliable execution.

## Chapter 5: Laziness

### A useful design principle under an imprecise label

This chapter presents defaults and habits as ways to make the desirable action the path of least resistance. Its practical orientation is strong: rather than demanding that a person win the same internal argument every day, change the arrangement in which the decision occurs. The word "laziness," however, groups together inertia, decision cost, limited attention, preference for efficiency, and learned routines. Those phenomena are related, not identical. [1, ch. 5]

The generic-prescribing example is a legitimate large effect. Prescriptions for eligible generic medications increased from **75.3% to 98.4%** after the electronic-record change, closely matching the book's rounded figures. The study was a before–after comparison within one health system, not a randomized demonstration that doctors were lazy. Its authors also note that pharmacies may already substitute generics, so changes in prescriptions need not translate one-for-one into downstream savings. [21]

The book's footnote deserves credit: it recognizes that defaults can convey recommendations and other signals, not only exploit inertia. Readers should not infer that the author's complete account attributes every default effect to lack of effort. [1, ch. 5, footnotes]

### A serious error in the opioid-default footnote

A footnote to chapter 5, "Laziness," says the nudge unit cut opioid prescriptions in half by defaulting to ten pills rather than a usual thirty-day dose. The linked Delgado study examined a transition from **no quantity default** to a ten-tablet default. It found no statistically significant change in mean quantity and relatively small median reductions. It did not demonstrate the claimed halving. The discrepancy concerns both the comparison and the magnitude; it should be corrected, not merely softened. [1, ch. 5, footnotes] [22]

This does not refute defaults as a class. It illustrates why a convincing general principle cannot substitute for verifying a specific example. A system change can strongly alter the distribution of selected quantities without producing the advertised reduction in total prescribing.

### Habit is not simply "repetition that continues"

The chapter sometimes moves between repeated behavior, skilled performance, cue-triggered automaticity, and the persistence of exercise after incentives cease. These can coexist, but they are not interchangeable measures. Continued gym attendance could reflect a new preference, better scheduling, learned information, or habit. The outcome alone does not identify which explanation is correct.

The Galla–Duckworth studies support associations among self-control, beneficial habits, and better outcomes. Their mediation results are informative, but do not establish that everyone who appears self-controlled lacks any superior capacity for resisting temptation. Milkman's categorical contrast between impressive willpower and habits is stronger than the studies' qualified proposal that habits are an important pathway. [1, ch. 5] [39]

The practical insight remains valuable: observing high performers only at their moment of action can hide the preparation that makes that action easier. But readers should avoid replacing one oversimplification—success is willpower—with another—success is only habit.

#### Flexible incentives do not invalidate consistent cues

The Google experiment is among the chapter's best examples. Participants in all conditions selected a preferred two-hour window and received reminders. Incentives varied in amount and in whether they required exercise within that window. Flexible incentives produced more persistent attendance; a comparison of different payment levels also addressed the possibility that one group simply exercised more during treatment. The book explains this design feature in a footnote. [23] [1, ch. 5, footnotes]

The appropriate inference concerns **how incentives reward completion**. It is not that consistent cues are useless, nor that deliberately varying the context strengthens every habit. A person can have a preferred time and still permit a fallback. The finding also does not directly demonstrate that the flexible group's behavior became more automatic.

**Revised takeaway:** automate an appropriate decision where possible. For actions that require effort, use a clear anchor but reward completion outside the ideal window as well. Track whether the behavior happens and how much effort it requires. Treat a broken streak as a recovery problem, not proof that a useful routine has vanished.

## Chapter 6: Confidence

#### The chapter with the largest gap between useful advice and evidentiary certainty

Milkman argues that self-doubt can prevent action, that advice-giving can improve the adviser's performance, and that expectations, growth mind-set, self-affirmation, and allowances for failure can support persistence. The chapter is right to distinguish underconfidence from overconfidence and to recognize that setbacks can disrupt pursuit of otherwise attainable goals. Its recommendations, however, are supported by markedly different kinds of evidence. [1, ch. 6]

A central causal problem runs through the chapter: success can increase confidence as well as confidence supporting action. People with better skills, stronger support, or fewer constraints may also have more justified confidence. Observing that confident people do better does not establish how much would change if confidence alone were raised. A useful intervention must be evaluated on its actual effect, not on the appeal of its proposed mediator.

#### Advice-giving: positive evidence, limited transfer

The large school experiment randomized **1,982 students** and found improved report-card grades in math and a self-selected target class over an academic quarter. It was preregistered, and the book explicitly describes the gains as small rather than miraculous. This is a good example of calibrated communication within an otherwise enthusiastic chapter. [24] [1, ch. 6]

It is not equivalent to showing that every underperforming employee should become a mentor or that advice is generally more useful to the giver than the receiver. The intervention could work through reflection, commitment, retrieval of strategies, or confidence. The grade outcome does not by itself isolate one of these explanations. Nor does a short advice exercise confer competence to advise others in consequential domains.

The practical recommendation is best applied where the person already has some relevant knowledge: invite them to articulate strategies, then examine which they can actually use. Solicited advice and collaborative problem-solving remain compatible with receiving instruction. A novice's knowledge gap should not be reframed as a confidence problem merely because empowering them sounds appealing.

#### Housekeepers: an intriguing small study, not a settled physiological mechanism

The exercise-mindset study involved **84 hotel workers across seven hotels**, with treatment assigned at the hotel level. It reported improvements in the informed group over four weeks. But seven randomized clusters provide less independent treatment information than eighty-four individually randomized participants, and unchanged exercise or diet was not established through continuous objective monitoring. [25]

Milkman first describes unchanged routines, then suggests increased vigor as an explanation. The latter is plausible, but it is not a measured demonstration. "Expectations affected behavior or experience" and "belief directly improved physiology without behavioral change" are different claims. The design cannot make all competing pathways disappear. The safest reading is that a small cluster study reported promising changes after reframing work as exercise; the mechanism and generalizability remain uncertain. [1, ch. 6]

A relevant occupational replication attempt was published in 2011. Stanforth and colleagues gave university building-service workers either an exercise message about their work or job-safety information. There were 53 initial participants; the reported three-time-point analysis used the 39 who completed all assessments. At four and eight weeks, the study did not reproduce a differential benefit for weight, BMI, or body fat, and it found no group-by-time benefit for waist circumference. Systolic and diastolic blood-pressure results favored the exercise-message group. The results therefore cannot fairly be described as all-null. [41]

This was not an exact repeat of the hotel-housekeeper experiment. It used a different worker population and comparison condition, and north and south campus work locations—not individual workers—were randomly assigned to the two conditions. There were only two location clusters, substantial attrition, and limited power to detect small effects. Both groups became more likely to describe themselves as regular exercisers, and activity was not objectively tracked throughout. The findings weaken confidence in a robust belief-alone weight or body-composition effect; they do not prove a zero effect or establish that the favorable blood-pressure contrast was caused independently of behavior. Because this paper preceded the 2021 book, the contrary evidence was available at publication. That chronology does not establish that Milkman knew about or deliberately ignored it. [41]

This review does not classify the finding as fraudulent, disproven, or impossible. It classifies the book's confidence as greater than the design warrants. A proper update would prioritize independently replicated, well-powered studies with objective activity measures and appropriate analysis of hotel-level assignment.

#### Milkshakes: an acute biomarker is not a life-outcome demonstration

In the cited crossover study, **46 participants** consumed the same 380-calorie drink under different labels, and the researchers observed different ghrelin responses. That is relevant evidence that expectations can accompany different physiological responses. It does not establish sustained weight loss, improved metabolic health, or a general rule that belief determines bodily outcomes. [26]

The corrected author-hosted paper reports an acute ghrelin pattern, not a significant hunger main effect or hunger interaction. Its 2011 correction repairs a mistaken reference, not its results, and the paper is not identified as retracted in the checked records. This audit did not locate an independent direct replication of the identical-drink label/ghrelin comparison. Related placebo-capsule work tests a different intervention and has mixed marker-specific findings; it is not that replication. [56, 57]

The chapter's transition from this laboratory result to broad statements about expectations should preserve the evidentiary distance. A mechanism can be interesting without having a demonstrated intervention payoff at the scale a self-help reader cares about. Conversely, requiring that distinction does not deny the possibility of psychologically mediated physiological effects.

#### Growth mind-set: modest conditional evidence, not "vastly more" achievement

The national experiment highlighted by the book found a roughly **0.10-grade-point** improvement among lower-achieving students and an increase in advanced mathematics enrollment from approximately **33% to 36%** in the analyzed enrollment sample. These are meaningful results for a brief intervention, especially at scale. They do not imply that the intervention transforms students generally or produces large gains in everyone. [27]

The chapter's accurate description of some study-specific results sits alongside broader language about much greater effort and achievement. Subsequent reviews make that mismatch more important. Macnamara and Burgoyne estimated a very small overall academic effect, with no clear benefit in their highest-quality subset. Burnette and colleagues reported a somewhat larger but still small academic effect in targeted, high-fidelity settings. These analyses differ in inclusion, coding, and the questions they emphasize; they are not interchangeable votes on a single universal proposition. [28, 29]

The balanced conclusion is neither "growth mindset is a proven transformation technology" nor "beliefs about learning never matter." A brief intervention may help some students in supportive circumstances. It should complement instruction and opportunity, not be used to explain away their absence. The evidence is much stronger for **small, conditional possibilities** than for sweeping claims about human potential.

#### Emergency reserves and self-affirmation: preserve the actual construct

Allowing a limited number of emergency exceptions is a practical way to avoid making a demanding goal brittle. The cited research supports preference and persistence benefits for some reserve-based goal structures. This review verified the study's direction and design summary, but did not independently reconstruct the book's exact CAPTCHA success percentages; those figures are therefore not certified in the numerical audit. [35] [1, ch. 6]

Self-affirmation should not be reduced to generic self-praise. The relevant literature often uses reflection on important values to reduce defensiveness under threat. That differs from simply telling oneself that success is assured. The book's broad summary—focus on experiences that make one proud—is a simplification of a more specific intervention family. [40] [1, ch. 6]

Large educational replications include both earlier successes and a later precise null. Hanselman and colleagues used earlier procedures and sites with a new cohort; for the focal Black and Hispanic students, including multiracial students, the later results did not show a benefit and ruled out gains larger than 0.10 GPA standard deviations. The proposed moderators did not adequately explain the difference. This limits confidence in reliably transferring a values-affirmation intervention to another school cohort. It does not establish that all values reflection fails or validate generic positive self-talk. [59]

Finally, the statement that underconfidence "can only" hinder success is too absolute. Uncertainty can prompt preparation, advice-seeking, or a better-calibrated choice of goal. Whether it is harmful depends on what the person believes, what the task requires, and what behavior follows. The appropriate objective is **accurate confidence sufficient for useful action**, not maximal confidence.

**Revised takeaway:** respond to setbacks with a concrete next attempt, useful feedback, and a proportionate expectation of improvement. Build confidence from practice and evidence. Use advice-giving and values reflection as possible aids, not substitutes for competence, resources, or reality testing.

## Chapter 7: Conformity

#### A notably candid account of social influence

This chapter explains both informational and social-approval reasons for following others. It encourages readers to seek useful examples, learn from peers, and use norms or visibility to support action. It also describes failed interventions, warns about coercion, and proposes giving people the chance to earn recognition rather than exposing them to unwanted shame. This combination is a substantial strength. [1, ch. 7]

The Air Force Academy story is especially valuable. The intervention did not simply correlate successful peers with successful students: researchers manipulated group composition in an attempt to improve outcomes and found harm among the students they intended to help.

Evidence suggested that people reorganized their social interactions rather than mixing as the designers expected. The book reports this failure clearly. [30]

### Peer effects are not a ready-made allocation policy

The lesson is deeper than "do not compare yourself with someone too successful." A policy changes the setting in which people choose friends, seek help, and interpret their position. A relationship observed under one assignment system may not survive when a designer optimizes the assignment using that relationship. In this case, the people affected by the intervention responded to it. [30]

That makes the chapter a useful caution against the book's own strongest engineering metaphor. Knowing that an obstacle or social association exists does not always tell the designer how to intervene. The system can change when acted upon, and a seemingly well-targeted intervention can create a new problem.

The retirement-saving experiment offers a parallel warning. Providing peer information reduced saving in a particular low-saving, nonparticipant group; the authors interpret the pattern as consistent with discouragement from upward comparison. The book fairly presents this as a backfire rather than an embarrassing result to omit. However, discouragement is not established as the uniquely possible pathway. [31] [1, ch. 7]

### Copying is a hypothesis-generation strategy, not proof of causation

The copy-paste study supports asking people to seek and adapt a strategy from an acquaintance: a preregistered experiment found more exercise the following week. That is a useful, relatively low-cost intervention. It is also a short-horizon result, not a guarantee that imitating a successful person's routine reproduces their success. [32]

Successful people have many visible behaviors. Some contribute to success, some are incidental, and some are affordable only because success has already occurred. The reader's task is to identify a transferable action that fits their constraints—not to treat every observed routine as an active ingredient. A method that works for a colleague with flexible hours may fail for someone with a fixed shift and caregiving duties.

### Norm messages: verify the norm and monitor more than compliance

The hotel-towel study supports the direction of the claim that descriptive norms can influence conservation. Its abstract also reports stronger effects for a locally matched reference group. This review did not reconstruct every contrast underlying the book's 18% and 33% figures, so those exact numbers are not included among the certified calculations. [33] [1, ch. 7]

Two German hotel field tests did not reproduce consistent normative-message superiority over an environmental appeal or a stable same-room advantage. High baseline reuse, floor/week allocation and a small second sample constrain interpretation. Their no-message comparison was a preceding baseline week, not a concurrent randomized control. The paper's

correction restored altered Greek characters; it did not change effect estimates or retract the work. The original hotel finding remains an original result, not a reliably portable estimate. [51, 52]

Dynamic norms—information about a behavior becoming more common—also have direct experimental support. But an upward trend is not a promise that a behavior will become universal. The message must be accurate, relevant, and not cherry-picked to manufacture social pressure. [34] [1, ch. 7]

Later British online work with 846 participants did not reproduce dynamic-norm effects on motivational outcomes. A preregistered 2024 online study with 1,294 participants did not establish a main dynamic-over-static advantage in initial motivational outcomes or one-week changes. Its consumption-change contrast moderately favored the null under its registered prior, while intentions and some interaction contrasts were inconclusive. An initial visual-cue interaction did not demonstrate a dynamic-norm advantage when visual cues were present. These outcomes are intentions and reported consumption, not observed café purchases. A 2026 synthesis of 79 studies in 49 papers reports very small, variable average effects. Full-text access to that synthesis was unavailable; no unreported moderator estimates are inferred. [53, 54, 55]

Social visibility creates a further trade-off. A program can improve a measured action while making participants resentful, anxious, or less trusting. The book recognizes this in its discussion of voting-pressure mailings. For institutional use, the appropriate evaluation includes consent, privacy, complaints, avoidance, and the distribution of burdens—not just the target action. [1, ch. 7]

**Revised takeaway:** use relevant peers to discover workable strategies and make genuine participation visible with consent. Do not fabricate norms, infer that social proximity alone produces improvement, or treat compliance as a complete measure of benefit.

## Chapter 8: Changing for Good

#### The book's most important qualification arrives at the end

The conclusion acknowledges the difficulty of persistence. Milkman and Duckworth disagree over whether a large gym study should count as a success: one emphasizes short-term improvement; the other, the limited carry-over once the program stops. Milkman uses the conversation to argue that recurring barriers may require ongoing treatment. She also encourages people to revise the route to a goal when a particular strategy remains unpleasant or ineffective. [1, ch. 8]

That is more nuanced than a claim that readers merely need one clever nudge. It also prevents the book from being fairly summarized as promising effortless permanent change. A critical review should recognize that the author herself supplies the central caveat that many readers will need most.

### The published megastudy supports both sides of the conversation

The subsequent Nature publication involved **61,293 gym members** and **54 four-week programs**. Approximately 45% of interventions significantly increased weekly visits during intervention, whereas only 8% generated detectable significant changes afterward. The first number supports optimism about finding useful tools; the second supports restraint about durable transformation. This is publication of the project discussed in the book, **not an independent replication** of it. [36]

The proportion of interventions attaining statistical significance is not the proportion of people who formed a habit. Nor is it a precise estimate of how many strategies "really work": it depends on effect size, uncertainty, the comparison group, and the decision threshold. Programs can have small effects that are not detected, while selecting the largest observed effect can exaggerate what a future deployment will achieve.

### Persistence and continued treatment are both legitimate outcomes

The energy-report study is informative because it includes discontinuation. After reports stopped following two years of exposure, part of the conservation effect remained and part decayed, at roughly 10–20% of the intervention effect per year. Continuing reports maintained more conservation. The book uses this accurately, provided "effect decay" is not misread as a 10–20% annual rise in household energy consumption. [37]

Neither complete permanence nor continuous effort should be the universal standard. A payroll default may persist because the system keeps executing it. A checklist must be used each time because each procedure is new. An acquired skill may persist without a reminder. The relevant questions are what maintains the behavior, what maintaining it costs, and what happens when that support is removed.

### Where the analogy overreaches

Describing ordinary human tendencies as symptoms of a chronic disease is a metaphor, not a medical finding. It can usefully discourage premature withdrawal of support. It can also imply that vigilance must be endless even when the goal, context, or intervention burden should be reconsidered. Readers should keep the author's final invitation to reassess goals as prominent as the instruction to maintain techniques. [1, ch. 8]

**Revised takeaway:** budget for maintenance, monitor whether support is still needed, and revise the arrangement when its costs exceed its benefits. Judge success by sustained useful outcomes, not by whether the action has become completely effortless or whether an intervention can be stopped forever.

# 06 Claim, number, and source audit

## Numerical and source-support audit

### Reading the audit

The twenty entries below are a targeted stress test, not a random sample. Each identifies a specific passage, a supporting primary source, and a bounded assessment. Earlier item grades are expressed as qualitative source-relationship labels rather than summed into a misleadingly comparable percentage. The chapter discussions supply the context needed to interpret each entry; no error-prevalence estimate is made.

The audit deliberately treats several apparent discrepancies as **not errors**. The initial temptation-bundling percentage can be reconstructed from raw means; the flu-shot result uses an adjusted estimate; the generic-prescribing numbers are fair rounding. Conversely, an incorrect endpoint cannot be repaired merely because the paper supports a broadly similar practical idea.

**A01** **Faithful at the stated scope**

### Fresh-start savings invitations

Book: ch. 1 · Primary source: [5]

The 20–30% relative improvement is broadly consistent with the reported take-up and contribution outcomes. It should be paired with absolute changes.

**Assessment:** Relative framing, not an arithmetic error.

---

**A02** **Not supported by the cited source**

### Calorie labelling does essentially nothing

Book: ch. 1 · Primary source: [3]

The cited study reports a 6% decrease in calories purchased per transaction, mainly from food. It does not support a blanket near-zero conclusion about consumption.

**Assessment:** Direct mismatch; purchasing is not consumption.

**A03** **Substantial support problem**

**“Carrot-eater” labels change our actions**

Book: ch. 1 · Primary source: [6]

Children judged the stability of a hypothetical person’s characteristics. Their own eating behavior was not the outcome.

**Assessment:** Wrong outcome used as support.

**A04** **Faithful at the stated scope**

**Initial 55% exercise increase with bundling**

Book: ch. 2 · Primary source: [8]

Table 2 reports 1.16 versus 0.75 weekly visits: a 54.7% raw relative difference. The book acknowledges subsequent decay.

**Assessment:** Defensible rounding; first-week effect only.

**A05** **Partial or overextended support**

**Teaching bundling yields lasting behavior change**

Book: ch. 2 · Primary source: [9]

The study supports some benefit, but includes free audiobooks and a broader program. The no-audiobook comparison used random assignment with unequal, changing probabilities; the authors call its unpreregistered, weighted pooled analysis quasi-experimental. The incremental randomized teaching effects are more modest.

**Assessment:** Package/design compression.

## A06

**Partial or overextended support**

### Locked savings accounts produce 80% more saving

Book: ch. 3 · Primary source: [12]

The positive bank-balance result is genuine. The paper scales the approximately 411-peso effect to preintervention balances; the book's $100-versus-$180 gloss is not the same estimand.

**Assessment:** Denominator and bank-versus-household scope.

## A07

**Substantial support problem**

### Self-set deadlines yield 50% fewer errors

Book: ch. 3 · Primary source: [13]

The reported proofreading endpoint was errors detected, not errors made in authored work. That source-description problem remains. More importantly, the paper is now **retracted** and receives no affirmative supporting credit. A close 2026 replication did not reproduce its performance results, with protocol differences disclosed. [42, 43]

**Assessment:** Incorrect endpoint and percentage interpretation; the underlying source is retracted and cannot support the claimed benefit.

## A08

**Faithful at the stated scope**

### Cash commitments help smokers quit

Book: ch. 3 · Primary source: [14]

Random offer increased six-month biochemically verified cessation by about 3 percentage points, with an effect at 12 months. The book reports low uptake.

**Assessment:** Strong support for the bounded direction of effect.

## A09

**Mostly faithful; qualification needed**

### Public clinician pledges cut inappropriate prescribing

Book: ch. 3 · Primary source: [16]

A 14-clinician trial supports a substantial reduction. The intervention was a displayed poster package; it does not isolate private guilt or prove universal pledge efficacy.

The larger English poster trial had no primary overall-dispensing benefit, with different delivery, engagement and measurement. It limits scale-up rather than identically retesting the US inappropriate-prescribing outcome. [50]

**Assessment:** Positive trial; small number of clinicians and bundled mechanism.

## A10

**Faithful at the stated scope**

### Planning prompts increase flu vaccinations by 13%

Book: ch. 4 · Primary source: [17]

The adjusted 4.2-percentage-point effect against a 33.1% baseline is approximately 12.7% relatively. The book distinguishes planning from booking an appointment.

**Assessment:** Accurate relative effect.

## A11

**Partial or overextended support**

### Surgical checklists cut complications/mortality by 35–45%

Book: ch. 4 · Primary source: [19]

The original before–after study reports large reductions, but cannot warrant a general causal effect of that size. Independent Ontario findings also limit generalization.

The Norwegian randomized rollout adds favorable complication and stay-length evidence, without significant overall mortality improvement. Original, Ontario and Norwegian results are implementation- and outcome-specific. [60]

**Assessment:** Original association generalized too strongly.

**A12** **Mostly faithful; qualification needed**

**Generic prescribing rises from 75% to 98%**

Book: ch. 5 · Primary source: [21]

The reported study rates are 75.3% and 98.4%. A large within-system observational change supports the example, not a universal effect or an identified laziness mechanism.

**Assessment:** Accurate numbers; observational setting.

**A13** **Not supported by the cited source**

**Opioid prescriptions cut in half by a ten-pill default**

Book: chapter 5, “Laziness,” footnote on opioid prescribing · Primary source: [22]

The cited study reports no significant change in mean quantity and only modest median changes. Its prior system had no quantity default, not a standard thirty-day dose.

**Assessment:** Direct conflict with the linked study.

**A14** **Mostly faithful; qualification needed**

**Flexible incentives create more persistent exercise**

Book: chapter 5, “Laziness,” main text and footnote on incentive amounts · Primary source: [23]

The incentive trial supports more persistent attendance under flexible rewards. It does not directly measure automaticity or establish superiority for every kind of habit.

**Assessment:** Good policy evidence; broader habit language needs qualification.

**A15** **Faithful at the stated scope**

**Giving advice improves students’ grades**

Book: ch. 6 · Primary source: [24]

A preregistered randomized study with 1,982 students found better math and target-class grades over a quarter. The book explicitly calls the improvements small.

**Assessment:** Accurate bounded trial description.

## A16

**Partial or overextended support**

### Housekeeper expectations improve health without changed routines

Book: ch. 6 · Primary source: [25]

The original study reports the improvements, but randomization involved seven hotels and unchanged activity was not established with continuous objective measurement.

The 2011 occupational replication attempt did not reproduce differential weight, BMI, or body-fat benefits; waist circumference also showed no group-by-time benefit. Blood-pressure results favored the exercise-message group. Its different population, two-location assignment, attrition, and lack of objective activity tracking limit the inference. This mixed result strengthens the existing caution about mechanism and generalizability; it does not establish that every original result was false. [41]

**Assessment:** Tentative evidence narrated too confidently.

## A17

**Mostly faithful; qualification needed**

### Milkshake expectations alter ghrelin response

Book: ch. 6 · Primary source: [26]

The small crossover study supports a difference in the acute measured ghrelin response to differently labelled identical drinks. It is not a long-term health-outcome test.

The published correction changes a reference, not the result. A direct independent label/ghrelin replication was not established by the source audit; related placebo-appetite trials should not be counted as exact repeats. [56, 57]

**Assessment:** Reasonably faithful; narrow biomarker evidence.

## A18

**Partial or overextended support**

### Growth mind-set produces broadly much greater achievement

Book: ch. 6 · Primary source: [27]

The large national experiment supports a modest effect for lower-achieving students and an advanced-math enrollment effect, not the surrounding sweeping magnitude language.

**Assessment:** Narrow result extended beyond its demonstrated scope.

## A19

**Faithful at the stated scope**

### Engineered peer groups can backfire

Book: ch. 7 · Primary source: [30]

The academy experiment found a negative effect on the intended beneficiaries and evidence of self-segregation. Milkman explains the failure rather than hiding it.

**Assessment:** Strongly supported cautionary example.

## A20

**Faithful at the stated scope**

### Energy-report effects decay after discontinuation

Book: ch. 8 · Primary source: [37]

The primary research reports persistence alongside roughly 10–20% annual decay of the intervention effect after reports stop.

**Assessment:** Accurate durability caveat; not 10–20% growth in household energy use.

### Reproducible calculations

| Item | Calculation from displayed or reported values | Interpretation |
|---|---|---|
| First-week bundling | (1.16 − 0.75) ÷ 0.75 × 100 = **54.7%** | Supports the book's rounded 55%; raw first-week gym visits, not lifetime exercise. |
| Flu-shot planning | 4.2 ÷ 33.1 × 100 = **12.7%** | Supports the rounded 13%; adjusted effect divided by the comparison baseline. |
| Generic prescribing | 98.4 − 75.3 = **23.1 percentage points** | Not a 23.1% relative increase; the book's 75%→98% rounding is fair. |
| Fresh-start take-up | 2.3 ÷ 7.6 × 100 = **30.3%** | Illustrates why a large relative change can be only a few percentage points. |
| Baseball example | (.294 − .246) ÷ .246 × 100 = **19.5%** | The book says 29%; its own displayed numbers do not support that arithmetic. |

Sources for the first four rows are the corresponding primary reports; the baseball row is an internal arithmetic check, not an independently reconstructed baseball dataset. [8, 17, 21, 5] [1, ch. 1]

### What was not numerically certified

This review does not certify every numerical statement in the book. In particular, it did not reconstruct the exact contrasts behind the towel-reuse percentages, CAPTCHA success rates, the promotional piano-staircase statistic, individual biographical outcomes, or every survey

and historical prevalence estimate. Where only a primary abstract was retrieved, the bibliography says so. A reproduced abstract is enough to establish that a study measured a different outcome, but not enough to independently audit its full statistical analysis.

The book's bibliography is generally traceable: claims frequently lead to identifiable research rather than uncited authority. The audit nevertheless shows why traceability and accuracy are different virtues. A real DOI can point to a study that answers a different question or contradicts the numerical claim made beside it.

## 07 Practical use, limitations, and safeguards

### Practical Value · 75%

The assessment concerns the book's ordinary behavior-change applications, not its effectiveness as medical treatment. The strongest evidence concerns implemented techniques; several benefits are modest, setting-dependent or supported only while an intervention remains in place.

Practical Value is 75% because multiple useful interventions have direct evidence, the book treats maintenance seriously, and many uses have reasonable burdens. It is not higher because matching, transfer, and net benefit are not established for every technique or reader. Criterion-by-criterion rationale.

**Ease of application — not included in the score:** Fairly easy to begin; some strategies need continuing support.

The implementation suggestions that follow are the reviewer's synthesis, not evidence that the book already supplies every safeguard or that this review has validated a replacement program. The score evaluates the book itself, not the reviewer's improved version of its advice.

### What a more rigorous reading changes

#### 1. Separate "the intervention worked" from "we know why"

Consider an intervention that provides a free audiobook, a planning exercise, reminder messages, and incentives. A positive result identifies the package's effect relative to its comparator. It does not assign a percentage of that effect to enjoyment, memory, commitment, or self-image. Similarly, an advice-giving exercise can improve performance without establishing confidence as its sole mediator. [9, 24]

Mechanism claims need evidence that can distinguish competing pathways. Randomizing a mediator, manipulating intervention components, checking the temporal sequence, and addressing mediator–outcome confounding can help. A statistically significant indirect association alone cannot make the causal explanation unique. These are standards for interpreting evidence, not reasons to abandon interventions with useful observed effects.

For practitioners, the implication is operational: reproduce the features supported by the study before claiming that an attractive mechanism is sufficient. Removing reminders or simplifying a multi-part program may save cost, but it is a new intervention requiring its own test.

### 2. Separate the behavior from the outcome that ultimately matters

Gym entry is not fitness. Bank balance is not net wealth. A prescription written as generic is not necessarily a different drug dispensed. A hormone response is not long-term health. Each may be a useful intermediate outcome, but the distance to the ultimate objective should remain visible. [8, 12, 21, 26]

A program can therefore succeed on its measured endpoint without establishing a net benefit. This is particularly important when the intervention creates an offsetting cost: a deposit commitment can reduce liquidity, a public leaderboard can impose unwanted scrutiny, and a study plan can displace another important task. Some trade-offs are acceptable. They should be evaluated rather than hidden by the label "good behavior."

### 3. Distinguish average effects from personalized treatment selection

The book's target audience may find the matching framework especially appealing because it promises a way to choose among many tools. But the evidence needed for personalization is demanding. A moderator must predict a **difference between intervention effects**, not just identify a group that tends to struggle. It must also be measured reliably enough to guide a real decision.

A useful diagnostic aid would need test–retest reliability, agreement between users or assessors where appropriate, clear behavioral definitions, and evidence that its assignments improve outcomes beyond a good general strategy. A useful adaptive system would additionally need a rule for when to switch strategies. None of these properties follows from having named seven plausible obstacles.

This is a criticism of the evidentiary bridge, not of the aspiration. The next empirical step would be an independently evaluated matching policy with a prespecified comparator and burden-equated treatment options. Until then, treat the categories as prompts to investigate, not as validated psychological types.

### 4. Distinguish a mature evidence base from a sequence of attractive results

A large field experiment is more informative than an anecdote, but it is not the end of the inquiry. Independent replication, varying settings, transparent analysis plans, and realistic implementation matter. A finding can also be too small to make an expensive program worthwhile, even when statistically detectable.

Evidence from two government nudge units illustrates the scale issue without implying that all nudges are ineffective. DellaVigna and Linos examined 126 randomized trials covering about 23 million people and reported an average effect around **1.4 percentage points**, compared with **8.7 points** in their academic-journal comparison sample. Their published

model attributed much of the gap to selective publication and design differences. These figures describe the studied communication interventions, not every default, financial incentive, or behavior-change technique. [38]

The implication is a realistic prior: an easy intervention can be worthwhile at scale even when its effect is much smaller than a memorable published example. Do not budget on the assumption that the most vivid result in the book will be your expected outcome.

#### 5. Take the book's counterexamples seriously

Milkman does not merely present triumphs. The lost temptation-bundling effect, an unsuccessful sales game, failed engineered peer groups, discouraging social comparisons, and weak post-program gym effects are part of the book's actual argument. The review's criticisms should not erase that record. [1, ch. 2; ch. 7; ch. 8]

The problem is subtler: caveats are sometimes carefully developed in one passage while neighboring summaries return to stronger, more general formulations. A reader who remembers only the end-of-chapter takeaway may retain more certainty than the supporting narrative warrants. The editorial fix is to put the conditions **inside** the takeaway, not merely elsewhere in the chapter.

## Practical value, difficulty, and safeguards

#### Practical strengths and constraints

**Usable actions.** The book supplies concrete moves: choose a cue, change a default, pair an activity with an enjoyable accompaniment, create a fallback, or ask a relevant peer for a strategy. Many can be tried without specialist knowledge. The limitation is that the organizing categories overlap and do not reliably tell a reader which intervention will be best. [1, chs. 1–7]

**Benefit relative to burden.** Several techniques have positive real-world evidence, and small low-cost effects can be worth obtaining. But implementation often requires infrastructure, repeated attention, coordination, or incentives. Reading an account of a successful program is not the same as receiving that program. The best-supported recommendation is selective, outcome-checked use rather than installing the entire toolkit at once. [17, 9, 11, 37]

**Safeguards and autonomy.** Milkman warns about rigid routines, excessive planning, coerced games, financial penalties, and social pressure. Yet the book could more consistently distinguish the person's own goal from an institution's preferred behavior, and could offer clearer stop rules for interventions that impose distress or cost. The chapter-three enthusiasm for hard commitments and the chapter-six confidence rhetoric particularly need such boundaries. [1, chs. 2–4; ch. 6; ch. 7]

**Difficulty:** fairly easy to begin; variable, sometimes substantial, to maintain. A calendar cue may take minutes to establish. A workable accountability relationship, redesigned work process, or continuing program can require coordination and upkeep. The book is useful

precisely because some solutions reduce repeated effort, but "low marginal cost" should not be confused with no setup cost or no burden.

### A decision rule for using the book

The following is **this review's synthesis**, not a newly validated protocol and not a claim that Milkman tested this exact sequence.

Start by defining one observable action and its purpose. "Get healthy," "be confident," and "stop being lazy" are not useful measurement rules. "Begin a twenty-minute walk after lunch on four workdays" is observable, although the target itself still needs to suit the person's health and circumstances. The desired action should serve a goal they actually endorse.

Next, examine a few real occasions when the action did not happen. Was it unavailable? Was the needed skill missing? Was a competing obligation more important? Was the cue not noticed? Was the immediate experience unpleasant? This avoids treating every failure as an internal motivation defect. A good explanation should distinguish the failed occasion from occasions when the action did occur.

Then choose the least burdensome plausible intervention. Improve access or remove a step before adding penalties. Use a cue when action is wanted but not initiated. Improve enjoyment when the activity is needlessly aversive. Add a fallback when the ideal time often fails. Seek instruction when the problem is skill rather than motivation. Reserve social pressure and financial commitments for situations where their costs and consent are clear.

Finally, measure the target behavior and one or two costs over a prespecified period. Did completion improve? Did quality deteriorate? Did the arrangement create avoidance, anxiety, expense, or conflict? A personal experiment is not a randomized clinical trial, so do not overinterpret a few observations; nevertheless, explicit tracking is more informative than merely asking whether a strategy sounds persuasive. Retain useful support and change what does not earn its burden.

### Four safeguards that belong beside the advice

**Do not make financial failure more damaging than behavioral failure.** A commitment should not threaten essentials or punish events the person cannot control. The relevant goal is a helpful incentive, not maximum pain. Distinguish a self-imposed arrangement from pressure exerted by someone with authority.

**Do not turn a streak into a verdict.** A tracking tool should make recovery easier. When it instead encourages concealment, unsafe compensatory behavior, or abandonment after one miss, change the metric or the rule. Completion counts can be useful; they are not measures of personal worth.

**Do not substitute confidence for feedback.** Encouragement should coexist with accurate information about performance, uncertainty, and constraints. A person can believe improvement is possible while acknowledging that a specific target, timeline, or strategy is unrealistic.

**Do not substitute behavioral technique for appropriate professional care or structural change.** The book does not establish a treatment for a medical disorder, a comprehensive financial plan, or a remedy for impossible workloads. Use behavioral tools to support an appropriate substantive plan, not to avoid obtaining one.

## 08 What changes with time—and what should change in the book

### What was knowable in 2021, and what changed afterward

A fair review must not penalize a 2021 author for failing to cite research that did not yet exist. It must also not excuse an original source mismatch merely because the field later became more skeptical. The table separates those situations.

| Evidence timing | Finding | Consequence for this review |
|---|---|---|
| **Available before publication** | The 2011 calorie-posting article reported lower calories purchased. | The book's source mismatch was already present; it is not a hindsight criticism. [3] |
| **Available before publication** | The 2011 occupational replication attempt did not reproduce differential weight or body-fat benefits, although blood-pressure results favored the exercise-message group. | The housekeeper example needed a replication and generalizability qualification in 2021. The 2011 paper predates the book; it is not newly published research or proof that every original result was false. [41] |
| **Available before publication** | The 2016 voter-label field study did not reproduce the noun-over-verb advantage in its setting. | The label example needed a generalizability caveat in 2021. [7] |
| **Available before publication** | The 2014 Ontario checklist study did not reproduce large outcome improvements. | A universal checklist-effect claim was already too broad. [20] |
| **Available before publication** | Reward-versus-deposit cessation trials showed a major adoption trade-off. | Stronger conditional incentives were not equivalent to better population impact. [15] |
| **Available before publication** | The 2018 opioid-default paper did not report the claimed halving. | A direct reference error in the supplied edition, independent of later literature. [22] |
| **Published later in 2021** | The complete gym megastudy formalized the modest and uneven durability discussed in the book. | Supports the final chapter; it is the same research program, not an independent replication. [36] |
| **2022 publication** | Two nudge units produced smaller average effects than the academic comparison sample. | Strengthens the case for realistic implementation priors; not a universal verdict on all behavioral tools. [38] |
| **2023 syntheses** | Growth-mindset meta-analyses differed, with small, quality- and context-sensitive academic effects. | Weakens broad magnitude claims; does not establish that every intervention is ineffective. [28, 29] |
| **2024 trial** | BE ACTIVE found sustained activity gains during a year-long program, with the clearest follow-up evidence for the combined intervention. | Positive update for selected gamified/incentive programs, not for arbitrary games. [11] |
| **Online 2024 / issue 2025** | A larger planning synthesis reported benefits and publication-bias sensitivity. | Supports planning while lowering confidence in uniformly large effects. [18] |

| Evidence timing | Finding | Consequence for this review |
|---|---|---|
| **2025 synthesis** | Cochrane found small effects on calories selected/purchased and less certain effects on consumption. | Reinforces the need to replace “essentially no effect” with a measured, endpoint-specific account. [4] |
| Published after the 2021 book | Close deadline replication in July 2026; formal retraction on 2 September 2026. | Withdraw reliance on the original paper. These 2026 developments occurred after the book’s publication. They do not establish what the book’s author knew in 2021. [42, 43] |
| Evidence published before the book | 2017 medication reminders and affirmation replication; 2014 hotel tests; 2015 Norwegian checklist publication; 2020 English poster trial and presidential-election report. | Add contrary and favorable boundaries, keeping outcomes, original protocols, later versions, and comparisons distinct. [58, 59, 51, 60, 50, 48] |
| 2023–2026 corrections and synthesis | Voter-table sign correction, broad noun-language synthesis, and later dynamic-norm studies. | Use corrected versions and small/mixed findings without declaring every identity or norm intervention false. [45, 49, 54, 55] |

Later evidence changes a current recommendation, not the historical facts about what an author could have known. Similarly, several later publications involve Milkman or the same collaborating research programs. Their larger samples and improved designs can increase confidence, but should not be counted automatically as independent replication by unrelated teams.

## Final assessment and editorial corrections

### What should be retained

Retain the problem-solving orientation, the distinction between wanting and doing, the concrete accounts of field experiments, and the emphasis on maintenance. Retain the counterexamples and the willingness to describe interventions that failed. Retain the idea that a person may need to change the route to an endorsed goal rather than intensify commitment to an unpleasant or unworkable route. These are central parts of the book, not merely favorable interpretations imposed on it. [1, Introduction; ch. 4; ch. 7; ch. 8]

### What should be corrected or qualified

The clearest original source-description corrections are the opioid-default claim, the carrot-eater outcome, the calorie-labelling characterization, the deadline-study endpoint and percentage, and the baseball arithmetic. The deadline source now also requires an explicit **retracted** status and withdrawal of evidentiary reliance. These are different kinds of problem: the small baseball arithmetic error, medically important opioid mismatch, and later deadline-data reliability finding should not be treated as equivalent failures. [22, 6, 3, 13, 42]

The most important conceptual revisions are different: mark the seven-obstacle matching framework as a design heuristic; distinguish intervention effects from mechanisms; report absolute changes and follow-up horizons beside striking relative percentages; and distinguish persistent performance from automatic habit. None of these requires making the book dull or stripping out practical advice. They require making the claim match the evidence.

### The judgment

**How to Change is worth reading. It is not worth treating as a scientific authority that has already resolved how to select the best intervention for every person and problem.** Its useful components survive a critical reading. Its weakest passages become much less troublesome once their status is clear: a specific error to correct, a plausible mechanism not yet isolated, a small study requiring replication, or an intervention whose benefit depends on context.

The strongest way to use the book is also the most consistent with its best passages: identify a concrete problem, make a proportionate change, observe what happens, and keep what works. The weakest way is to memorize a collection of named effects and assume that knowing the name supplies the cause, the expected effect size, and the right prescription.

## 09 How the ratings were determined

The headline scores summarize three different editorial judgments. **Scientific Accuracy** evaluates three central propositions; **Reference Accuracy** evaluates traceability, description and inference in the examined evidence; **Practical Value** evaluates likely benefits, applicability and net value. Ease of application is reported separately and is not included in the average.

The source audits are not identical probability samples. Reference Accuracy therefore uses the same three anchored criteria in every review—not a percentage of references that passed an audit. Both faithful and problematic source use inform the assessment, and inaccessible passages are not counted as errors. See Sections 02, 06 and 10 for scope and coverage.

**Overall rating: 72%.** The three categories receive equal weight. This is an editorial convention, not an empirically validated estimate of overall truth or benefit. The percentage does not override the recommendation or the consequential caution on the cover.

### Scientific Accuracy · 75%

Do the book's main claims hold up?

#### Matching a strategy to the obstacle beats an all-purpose approach.

A useful design heuristic, but the seven-obstacle taxonomy is not itself a validated diagnostic or treatment-matching instrument. The relevant comparison is matched treatment against a credible alternative, not merely intervention against no intervention. [Section 04]

#### Behavioral tools can improve concrete action.

Randomized field interventions support this carefully bounded existence claim. The grade does not extend to every technique, every population, or a promise of lasting individual transformation. [Section 04]

**Maintaining change often requires continuing support and adaptation.**

Maintenance is an important corrective to one-shot self-help. Persistence after intervention and the need for continued support vary by behavior; neither permanent support nor permanent automaticity is established universally. [Section 04]

## Reference Accuracy · 67%

Do the cited sources support what the book says?

**Traceability**

The checked claims generally lead to identifiable named studies or notes; traceability does not redeem an inaccurate summary. [Section 06]

**Accurate description**

The opioid, carrot-label, calorie, and baseball examples require consequential outcome or numerical corrections. [Section 06]

**Appropriate inference**

The account sometimes carries narrow findings into mechanisms or general prescriptions not identified by the linked studies. [Section 06]

## Practical Value · 75%

Is the advice likely to be worthwhile for its intended reader?

The assessment concerns the book's ordinary behavior-change applications, not its effectiveness as medical treatment. The strongest evidence concerns implemented techniques; several benefits are modest, setting-dependent or supported only while an intervention remains in place.

**Evidence of intended benefit**

Several consequential recommendations have experimental support for concrete behavior, including planning prompts, incentive arrangements and commitment offers. Credit is for these bounded interventions, not for an established effect of reading the book or a guaranteed personal transformation. Variation among techniques and the distance from intermediate behavior to ultimate welfare rule out the highest grade. [17; 14; 15; 36]

#### Applicability and durability

The book explicitly recognizes context, failure and maintenance, and its closing chapter credits continuing support rather than promising universal permanence. Its useful advice can often be adapted to an ordinary goal. However, the seven-obstacle matching system is not validated, and some favorable field results depend on infrastructure or bundled support a reader may not have. [1; 9; 23; 36]

#### Benefits relative to burdens and risks

Many interventions are inexpensive and reversible, and the text gives meaningful cautions about coercion, financial commitments, rigid routines and social comparison. These make selective use reasonably defensible. Stronger stopping rules, attention to who chose the goal, and accounting for liquidity and ongoing burden would improve the advice. [1; 12; 15; 31]

**Practical Value is 75% because multiple useful interventions have direct evidence, the book treats maintenance seriously, and many uses have reasonable burdens. It is not higher because matching, transfer, and net benefit are not established for every technique or reader.**

## Calculation and scoring anchors

The full audit uses five anchored grades, from 0 to 4. These are the inputs behind the percentages, not an additional reader-facing rating system. At a given criterion, 0 denotes a demonstrated fundamental failure or contradiction; 1 major limitations; 2 partial or mixed support; 3 substantial support with qualifications; and 4 strong support at the defined scope. The series methodology gives category-specific anchors and missing-data rules.

**Scientific Accuracy:** (2 + 4 + 3) ÷ 12 × 100 = 75.000% ➜ **75%**.

**Reference Accuracy:** (4 + 2 + 2) ÷ 12 × 100 = 66.667% ➜ **67%**.

**Practical Value:** (3 + 3 + 3) ÷ 12 × 100 = 75.000% ➜ **75%**.

**Overall:** (9 + 8 + 9) ÷ 36 × 100 = 72.222% ➜ **72%**. This is exactly the average of the three *unrounded* category scores. Only the final displayed values are rounded.

**Uncertainty.** A one-grade disagreement on a single criterion changes its category by about 8.3 percentage points and the overall rating by about 2.8 points before rounding. This is a sensitivity calculation, not a confidence interval. Independent duplicate scoring and measured inter-rater reliability are not documented. Small differences between books should not be interpreted as precise scientific rankings.

**Missing evidence.** No category in this edition is withheld as unrateable, but some individual source passages remain unresolved. Those gaps retain their explicit access labels; they are not zeroes or hidden failed checks. If a whole required criterion could not responsibly be judged, the category and overall score would be withheld rather than silently changing the denominator.

## Housekeeper evidence: scoring rationale

The 2011 replication is part of the adverse inference case. Appropriate inference is **2 / 4 (Mixed)**: consequential overreach coexists with substantial faithful material, including bounded planning results, cash-commitment evidence, and candid reporting of failed peer-group engineering and fading effects. The housekeeper mechanism is tentative and overextended. The replication does not clearly place this whole criterion at the **1 / 4 (Major limitations)** anchor. This is a criterion-level judgment under the methodology, not a fixed deduction for one unfavorable finding.

The housekeeper illustration does not test or contradict the three selected scientific propositions at their stated scope, and it is not the basis for the book's ordinary behavior-change Practical Value assessment. The original source is traceable, and contrary results do not by themselves make the original result description false. The method does not introduce a separate subscore, hidden cap, post hoc central-claim replacement, or duplicate penalty. The nine inputs are **2, 4, 3; 4, 2, 2; 3, 3, 3**: Scientific Accuracy **75%**, Reference Accuracy **67%**, Practical Value **75%**, and Overall **72%**. [41]

## Study integrity: rationale for all nine inputs

The retracted deadline paper receives no affirmative credit. The three Scientific Accuracy inputs are **2, 4, 3**: the matching policy is only partly supported; the bounded existence claim has independent randomized vaccination, smoking-commitment and exercise-incentive support without the deadline paper; and continuing support is substantially supported with behavior-specific exceptions. The newer nulls are not contradictions of those carefully stated propositions. [17, 14, 23, 36, 37, 42]

The Reference Accuracy inputs are **4, 2, 2**. Traceability is strong even when a located source is unreliable. Accurate description is mixed because faithful comparisons coexist with consequential endpoint and numerical errors, including the deadline endpoint. Appropriate inference is mixed: substantial faithful use of bounded evidence coexists with consequential overreach. The withdrawn paper and transfer limits contribute to that adverse case, but the criterion-level balance fits the Mixed anchor. Later retraction is not treated as evidence that Milkman knowingly used unreliable data in 2021. No unexplained duplicate deduction is applied.

The Practical Value inputs are **3, 3, 3**. Several ordinary-use recommendations have direct benefit evidence independent of the deadline example. Context and maintenance gaps fit the applicability grade; they do not establish that the book's important ordinary uses are largely unsuitable. The assessment also considers the book's documented safeguards and burdens. Favorable Norwegian checklist evidence is credited alongside the null scale-ups, not used to claim general clinical efficacy. [60]

These criterion judgments yield **75% Scientific Accuracy, 67% Reference Accuracy, 75% Practical Value, and 72% overall**. A future loss of central supporting evidence or a changed anchor would change the relevant input. The method does not impose a fixed per-paper penalty, a separate score, a hidden cap, or a post hoc replacement central claim.

## 10 Verification record and remaining limits

**Research stages and access.** The first ledger below records targeted source checks documented on 8 September 2026. "Primary text" means the stated material was retrieved, not that an entire article, its supplements or its dataset was independently audited. A repeat metadata or abstract check does not upgrade access to full text. Targeted repeat checks and their limits are identified separately below.

| Check and source | Finding at the checked scope | Remaining limitation |
|---|---|---|
| **H01**<br>Delgado et al. (2018): opioid prescription defaults<br>Primary methods and results text retrieved | The study reports no change in mean tablets prescribed across the two emergency departments; changes in the median and the share of ten-tablet prescriptions do not establish a halving of overall prescribing.<br>Footnote on opioid prescribing in chapter 5, "Laziness"; visually inspected. | No original data reanalysis or independent human verification. |
| **H02**<br>Gelman & Heyman (1999): carrot-eater labels<br>Primary abstract retrieved | The measured outcome was children's judgments of the stability of hypothetical people's characteristics, not whether the children ate more carrots.<br>Carrot-eater passage in chapter 1, "Getting Started"; visually inspected. | No original data reanalysis or independent human verification. |
| **H03**<br>Bollinger, Leslie & Sorensen (2011): calorie posting<br>Primary abstract retrieved | The abstract reports a 6% reduction in calories per transaction, principally food rather than drinks. This is purchasing, not direct food consumption or weight change.<br>See numerical audit A02. | Abstract-level check; the full dataset and every model were not reanalysed. |
| **H04**<br>Milkman et al. (2021): exercise megastudy<br>Publication identity and date rechecked | The later Nature publication is dated December 2021. This current identity check does not independently revalidate all result estimates preserved from the supplied review.<br>Chapter 8 dossier. | Retrieval supplied metadata and references; substantive result verification is documented separately in the research record. |
| **H05**<br>Beshears et al. (2021): fresh-start savings invitations<br>Publication identity rechecked | The linked work is the retirement-savings fresh-start paper, not evidence validating the book's entire matching framework.<br>Chapter 1 dossier. | Metadata-level current check; numerical result discussion is inherited from the supplied review. |

### Edition identification and source limits

Edition details are taken from the book's title and copyright pages. They do not authenticate the reviewed copy against a publisher-controlled master or establish correspondence with every commercially distributed edition.

### What has not been independently certified

The consolidation is not independent human fact-checking, expert peer review, legal clearance, an exhaustive quotation audit, a complete retraction/correction search, or participant-level reanalysis. Historical anecdotes, private interviews, unpublished data, blocked sources, and any unresolved numerical claims retain their original limits. Repeated AI review does not create independent corroboration.

HTML is the canonical report text; the PDF is generated from that same text without abridgment. The package's validation record documents the executed structural, scoring, citation-target, text-coverage, and rendering checks and their results. A functioning reference link is not proof that the linked study supports an assertion. Nor does a clean PDF layout validate the science.

### Substantive access limits

### Validation boundary

The review checks selected source-to-claim links, key numerical interpretations, page locations, document integrity, and PDF/HTML content consistency. It does not independently reproduce the studies' analyses from raw data, authenticate datasets, verify every anecdote, establish a comprehensive retraction history, or substitute for an independent subject-matter reviewer. The report is AI-produced and source-checked. It is not independently human peer reviewed. No author response was solicited, and no endorsement by the author, publisher, or Red Pen Reviews is implied.

### Targeted repeat source checks

**V2-H01 · Milkman et al. (2021), Megastudies improve the impact of applied behavioural science.** Publisher abstract and publication record rechecked on 8 September 2026. The reported intervention effects concern gym attendance, with limited significant post-intervention persistence. This supports bounded component benefit and the importance of maintenance, not universal lasting change.

**Limit:** No new full-text or participant-level reanalysis; conclusions and source-access distinctions already in the review remain bounded. This check does not constitute a new comprehensive literature search.

Production checks for this edition cover score arithmetic, internal links, preservation of the substantive analysis, agreement of HTML and PDF text, and representative browser and PDF rendering. The accompanying production record reports the actual results; these checks do not establish scientific or legal clearance.

### 13 September 2026: targeted housekeeper evidence check

The full published Stanforth et al. (2011) paper was read, including its methods, both results tables, discussion, and limitations. The original Crum and Langer (2007) methods, results table, and discussion were also rechecked. The 2011 paper's identity, year, title, authors, and DOI were checked against its PubMed record. This access supports the population, assignment, follow-up, outcome, and limitation statements. No participant-level reanalysis, independent replication performed by this publication, or independent human subject-matter sign-off is claimed. [41]

This was a targeted source check, not a comprehensive literature search. The 2011 paper predates the 2021 book. It was examined on 13 September 2026. The reader explanation was checked against the companion appendix, and the nine scoring inputs were assessed under the common criteria.

### 13 September 2026: source-by-source study-integrity audit

The preparation record documents a 13 September 2026 audit of 41 reference-table entries and 20 claim-register entries, plus the named baseball, Mandatory Fun, explicit voting-pressure and piano-staircase examples. That record lists checks of publication records, primary notices, replications, reanalyses and syntheses. This review does not independently authenticate every documented retrieval or search. The deadline retraction was confirmed in the publisher's full notice even though the dataset notice screen had no matching original-reference hit. Notice absence is therefore not clearance or evidence that a result replicated.

New access varies by source: full notices and selected primary methods/results for deadline, voter, hotel and dynamic-norm work; a full medication-reminder author manuscript; primary abstract/discussion for affirmation; and structured abstracts for Norwegian checklist and the 2026 norm synthesis. The corrected milkshake methods/results and minor reference correction were checked without independently rerunning the original analyses. The new reference rows state the actual access. Original datasets and statistical models were not independently reproduced. Unidentified anecdotes, exact original towel percentages, CAPTCHA figures and promotional statistics remain uncertified.

The evidence cutoff is 13 September 2026. Research available before the book is distinguished from later developments. Neither source access nor repeated AI analysis is represented as independent human scientific review. Section 09 records each scoring decision. Dated source-access records identify when checks were made; they do not mean that every original check was repeated.

## 11 References and source access

These are the sources cited in the evidence dossiers. Their access notes identify the documented source work and distinguish abstracts and metadata records from full-text verification. Targeted repeat checks and unsuccessful access attempts are described in Section

10. Identifying a source, including a review article or institutional guidance, is not itself evidence that it supports every claim associated with it.

Book citations use chapters and named sections rather than edition-dependent page numbers. Reference numbers are local to this report. Codes such as T/full text, A/abstract, M/metadata, or U/unresolved are defined in the local source ledger; a source can have accessible metadata while a particular passage remains unverified. External links require internet access; the review text and internal navigation work offline.

## References and access notes

References identify the material actually used. "Primary text" means the stated methods or results were examined, not that raw data were independently reanalyzed. Earlier access notes record the 8 September review; additional checks and follow-up sources are dated 13 September 2026. Abstract-level access supports only findings reported there. Online and issue dates, corrected versions and retracted sources are identified when material.

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 01 | 01<br>Milkman, K. (2021). **How to Change: The Science of Getting from Where You Are to Where You Want to Be**. *Portfolio/Penguin. Ebook ISBN 9780593083765. Chapters, notes and footnotes inspected; citations identify chapters and named sections.*<br><br>Access: Full book, notes, and chapter footnotes. |
| 02 | 02<br>Red Pen Reviews (2025). **Review method, version 2.2**. *Official methodology*.<br><br>Primary source / version examined<br><br>Access: Full methodology. |
| 03 | 03<br>Bollinger, B.; Leslie, P.; Sorensen, A. (2011). **Calorie Posting in Chain Restaurants**. *American Economic Journal: Economic Policy, 3(1), 91–128*.<br><br>DOI: 10.1257/pol.3.1.91<br><br>Access: Publisher abstract. |

| Row | Source identity, bibliography, links, and access record |
| --- | --- |
| 04 | 04<br>Clarke, N.; Pechey, E.; Shemilt, I.; et al. (2025). **Calorie (energy) labelling for changing selection and consumption of food or alcohol**. *Cochrane Database of Systematic Reviews, CD014845*.<br><br>DOI: 10.1002/14651858.CD014845.pub2<br><br>Primary source / version examined<br><br>Access: Full structured review abstract and Cochrane evidence summary. |
| 05 | 05<br>Beshears, J.; Dai, H.; Milkman, K. L.; Benartzi, S. (2021). **Using Fresh Starts to Nudge Increased Retirement Savings**. *Organizational Behavior and Human Decision Processes*.<br><br>DOI: 10.1016/j.obhdp.2021.06.005<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |
| 06 | 06<br>Gelman, S. A.; Heyman, G. D. (1999). **Carrot-Eaters and Creature-Believers: The Effects of Lexicalization on Children's Inferences About Social Categories**. *Psychological Science, 10(6), 489–493*.<br><br>DOI: 10.1111/1467-9280.00194<br><br>Access: Publisher abstract; outcome and sample verified. |
| 07 | 07<br>Gerber, A. S.; Huber, G. A.; Biggers, D. R.; Hendry, D. J. (2016). **A field experiment shows that subtle linguistic cues might not affect voter behavior**. *Proceedings of the National Academy of Sciences*.<br><br>DOI: 10.1073/pnas.1513727113<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 08 | 08<br>Milkman, K. L.; Minson, J. A.; Volpp, K. G. M. (2014 [online 2013]). **Holding the Hunger Games Hostage at the Gym: An Evaluation of Temptation Bundling**. *Management Science, 60(2), 283–299*.<br><br>DOI: 10.1287/mnsc.2013.1784<br><br>Primary source / version examined<br><br>Access: Full primary PDF; Table 2 visually checked. |
| 09 | 09<br>Kirgios, E. L.; Mandel, G. H.; Park, Y.; et al. (2020). **Teaching temptation bundling to boost exercise: A field experiment**. *Organizational Behavior and Human Decision Processes*.<br><br>DOI: 10.1016/j.obhdp.2020.09.003<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |
| 10 | 10<br>Gallus, J. (2017 [online 2016]). **Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia**. *Management Science, 63(12), 3999–4015*.<br><br>DOI: 10.1287/mnsc.2016.2540<br><br>Access: Publisher abstract. |
| 11 | 11<br>Fanaroff, A. C.; Patel, M. S.; Chokshi, N.; et al. (2024). **Effect of Gamification, Financial Incentives, or Both to Increase Physical Activity Among Patients at High Risk of Cardiovascular Events: The BE ACTIVE Randomized Controlled Trial**. *Circulation*.<br><br>DOI: 10.1161/CIRCULATIONAHA.124.069531<br><br>Primary source / version examined<br><br>Access: Primary article results/methods and trial registry NCT03911141. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 12 | 12<br>Ashraf, N.; Karlan, D.; Yin, W. (2006). **Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines**. *Quarterly Journal of Economics, 121(2), 635–672*.<br><br>DOI: 10.1162/qjec.2006.121.2.635<br><br>Access: Published article excerpts and author working-paper methods/results; version-sensitive details identified in text. |
| 13 | 13<br>Ariely, D.; Wertenbroch, K. (2002). **Procrastination, Deadlines, and Performance: Self-Control by Precommitment**. *Psychological Science, 13(3), 219–224*.<br><br>DOI: 10.1111/1467-9280.00441<br><br>Historical author manuscript · Publisher retraction notice<br><br>Earlier access: Author manuscript; reported methods and Figure 2 inspected. Additional 13 September 2026 access: full publisher retraction notice and close-replication methods/transparency/discussion. **Status: retracted 2 September 2026; retained only as a historical source, with no affirmative evidentiary credit.** |
| 14 | 14<br>Giné, X.; Karlan, D.; Zinman, J. (2010). **Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation**. *American Economic Journal: Applied Economics, 2(4), 213–235*.<br><br>DOI: 10.1257/app.2.4.213<br><br>Access: Publisher abstract. |
| 15 | 15<br>Halpern, S. D.; French, B.; Small, D. S.; et al. (2015). **Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation**. *New England Journal of Medicine*.<br><br>DOI: 10.1056/NEJMoa1414293<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 16 | 16<br>Meeker, D.; Knight, T. K.; Friedberg, M. W.; et al. (2014). **Nudging Guideline-Concordant Antibiotic Prescribing: A Randomized Clinical Trial**. *JAMA Internal Medicine, 174(3), 425–431*.<br><br>DOI: 10.1001/jamainternmed.2013.14191<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |
| 17 | 17<br>Milkman, K. L.; Beshears, J.; Choi, J. J.; Laibson, D.; Madrian, B. C. (2011). **Using implementation intentions prompts to enhance influenza vaccination rates**. *Proceedings of the National Academy of Sciences, 108(26), 10415–10420*.<br><br>DOI: 10.1073/pnas.1103170108<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |
| 18 | 18<br>Sheeran, P.; Listrom, O.; Gollwitzer, P. M. (2025 [online 2024]). **The when and how of planning: Meta-analysis of the scope and components of implementation intentions in 642 tests**. *European Review of Social Psychology, 36(1), 162–194*.<br><br>DOI: 10.1080/10463283.2024.2334563<br><br>Primary source / version examined<br><br>Access: Full primary text excerpts, including publication-bias analysis and search limitations. |
| 19 | 19<br>Haynes, A. B.; Weiser, T. G.; Berry, W. R.; et al. (2009). **A surgical safety checklist to reduce morbidity and mortality in a global population**. *New England Journal of Medicine, 360, 491–499*.<br><br>DOI: 10.1056/NEJMsa0810119<br><br>Primary source / version examined<br><br>Access: Primary article methods/results and PubMed abstract. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 20 | 20<br>Urbach, D. R.; Govindarajan, A.; Saskin, R.; Wilton, A. S.; Baxter, N. N. (2014). **Introduction of Surgical Safety Checklists in Ontario, Canada**. *New England Journal of Medicine*.<br><br>DOI: 10.1056/NEJMsa1308261<br><br>Primary source / version examined<br><br>Access: Primary article methods/results and investigator-institution abstract. |
| 21 | 21<br>Patel, M. S.; Day, S. C.; Halpern, S. D.; et al. (2016). **Generic Medication Prescription Rates After Health System–Wide Redesign of Default Options Within the Electronic Health Record**. *JAMA Internal Medicine, 176(6), 847–848*.<br><br>DOI: 10.1001/jamainternmed.2016.1691<br><br>Access: Primary text / relevant results and methods. |
| 22 | 22<br>Delgado, M. K.; Shofer, F. S.; Patel, M. S.; et al. (2018). **Association between Electronic Medical Record Implementation of Default Opioid Prescription Quantities and Prescribing Behavior in Two Emergency Departments**. *Journal of General Internal Medicine, 33, 409–411*.<br><br>DOI: 10.1007/s11606-017-4286-5<br><br>Primary source / version examined<br><br>Access: Full primary PDF; results and figures visually checked. |
| 23 | 23<br>Beshears, J.; Lee, H. N.; Milkman, K. L.; Mislavsky, R.; Wisdom, J. (2021 [online 2020]). **Creating Exercise Habits Using Incentives: The Trade-off Between Flexibility and Routinization**. *Management Science, 67(7), 4139–4171*.<br><br>DOI: 10.1287/mnsc.2020.3706<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |

| Row | Source identity, bibliography, links, and access record |
| --- | --- |
| 24 | 24<br>Eskreis-Winkler, L.; Milkman, K. L.; Gromet, D. M.; Duckworth, A. L. (2019). **A large-scale field experiment shows giving advice improves academic outcomes for the advisor**. *Proceedings of the National Academy of Sciences*.<br><br>DOI: 10.1073/pnas.1908779116<br><br>Access: Publisher abstract; preregistration, sample, outcomes and horizon verified. |
| 25 | 25<br>Crum, A. J.; Langer, E. J. (2007). **Mind-Set Matters: Exercise and the Placebo Effect**. *Psychological Science, 18(2), 165–171*.<br><br>DOI: 10.1111/j.1467-9280.2007.01867.x<br><br>Primary source / version examined<br><br>Access: Primary article methods/results. |
| 26 | 26<br>Crum, A. J.; Corbin, W. R.; Brownell, K. D.; Salovey, P. (2011). **Mind over milkshakes: Mindsets, not just nutrients, determine ghrelin response**. *Health Psychology*.<br><br>DOI: 10.1037/a0023467<br><br>Access: Primary structured abstract.<br><br>Additional 13 September 2026 check: corrected author-hosted methods/results and APA citation-correction notice. Acute biomarker evidence only; the correction does not change results. Direct independent label/ghrelin replication was not established. No source-specific retraction was found in the checked records. Corrected author-hosted paper. |
| 27 | 27<br>Yeager, D. S.; Hanselman, P.; Walton, G. M.; et al. (2019). **A national experiment reveals where a growth mindset improves achievement**. *Nature, 573, 364–369*.<br><br>DOI: 10.1038/s41586-019-1466-y<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 28 | 28<br>Macnamara, B. N.; Burgoyne, A. P. (2023 [online 2022]). **Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices**. *Psychological Bulletin, 149(3–4), 133–173*.<br><br>DOI: 10.1037/bul0000352<br><br>Access: Primary abstract and author preprint results, including quality-restricted analysis. |
| 29 | 29<br>Burnette, J. L.; et al. (2023). **A systematic review and meta-analysis of growth mindset interventions: For whom, how, and why might such interventions work?**. *Psychological Bulletin, 149(3–4), 174–205*.<br><br>DOI: 10.1037/bul0000368<br><br>Access: Primary structured abstract. |
| 30 | 30<br>Carrell, S. E.; Sacerdote, B. I.; West, J. E. (2013). **From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation**. *Econometrica, 81(3)*.<br><br>DOI: 10.3982/ECTA10168<br><br>Primary source / version examined<br><br>Access: Publisher abstract. |
| 31 | 31<br>Beshears, J.; Choi, J. J.; Laibson, D.; Madrian, B. C.; Milkman, K. L. (2015). **The Effect of Providing Peer Information on Retirement Savings Decisions**. *Journal of Finance, 70(3), 1161–1201*.<br><br>DOI: 10.1111/jofi.12258<br><br>Access: Primary text / relevant results and methods. |
| 32 | 32<br>Mehr, K. S.; Geiser, A. E.; Milkman, K. L.; Duckworth, A. L. (2020). **Copy-Paste Prompts: A New Nudge to Promote Goal Achievement**. *Journal of the Association for Consumer Research*.<br><br>DOI: 10.1086/708880<br><br>Access: Publisher abstract. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 33 | 33<br>Goldstein, N. J.; Cialdini, R. B.; Griskevicius, V. (2008). **A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels**. *Journal of Consumer Research*.<br><br>DOI: 10.1086/586910<br><br>Access: Publisher abstract; exact book percentage contrasts not independently reconstructed. |
| 34 | 34<br>Sparkman, G.; Walton, G. M. (2017). **Dynamic Norms Promote Sustainable Behavior, Even if It Is Counternormative**. *Psychological Science*.<br><br>DOI: 10.1177/0956797617719950<br><br>Access: Publisher abstract. |
| 35 | 35<br>Sharif, M. A.; Shu, S. B. (2017 [online 2016]). **The Benefits of Emergency Reserves: Greater Preference and Persistence for Goals That Have Slack with a Cost**. *Journal of Marketing Research*.<br><br>DOI: 10.1509/jmr.15.0231<br><br>Access: Publisher abstract; exact book CAPTCHA percentages not independently reconstructed. |
| 36 | 36<br>Milkman, K. L.; Gromet, D.; Ho, H.; et al. (2021). **Megastudies improve the impact of applied behavioural science**. *Nature*.<br><br>DOI: 10.1038/s41586-021-04128-4<br><br>Primary source / version examined<br><br>Access: Primary text / relevant results and methods. |
| 37 | 37<br>Allcott, H.; Rogers, T. (2014). **The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation**. *American Economic Review, 104(10), 3003–3037*.<br><br>DOI: 10.1257/aer.104.10.3003<br><br>Access: Primary text / relevant results and methods. |

| Row | Source identity, bibliography, links, and access record |
| --- | --- |
| 38 | 38<br>DellaVigna, S.; Linos, E. (2022). **RCTs to Scale: Comprehensive Evidence From Two Nudge Units**. *Econometrica, 90*.<br><br>DOI: 10.3982/ECTA18709<br><br>Primary source / version examined<br><br>Access: Published abstract and author/NBER paper methods; published estimates used. |
| 39 | 39<br>Galla, B. M.; Duckworth, A. L. (2015). **More than resisting temptation: Beneficial habits mediate the relationship between self-control and positive life outcomes**. *Journal of Personality and Social Psychology*.<br><br>DOI: 10.1037/pspp0000026<br><br>Access: Primary article abstract and methods/results excerpts. |
| 40 | 40<br>Cohen, G. L.; Sherman, D. K. (2014). **The Psychology of Change: Self-Affirmation and Social Psychological Intervention**. *Annual Review of Psychology, 65, 333–371*.<br><br>DOI: 10.1146/annurev-psych-010213-115137<br><br>Access: Primary scholarly review abstract and relevant sections. |
| 41 | 41<br>Stanforth, D.; Steinhardt, M.; Mackert, M.; Stanforth, P. R.; Gloria, C. T. (2011). **An investigation of exercise and the placebo effect**. *American Journal of Health Behavior, 35(3), 257–268*.<br><br>DOI: 10.5993/ajhb.35.3.1 · PubMed record · Full published paper<br><br>Access: Full published paper, including methods, both results tables, discussion, and limitations, read on 13 September 2026. Occupational partial replication; not an exact hotel-housekeeper repeat. Original participant-level data were not reanalysed. |
| 42 | 42<br>**Retraction: Procrastination, Deadlines, and Performance: Self-Control by Precommitment**. Publication/version: 2026-09-02.<br><br>DOI: 10.1177/09567976261488042 · Primary publication or notice<br><br>13 September 2026 access: Full publisher retraction notice. Formal retraction notice; original findings receive no supporting credit. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 43 | 43<br>Hyndman & Bisin. **Replication of Procrastination, Deadlines, and Performance**. Publication/version: 2026-07-15.<br><br>DOI: 10.1177/09567976261460772 · Primary publication or notice<br><br>13 September 2026 access: Published abstract, full methods, transparency statement and discussion. Follow-up source, not a new rerun performed by this publication. |
| 44 | 44<br>Gerber, Huber & Fang. **Do Subtle Linguistic Interventions Priming a Social Identity as a Voter Have Outsized Effects on Voter Turnout?**. Publication/version: 2017 online; 2018 issue.<br><br>DOI: 10.1111/pops.12446 · Primary publication or notice<br><br>13 September 2026 access: Primary abstract, later primary report summarizing design/results, and full 2023 correction. Follow-up source, not a new rerun performed by this publication. |
| 45 | 45<br>**Correction to Gerber, Huber & Fang: Do subtle linguistic interventions priming a social identity as a voter have outsized effects on voter turnout?**. Publication/version: 2023-09-11 online; 2024 issue.<br><br>DOI: 10.1111/pops.12931 · Primary publication or notice<br><br>13 September 2026 access: Full publisher correction and corrected Table 3. Follow-up source, not a new rerun performed by this publication. |
| 46 | 46<br>Bryan, Yeager & O'Brien. **Replicator degrees of freedom allow publication of misleading failures to replicate**. Publication/version: 2019-11-25.<br><br>DOI: 10.1073/pnas.1910951116 · Primary publication or notice<br><br>13 September 2026 access: Primary reanalysis text and author institutional record. Follow-up source, not a new rerun performed by this publication. |
| 47 | 47<br>**Correction for Bryan et al.: Replicator degrees of freedom**. Publication/version: 2021-08-09.<br><br>DOI: 10.1073/pnas.2111799118 · Primary publication or notice<br><br>13 September 2026 access: Full one-page primary notice in indexed institutional PDF. Follow-up source, not a new rerun performed by this publication. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 48 | 48<br>Gerber, Huber & Fang. **Voting behavior is unaffected by subtle linguistic cues: evidence from a psychologically authentic replication**. Publication/version: 2020 online; 2023 issue.<br><br>DOI: 10.1017/bpp.2020.57 · Primary publication or notice<br><br>13 September 2026 access: Primary published PDF methods, results and discussion. Follow-up source, not a new rerun performed by this publication. |
| 49 | 49<br>Witkowska et al.. **The Grammar of Persuasion: A Meta-Analytic Review Disconfirming the Role of Nouns as Linguistic Cues of Subsequent Behavior**. Publication/version: 2024-03-14.<br><br>DOI: 10.1177/0261927x241234845 · Primary publication or notice<br><br>13 September 2026 access: Primary publisher abstract and included-reference list; full analysis not accessed. Follow-up source, not a new rerun performed by this publication. |
| 50 | 50<br>Sallis et al.. **Prescriber Commitment Posters to Increase Prudent Antibiotic Prescribing in English General Practice: A Cluster Randomized Controlled Trial**. Publication/version: 2020.<br><br>DOI: 10.3390/antibiotics9080490 · Primary publication or notice<br><br>13 September 2026 access: Primary abstract, results and full discussion, including replication limits. Follow-up source, not a new rerun performed by this publication. |
| 51 | 51<br>Bohner & Schlüter. **A Room with a Viewpoint Revisited: Descriptive Norms and Hotel Guests' Towel Reuse Behavior**. Publication/version: 2014-08-01; corrected article republished 2014-08-06.<br><br>DOI: 10.1371/journal.pone.0104086 · Primary publication or notice<br><br>13 September 2026 access: Primary corrected article methods, both results sections and discussion. Follow-up source, not a new rerun performed by this publication. |
| 52 | 52<br>PLOS ONE Staff. **PLOS ONE Staff: Correction: A Room with a Viewpoint Revisited**. Publication/version: 2014-08-20.<br><br>DOI: 10.1371/journal.pone.0106606 · Primary publication or notice<br><br>13 September 2026 access: Full publisher notice. Follow-up source, not a new rerun performed by this publication. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 53 | 53<br>Aldoh, Sparks & Harris. **Dynamic Norms and Food Choice: Reflections on a Failure of Minority Norm Information to Influence Motivation to Reduce Meat Consumption**. Publication/version: 2021-07-26.<br><br>DOI: 10.3390/su13158315 · Primary publication or notice<br><br>13 September 2026 access: Primary indexed abstract and methods excerpt; direct page retrieval rate-limited. Follow-up source, not a new rerun performed by this publication. |
| 54 | 54<br>Aldoh, Sparks & Harris. **Shifting norms, static behaviour: effects of dynamic norms on meat consumption**. Publication/version: 2024-06-26.<br><br>DOI: 10.1098/rsos.240407 · Primary publication or notice<br><br>13 September 2026 access: Primary abstract, preregistered analysis plan, main contrasts, interaction/simple-effects results and data-access statement; models not independently rerun. Follow-up source, not a new rerun performed by this publication. |
| 55 | 55<br>Fernanda M. Reintgen Kamphuisen, Thijs Bouman, & Ellen van der Werff. **Unpacking dynamic norms: Identifying key elements of dynamic norms and a meta-analysis of their effects on pro-environmental behaviour**. Publication/version: 2026-05-06 online corrected proof.<br><br>DOI: 10.1016/j.jenvp.2026.103060 · Primary publication or notice<br><br>13 September 2026 access: Primary indexed abstract and highlights; full analysis not accessed. Follow-up source, not a new rerun performed by this publication. |
| 56 | 56<br>**Correction to Crum et al.: Mind over milkshakes**. Publication/version: 2011.<br><br>DOI: 10.1037/a0024760 · Primary publication or notice<br><br>13 September 2026 access: Primary APA correction and boxed correction in author-hosted article. Follow-up source, not a new rerun performed by this publication. |
| 57 | 57<br>Hoffmann et al.. **Effects of Placebo Interventions on Subjective and Objective Markers of Appetite—A Randomized Controlled Trial**. Publication/version: 2018.<br><br>DOI: 10.3389/fpsyt.2018.00706 · Primary publication or notice<br><br>13 September 2026 access: Primary abstract, methods/results excerpts and author manuscript. Follow-up source, not a new rerun performed by this publication. |

| Row | Source identity, bibliography, links, and access record |
|---|---|
| 58 | 58<br>Dai et al.. **Effectiveness of Medication Adherence Reminders Tied to Fresh Start Dates: A Randomized Clinical Trial**. Publication/version: 2017-02-08 online.<br><br>DOI: 10.1001/jamacardio.2016.5794 · NIH author manuscript · Final published article<br><br>13 September 2026 access: Full NIH author manuscript; manuscript title differs from publisher title but DOI and authors match. Follow-up source, not a new rerun performed by this publication. A further check of the final published article read its methods and results and inspected its participant-flow figure. The final report distinguishes 15,011 randomized participants, 13,323 reminder recipients and 10,318 analyzed participants. |
| 59 | 59<br>Hanselman et al.. **New Evidence on Self-Affirmation Effects and Theorized Sources of Heterogeneity From Large-Scale Replications**. Publication/version: 2017.<br><br>DOI: 10.1037/edu0000141 · Primary publication or notice<br><br>13 September 2026 access: Primary abstract and discussion in full NIH author manuscript. Follow-up source, not a new rerun performed by this publication. |
| 60 | 60<br>Haugen et al.. **Effect of the World Health Organization checklist on patient outcomes: a stepped wedge cluster randomized controlled trial**. Publication/version: 2014 online; 2015 issue.<br><br>DOI: 10.1097/sla.0000000000000716 · Primary publication or notice<br><br>13 September 2026 access: Full primary structured abstract and comment links; full original article not accessed. Follow-up source, not a new rerun performed by this publication. |

## 12 Editorial Review Notice, Disclosures, and Corrections

### Purpose and scope.

This report is published for criticism, commentary, education, and discussion of matters of public interest. It evaluates the particular work, edition, claims, and evidence identified in the report. Its conclusions should not be extended to editions, claims, publications, or professional activities that it does not examine. It is not a comprehensive audit of the author’s work or a certification of scientific accuracy.

### Factual reporting and editorial judgment.

This review contains both factual reporting and editorial analysis. Quotations, descriptions of source material, and reported research findings are presented as factual information. Assessments of evidentiary strength, methodological quality, interpretation, balance, and

practical value are the reviewer's evaluative conclusions, based on the sources, reasoning, and criteria identified in the report. Readers are encouraged to examine those materials and assess the conclusions for themselves. Evaluative judgments are not intended to imply possession of undisclosed facts concerning anyone's conduct or motives.

### Meaning of critical assessments.

Descriptions such as "unsupported," "overstated," "misleading," or "inconsistent with the evidence" concern the specific claim, presentation, or evidentiary relationship discussed, for the reasons explained in the accompanying analysis. They do not, by themselves, assert that an author or other person knowingly made a false statement, intended to deceive, fabricated evidence, or engaged in professional misconduct. An assessment that the available evidence does not adequately support a claim is distinct from a finding that the claim is necessarily false.

### Interpretation of scores.

Scores, grades, rankings, and recommendation categories summarize the reviewer's application of the stated rubric. They are not measurements of the author's honesty, character, intentions, or overall professional competence. They are not estimates of the percentage of a book that is true or false, probabilities that an author is correct, or representations of scientific consensus. The accompanying analysis explains the basis and limitations of each assessment.

### Evidence, timing, and verification limits.

Conclusions reflect the materials examined and the state of the evidence assessed through the literature-search cutoff stated in this report. Relevant limitations—including unavailable sources, incomplete access to underlying data, and unresolved verification questions—are identified in the methodology or accompanying analysis. An inability to locate or independently verify a source does not, by itself, establish that the source does not exist or that a claim was fabricated. Evidence published after the reviewed work is distinguished from evidence available when it was written; later developments do not, by themselves, establish what an author knew or should have known at the time. Independent replication of studies, reanalysis of original datasets, and expert peer review are not claimed unless expressly documented.

### Preparation and review disclosure.

This paired reader review and technical appendix was prepared with ChatGPT at Jason Hreha's request. AI assisted research, analysis, drafting, scoring and production, including the source-specific study-integrity checks. The evidence cutoff is 13 September 2026. The source-access record and criterion-level scoring rationale describe the basis and limits of the assessment.

The publisher supplied an additional critique, *The Audit, Audited*, reporting approximately 110 checks of draft reviews. Its author, tool identity and independent human-review status were not established in the supplied material. Its reported checks are not relabeled as independent

human fact-checking or complete verification. Selected locator, version and cross-book template issues were checked during preparation. No comprehensive literature search or independent replication is claimed.

No independent human subject-matter sign-off or media-law clearance is documented. Jason Hreha is not represented as having personally verified every assertion. No response from a reviewed author or publisher was solicited during preparation. Source-target, arithmetic, content-preservation and rendering tests are production checks, not scientific peer review. The extent and limits of source access are recorded in Section 10 and the references.

### Relevant interests and relationships.

Jason Hreha leads The Behavioral Scientist, is the author of the commercially available behavior-change book *Real Change*, and offers behavioral-science consulting and related services. These are relevant commercial and intellectual interests, including potentially competing books, explanations or products. The publication's book page and services/contact page were checked on 8 September 2026.

The preparation record does not independently establish the presence or absence of additional relationships with the reviewed author or publisher, funding, compensation, review-copy arrangements, sponsorships or affiliate arrangements. No blanket absence-of-interests statement is made. The links included in this report are source and navigation links, not newly created affiliate links.

### Affiliation and attribution.

References to individuals, organizations, institutions, publications, and trademarks identify the subjects or sources of discussion and do not, merely by their inclusion, imply sponsorship, approval, or endorsement. This is not an official Red Pen Reviews review and is not affiliated with or endorsed by Red Pen Reviews. Any use or adaptation of its publicly described approach is identified in the methodology; the ratings in this report are not ratings issued by Red Pen Reviews. Copyright in quoted or reproduced third-party material, where applicable, remains with the respective rights holders. Its inclusion in this report does not grant permission for separate reuse.

### General information, not individualized advice.

This report does not provide individualized medical, psychological, legal, financial, or other professional advice, and reading it does not establish a professional-client relationship. Discussion of a recommendation's evidentiary support does not establish its suitability for any particular person or circumstance.

### Corrections and substantive responses.

The publisher welcomes factual corrections, relevant additional evidence, and substantive responses from authors, publishers, researchers, and readers. Please send submissions through **The Behavioral Scientist contact form**, identifying the passage at issue, the proposed correction or clarification, and supporting sources. The publisher will assess submissions and correct substantiated factual errors. Material corrections and updates made after first

publication will be dated and explained on the public page. Internal draft revisions are not public updates. A person's failure to respond to a request for comment is not treated as agreement with the review.

## Publication details

| | |
|---|---|
| **Reviewer / preparation** | ChatGPT-assisted critical analysis, commissioned by Jason Hreha. The extent of AI and human review is disclosed above. |
| **Publisher and responsible editorial contact** | The Behavioral Scientist / Haystack Group LLC, as identified in the website's terms. Responsible editorial contact: Jason Hreha. |
| **Book and edition reviewed** | How to Change — Katy Milkman. 2021 Portfolio / Penguin ebook; ISBN 9780593083765. |
| **Evidence assessed through** | 13 September 2026. This narrative audit is not a guarantee of exhaustive literature coverage. |
| **Evidence assessment and contact** | The assessment summary below explains the current evidence judgments. To submit a correction, use The Behavioral Scientist contact form and identify the book and passage. |

## Evidence assessment summary

**Study integrity.** The formally retracted Ariely/Wertenbroch deadline paper receives no affirmative evidentiary credit; its historical endpoint-description error is documented without endorsing its data. The assessment includes the close deadline replication, voter tests and reanalysis, corrected source versions, null English poster and German hotel tests, mixed dynamic-norm/affirmation/fresh-start evidence, and favorable Norwegian checklist evidence. It also records the minor milkshake reference correction and direct-replication gap. The nine-input scoring rationale explains the 72% overall rating. AI assisted the research; no original-data reproduction, independent human scientific sign-off, misconduct finding, or legal clearance is claimed.

**Housekeeper evidence.** Stanforth et al. (2011) did not reproduce differential weight, BMI, or body-fat benefits; waist circumference also showed no group-by-time benefit. Blood-pressure results favored the exercise-message group. The assessment accounts for the different population and control condition, two-location assignment, attrition, limited power, and activity-measurement limits. The paper predates the book. The claim-register verdict is **Overstated**, not conclusively false or fraudulent. The criterion-level scoring rationale explains the nine inputs and **72%** overall rating. AI materially assisted this targeted source check; no comprehensive search, independent human fact-checking, or legal clearance is claimed.