BOOK EVIDENCE REVIEW

How to Change review

A book by Katy Milkman

Useful tools, uneven science

The Behavioral Scientist

4,729 words / 16 min read

Evidence assessed through 13 September 2026 · Book edition and source-copy details below

OVERALL RATING

72%

An editorial assessment, not a percentage of truth.

Scientific Accuracy	**75%**
Reference Accuracy	**67%**
Practical Value	**75%**

Read it this way: A useful behavior-change toolkit, with important corrections to its scientific storytelling.

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

Confidence in the overall appraisal: moderate. Specific evidence limits accompany the findings.

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. Read the methodology, book-specific disclosures, and full editorial notice with the review. Corrections and substantive responses are welcome.

Katy Milkman writes that labels such as "carrot eaters" influence how we act, not just how we describe ourselves.

There's a problem. The study behind the example didn't measure children eating carrots. It measured what children inferred about another person after hearing a label. That's an interesting finding about language. It answers a different question from the one the book puts in front of the reader. [1]

This is the tension running through *How to Change*. Some of its advice rests on real experiments with useful, observable results. Some of its scientific storytelling turns a narrower finding into a more exciting claim. You need to know which part you're reading.

My verdict: this is a useful book, especially for readers who want to move beyond the idea that failed behavior change always means insufficient willpower. Keep the practical tools. Be more selective about the explanations attached to them.

The best way to read this book is as a collection of options to test, rather than a diagnostic manual that already knows why you're stuck. That reading preserves much of its value and removes much of the overpromising.

What Milkman gets right about change

Milkman's starting point is sensible: different obstacles call for different responses. Someone who forgets an appointment has a different problem from someone who cannot afford it. Someone who dislikes an activity has a different problem from someone who enjoys it but cannot find the time. Treating these as a single shortage of motivation makes intervention design worse. [2]

The book organizes its recommendations around starting, impulsivity, procrastination, forgetting, inertia, confidence, and social influence. It then does something more valuable than adding an eighth clever trick: it asks how change survives after the intervention ends. [2]

Milkman is also more cautious than a quick dismissal of the book would suggest. She acknowledges fading effects, low uptake, unhelpful social pressure, and the possibility that rigid routines can backfire. Those qualifications are part of her argument, and she deserves credit for including them.

The strongest version of her argument is therefore practical: stop demanding the same thing from yourself in the same circumstances, and start changing something about the circumstances. Make the next action easier to remember, easier to perform, or more rewarding to repeat.

That is a productive place to begin. The remaining questions are about evidence, magnitude, and fit. Which change? For whom? Compared with what? And for how long?

Three claims worth separating

The book contains three different propositions that are easy to blur together.

First, matching the strategy to the obstacle improves results. As a design principle, this is promising. As a validated system for classifying readers and selecting their best intervention, the book's seven-obstacle framework remains unproven.

Second, behavioral tools can improve concrete actions. Some randomized field experiments really do improve vaccination, attendance, or other measured behaviors. This is the narrower claim that receives a 100% grade in the technical appendix: at least some of these tools can improve specific actions under the conditions studied. The grade reflects strong support for that claim. It does not mean every tool works or that behavioral science delivers universal transformation.

Third, maintaining a change often requires continued support. This is one of the book's strongest lessons, with the sensible qualification that different behaviors have different maintenance needs. [3]

The evidence that particular tools can help, and that useful changes may need continuing support, does not establish that the book can reliably tell each reader which tool to choose. You can accept the second and third propositions while treating the first as a promising idea rather than a finished personalization system. That distinction is central to this review's 75% Scientific Accuracy rating.

There is a useful way to test the first proposition properly. Measure the proposed obstacles before assigning an intervention, specify how those measurements determine the choice, and compare the matched approach with a credible alternative. That might be participant choice, a strong general program, or the best single intervention available. [3]

The key question is not just whether the framework identifies people who have more difficulty, but whether it predicts **which intervention works better for them**. Someone can report being forgetful and still benefit more from improved access than from another reminder. A useful matching system has to tell those possibilities apart; identifying the difficulty alone is not enough.

That comparison asks more of the framework than a collection of successful examples arranged under seven chapter headings. It is what would turn a sensible organizing idea into evidence-backed personalization. Until the comparison is available, use the framework to suggest possibilities, try them, and revise your choice in light of what happens.

The carrot story changes the outcome

The carrot example is a clean illustration of how scientific communication goes wrong without requiring an elaborate statistical dispute.

The researchers told children about hypothetical people. Some descriptions used a noun label, such as describing someone as a carrot-eater. Others described the behavior without giving the person a category label. The children then judged how stable the characteristic would be across time and situations. [1]

The study tells us something about how children interpret a description. It doesn't show that labeling a child changes the child's eating behavior. To establish that, you would need to observe the relevant behavior after the labeling intervention and compare it with an appropriate alternative.

Why care? To see the practical difference, imagine a parent trying to encourage a child to eat more vegetables. The parent needs to know whether a label changes what the child eats. The source is real, but it offers evidence about other children's judgments, not that outcome. That gap is the problem with using it to support Milkman's broader claim about labels and behavior.

The nearby voter-identity example also needs context. The original finding favored asking someone to be a "voter" rather than to "vote." Several later field tests did not reproduce a reliable large advantage. Bryan and colleagues argued that differences in timing, participants, and analysis explained some failures. Their 2019 reanalysis of existing data tested many author-defined model choices and reported a benefit, with stronger estimates among people contacted one or two days before Election Day. A closer preregistered test during the 2016 presidential election still did not find a noun-wording advantage. A 2024 review found a small average effect, but its diagnostic checks did not establish that nouns reliably change behavior. That is a mixed record, not a dependable turnout technique. It also does not retest the carrot study's separate finding about children's judgments. See the replication evidence and the researchers' disagreement.

The better takeaway is modest: language can shape interpretation, and identity language may sometimes influence behavior. But a parent, teacher, or designer should test the actual outcome they care about. A label's effect on how people think about an activity is only one step along the way.

The opioid example needs a substantial correction

An endnote says a prescribing default cut prescriptions of habit-forming opioids in half. The cited study tells a less dramatic story.

After a ten-tablet default was introduced in two emergency departments, prescriptions clustered more heavily around ten tablets. At each hospital, the median number prescribed—the middle value when prescription quantities are ordered—fell slightly. But the mean, or arithmetic average, did not change by a statistically significant amount across the two departments. The earlier system required clinicians to enter a quantity; it wasn't simply a thirty-day default replaced with ten pills. [5]

That matters because several quantities are being confused: prescriptions, tablets, days of treatment, and the distribution of prescription sizes. They describe different outcomes. The source supports an effect on prescribing patterns. It doesn't support the headline that overall opioid prescribing was halved by the intervention.

This is a more consequential error than a stray percentage in a sports anecdote. A reader could come away thinking a low-cost interface change produced a huge reduction in a medically important outcome, when the evidence supports a narrower change.

Defaults remain worth studying and, in appropriate settings, using. But their effects must be described accurately. If prescriptions become more concentrated around a preferred quantity while other prescription sizes also change, the pattern can shift without much movement in the average. A greater share of prescriptions at the preferred quantity is not the same result as a large reduction in prescribing overall.

The appropriate correction is straightforward: describe the change in prescription quantities that the study actually observed, retain the comparison and setting, and remove the halving claim. The intervention doesn't need an inflated result to be interesting. In applied work, knowing the actual size and shape of an effect is more useful than having a dramatic story to repeat.

A good direction can still come with the wrong number

The book describes a baseball player's batting average rising from .246 to .294 as a 29% improvement. Using those figures, the increase is about 19.5%: the difference, .048, divided by the starting value, .246. That is an arithmetic correction, not a dispute about the psychology of fresh starts. [6]

The example also has a second limitation. One player's improvement can illustrate a story, but it cannot isolate a psychological reset as its cause. A change can be real while the explanation remains uncertain.

The deadline example now raises a more serious concern than a misplaced percentage. The book describes self-imposed deadlines in terms of fewer errors, but the experiment asked participants to find planted errors in supplied texts. Detecting more errors and making fewer errors are different measures. Even "50% more detected" and "50% fewer left" use different denominators. That source-description problem remains, but it is no longer the main reason not to rely on the study. [7]

On 2 September 2026, *Psychological Science* retracted Ariely and Wertenbroch's 2002 paper. The journal could no longer attest to the reliability of its data and findings, and both authors agreed to the retraction. A close replication published earlier in 2026 also failed to reproduce the original deadline-performance results. Its researchers could not obtain the original instructions and had to change some procedures, so it was not an identical repeat. But the retraction means we should stop using the original paper as affirmative evidence that these deadlines improve performance—not replace its headline with a more carefully worded endorsement. Read the notice and replication details.

The book appeared in 2021, before these developments. They change what we can responsibly recommend now; they do not show that Milkman knew the source would later be retracted. Nor do they erase independent evidence for vaccination plans, smoking-cessation commitments, or other interventions. Different studies need separate judgments.

The calorie-labeling passage moves in the opposite direction. It suggests that calorie information changes behavior essentially not at all, despite citing a study reporting a roughly 6% reduction in calories purchased per transaction. A later review also supports a small average reduction in calories

selected or purchased. The sensible criticism is that the effect is modest, not that the evidence shows nothing happened. [8]

These examples teach a useful habit for reading research-based books: ask what was counted. Calories purchased, calories eaten, and body weight are three outcomes. Finding mistakes, making mistakes, and completing work on time are three more. The more exciting the summary, the more important it is to keep the original measurement attached.

Planning is one of the book's stronger tools

The vaccination example is worth keeping precisely because it survives careful scrutiny.

In a workplace experiment, a prompt to write down both a date and time for getting a flu shot improved vaccination uptake. The control rate was 33.1%, and the observed difference between the groups was 4.0 percentage points. The researchers' adjusted estimate was 4.2 points. Relative to the control baseline, that adjusted difference was about 12.7%, which supports the book's rounded 13% figure. [9]

Four additional vaccinations per hundred eligible people may sound less spectacular than a 13% increase. Both descriptions can be accurate, but the first makes the result easier to picture. Neither deserves to be dismissed simply because it isn't a transformation for every participant.

Notice what the intervention had to work with: a convenient vaccination opportunity already existed, and participants received a reminder. The planning prompt was an addition to that arrangement, helping some people follow through on an action they could already take. The experiment was not trying to solve lack of access or deep opposition to vaccination with a blank calendar.

This is a good model for practical behavioral science: a defined population, a specific bottleneck, a small change, and an observable outcome.

The same standard applies to checklists. The book's original surgical example was a before-and-after implementation program, not a randomized test of a list alone. Ontario's later rollout did not reproduce its large outcome improvements. A randomized rollout in two Norwegian hospitals, however, found fewer complications and shorter stays, without a statistically significant overall mortality reduction. The evidence is not that checklists never help. It is that a checklist's value depends on how it changes care, and that the outcome and setting matter. See the original, Ontario, and Norwegian evidence.

The fresh-start savings experiment offers a similar lesson. Among 6,082 employees at four universities, describing a future starting date as a fresh start increased take-up by about 2.2–2.3 percentage points against a 7.6% baseline. The invitation was connected to payroll arrangements that could implement the choice. [10]

The timing mattered within a system that could implement the choice. The invitation did not merely make the starting date sound meaningful; it was attached to a usable enrollment process. A birthday message floating in isolation would be a different intervention, without that same route from invitation to action.

That distinction is supported by a less successful fresh-start application. A randomized medication-reminder trial did not improve adherence when reminders were timed or framed around birthdays and New Year. It tested a different action, comparison, and delivery system from the savings experiment. It does not overturn that savings result. It does tell us not to assume that attaching a meaningful date to any reminder will improve follow-through. See the medication and savings comparisons.

That's the practical principle I would carry forward: when the desired action is worthwhile and available, make the next step concrete. Don't ask a reminder to solve a problem that requires money, access, skill, or a different goal.

Temptation bundling has evidence—and a clock

Pairing something enjoyable with something worthwhile is an appealing idea. Milkman calls one version *temptation bundling*: reserve an indulgence for the activity you otherwise tend to avoid.

The original gym experiment randomized 226 participants to gym-restricted audiobooks, a suggestion to restrict audiobook use themselves, or a control group. In the first week, the gym-restricted audiobook group made 1.16 gym visits compared with 0.75 in the control group—about 55% more. That number is defensible. The roughly 51% figure often quoted from the paper is a different, adjusted estimate, so the two percentages reflect different calculations rather than an obvious contradiction. [11]

Milkman also reports that the benefit faded around a Thanksgiving interruption. This is important scientific honesty. She does not simply take the initial result and pretend it continued unchanged forever.

The key outcome was gym attendance. That's useful information, but it leaves open how much people exercised, how their fitness changed, and whether their behavior became automatic. A gym entry is an observable action, not a complete account of the benefit.

The larger follow-up is encouraging, but it did not test the teaching of bundling in isolation. Teaching was compared with audiobook provision within a wider program that also included other supports. Both comparisons used random assignment. For the no-audiobook comparison, however, assignment probabilities were unequal and changed across enrollment cohorts; the authors used statistical weighting and called the pooled analysis quasi-experimental. The added effect of teaching was more modest than a broad summary of the whole program can make it sound. That distinction matters if you want to try the idea without the accompanying program: the benefit of the full arrangement is not automatically the benefit of learning the technique alone. [12]

For a reader, the sensible experiment is simple: make an activity more enjoyable without interfering with its purpose. An absorbing audiobook might fit a walk. It might fit a treadmill session. It may be a poor partner for an activity requiring full attention.

Then look at what actually changes. Do you start more often? Stay longer? Enjoy it more? Does the arrangement survive a disrupted week? There is no prize for faithfully preserving a clever pairing that makes the activity worse. The book is most useful when it encourages this kind of testing rather than loyalty to the name of a technique.

Stronger pressure can produce a weaker program

Commitment devices make future behavior easier to enforce by changing the consequences of backing out. The book includes financial stakes, locked savings, and public commitments. Some examples have meaningful experimental support. [13]

But a stronger consequence is not automatically a better intervention. You also have to ask whether people will adopt it, whether they can afford it, and what happens when circumstances change.

Consider the smoking-cessation comparison discussed in the appendix. Reward programs were accepted by 90% of the people offered them, versus 13.7% for deposit programs. When people were assigned to the offers, six-month abstinence was higher under rewards. The conditional estimates were stronger for deposits among people willing to accept either arrangement. These answer different questions: how the offer performs across the people assigned to it, and how the arrangements compare for people willing to accept either. [14]

This is not a direct test of every “hard” versus “soft” commitment. It is a clear demonstration of the decision problem: the strongest tool among willing users can be the weaker offer for a whole population. A program that most people decline has a limited route to helping them.

Public commitments also deserve a setting-specific judgment. A small US trial found less inappropriate antibiotic prescribing after clinicians displayed signed commitment posters. A larger English trial did not improve its primary overall dispensing outcome. The posters, delivery, engagement, and outcome measure differed, so this was not an exact repeat. It is still an important warning against assuming that a promising local intervention will produce the same result when sent out across many practices. See the two poster trials.

The savings example raises another practical issue. More money held in a particular account is not necessarily the same as more total wealth or greater financial security. Restricted access may help with one temptation while creating a problem during an emergency. There is also a numerical distinction: the study scales its reported effect to account balances before the intervention, while the book’s illustration makes it sound like a comparison with the amount a control customer saved. Those are different baselines, so the illustration does not express the effect in the same way. [15]

Milkman deserves credit for discussing uptake and some of these costs. The revision I would make is to move those considerations from the edge of the story to the center of the decision.

Before adding a penalty, ask whether the action is the right one, whether the obstacle is actually temptation, and whether a lower-cost change would work. Stronger pressure is an intervention choice. It is not evidence of a more serious commitment to improvement.

Confidence and social influence require particular care

It is one thing to say that a positive message can encourage someone to act. It is another to say that the message changes their physiology in the way a book suggests. The confidence chapter needs us to keep those claims separate.

The housekeeper study and the milkshake experiment show why the measured outcome matters. In the housekeeper study, hotel workers were given information reframing their work as exercise, and the study reported health improvements. But unchanged activity was not established through continuous objective monitoring. A health change after reframing therefore doesn't by itself prove that expectations caused it independently of all behavior. In the milkshake experiment, participants consumed the same drink under different labels, and researchers observed different acute responses in the hormone ghrelin. That is not a long-term weight-change trial. These distinctions narrow the book's strongest interpretations without requiring us to dismiss expectations as irrelevant. [16] [17]

There is another reason to be cautious about the housekeeper story. In a 2011 replication attempt, researchers gave university building-service workers a similar message about their work being exercise. They did not reproduce the weight or body-fat benefits, although blood-pressure results favored the exercise-message group. The study was small and differed from the hotel experiment, so it does not settle every original result. But it makes the housekeeper finding harder to treat as a reliable demonstration that changing a belief improves health without changing behavior. This contrary evidence was available before the book appeared in 2021.

The same distinction matters for growth mindset. Evidence can support modest benefits under particular conditions without supporting the much larger impression that changing a belief reliably produces much greater achievement. A possible benefit in a specific setting is not a promise of transformation. Effective instruction, prior motivation, context, and the particular intervention still matter. [18]

Values affirmation has a similar limit. Reflecting on an important value is different from repeating flattering statements about yourself. Some educational studies reported benefits, but a large later cohort using procedures from an earlier successful replication found no benefit and ruled out effects larger than a small threshold. That makes values reflection harder to recommend as a reliably transferable school intervention. It does not show that reflection never helps anyone. See the large replication and its limits.

Social influence can also go in the wrong direction. The book includes an unusually useful example in which deliberately engineered peer groups harmed the very students the arrangement was meant to help. That is more informative than a collection of success stories. It shows how a plausible social design can generate a different pattern of interaction from the one its designers intended. [19]

Even familiar success stories need this check. Later German hotel studies did not consistently find that towel-reuse norm messages beat an environmental appeal, or that a same-room norm reliably helped more. Later meat-reduction studies also failed to establish a dependable advantage for messages about a behavior becoming more common. These studies measured different things, from intentions and reported consumption to actual towel reuse; they do not all repeat a café-purchase experiment. Together with the broader evidence, they support modest, conditional expectations—not a portable effect you can assume will appear whenever you describe what other people do. See the norm studies and follow-up results.

The lesson is to look at how people respond, not just how convincing the proposed explanation sounds. A comparison can inspire one person and discourage another. Advice-giving can help in a studied educational setting without becoming a universal confidence treatment. An intervention's label tells you what its designer hoped would happen; the outcome tells you more about what happened.

For readers, this argues for a grounded question: does this change make the useful behavior more likely under my actual circumstances? "This should increase confidence" is a reason to investigate, not a substitute for looking.

The ending is better than the fantasy of permanent change

One of the most useful parts of *How to Change* is its willingness to discuss how hard it can be to make a temporary gain last.

The gym megastudy described in the book tested many brief programs. Its subsequent publication examined 54 programs with 61,293 participants. About 45% of the programs significantly increased attendance during the intervention; about 8% produced significant increases afterward. Those percentages describe programs meeting a statistical threshold, not the percentage of participants transformed or the fraction of each effect that remained. [20]

Milkman's discussion does not hide the basic problem. Producing a temporary increase is easier than producing a lasting one. Removing the support often removes much of the gain.

There are two ways to react. One is to declare any intervention that needs maintenance a failure. The other is to ask whether the continuing benefit is worth the continuing cost. The second question is generally more useful for deciding what to do.

Suppose an appointment reminder helps you attend, but you would still forget without it. Keeping the reminder active may be an excellent arrangement. The point is to attend the appointment, not to prove that you can become a person who never needs reminders.

Still, continued support is not a universal requirement. A one-time choice has different maintenance needs from a recurring action, and learning or a lasting environmental change may continue to matter after the initial intervention. Recurring interventions also differ: some lose influence when they stop, while others leave some effect behind. The energy-report research in the appendix provides a specific test of persistence rather than an answer for all behavior. [21]

The stronger practical standard is this: plan for maintenance, account for its burden, and test what happens when support changes. Milkman's closing chapter brings that standard into view. It is a substantial reason to recommend the book despite the corrections it needs.

How I would use the book

Start with an outcome, not allegiance to a technique. "I want to be more active" leaves room to choose an activity that fits. "I must become the kind of person who runs at dawn" adds assumptions that may have nothing to do with the outcome.

Next, describe the point of failure. Are you forgetting? Is the action unpleasant? Is access difficult? Are you attempting something you do not actually value? Be concrete enough that a change could plausibly address the problem.

Imagine someone who repeatedly skips an exercise class. An earlier reminder helps only if forgetting is the obstacle. A class closer to home addresses travel. A different activity addresses dislike. A more realistic schedule addresses competing demands. These are illustrative possibilities, not diagnoses supplied by the book.

Choose a low-burden change and give it a fair test. Track the outcome you care about, alongside the cost of producing it. A reminder that increases attendance while generating constant annoyance may still be worthwhile, or it may be inferior to a simpler arrangement. You need both sides of that comparison.

Keep: specific plans, appropriately timed prompts, well-chosen defaults, enjoyable pairings, and the willingness to maintain useful support.

Modify: broad promises, relative percentages without baselines, and any attempt to infer your best intervention from a chapter label alone.

Reject: the inaccurate outcomes and numerical descriptions identified above, and reliance on the retracted deadline findings. A useful overall message is no reason to keep repeating a false example or treating an unreliable source as support. [22]

For organizations, add one more question: whose goal is being optimized? A manager's preferred behavior may impose costs on the employee. More engagement may benefit a product while wasting a customer's time. Calling an intervention "behavioral" does not settle whether the target is worth pursuing.

This is my proposed way of applying the book, not a tested replacement program. Its advantage is that failure produces information. Instead of asking whether you were sufficiently disciplined, you ask whether the chosen action, support, and setting were a good match. That is a much more productive conversation.

Why the rating is 72%

Scientific Accuracy: 75%. The book's strongest claim—that particular behavioral tools can improve particular actions—has good experimental support. Its matching framework is less established, and the strength of evidence varies across techniques.

Reference Accuracy: 67%. Most important sources can be identified. Several are described or applied inaccurately, including the carrot, opioid, and deadline examples. The deadline paper is now retracted and receives no supporting credit. Clear citations help readers inspect the evidence; they do not erase a mismatch or establish that a source remains reliable.

Practical Value: 75%. There are useful, often manageable interventions here, plus an unusually sensible discussion of maintenance. Their value depends on the goal, the setting, uptake, and ongoing burden. [23]

The overall rating gives these three categories equal weight. It is a summary of the review's judgments, not a finding that 72% of the book is true.

Read it as a toolkit, and test its suggestions against the outcome you actually want. Keep the tools that earn their place. The best parts of *How to Change* support that approach; its weakest moments are the ones in which a compelling explanation gets ahead of the evidence.

> **Want to inspect the evidence?** The technical appendix contains the full claim register, chapter assessments, source audit, score calculations, limitations, and evidence assessment. The reader essay and appendix make the same numerical judgments; they serve different reading needs.

Evidence notes

Each note links to the relevant passage of the technical appendix, which identifies the relevant book passages, original research, and source-access limitations. These notes are a reading aid, not a new claim that every source was rechecked for this essay.

1. The book’s “Carrot-eater” claim and the study’s actual outcome. Detailed assessment, source references, and access limits in the technical appendix.
2. The strongest fair reading. Detailed assessment, source references, and access limits in the technical appendix.
3. Three central scientific claims. Detailed assessment, source references, and access limits in the technical appendix.
4. Gerber et al. (2016): voter-identity field experiment. Detailed assessment, source references, and access limits in the technical appendix.
5. The book’s opioid-halving claim compared with the ten-pill default study. Detailed assessment, source references, and access limits in the technical appendix.
6. Chapter 1: Getting Started. Detailed assessment, source references, and access limits in the technical appendix.
7. The deadline study’s error-detection task and the book’s claim of 50% fewer errors. Detailed assessment, source references, and access limits in the technical appendix.
8. The book’s calorie-labelling claim compared with purchasing outcomes. Detailed assessment, source references, and access limits in the technical appendix.
9. Planning prompts increase flu vaccinations by 13%. Detailed assessment, source references, and access limits in the technical appendix.
10. Fresh-start savings invitations. Detailed assessment, source references, and access limits in the technical appendix.
11. Initial 55% exercise increase with bundling. Detailed assessment, source references, and access limits in the technical appendix.
12. Teaching temptation bundling: program design, comparisons, and persistence. Detailed assessment, source references, and access limits in the technical appendix.
13. Cash commitments help smokers quit. Detailed assessment, source references, and access limits in the technical appendix.
14. Halpern et al. (2015): four smoking-cessation incentive programs. Detailed assessment, source references, and access limits in the technical appendix.
15. The book’s claim of 80% more saving: account balances, denominators, and scope. Detailed assessment, source references, and access limits in the technical appendix.
16. The housekeeper study: reported health changes, activity measurement, and mechanism limits. Detailed assessment, source references, and access limits in the technical appendix.
17. Milkshake expectations alter ghrelin response. Detailed assessment, source references, and access limits in the technical appendix.
18. The growth-mindset claim: conditional benefits and limits on broad achievement promises. Detailed assessment, source references, and access limits in the technical appendix.
19. Engineered peer groups can backfire. Detailed assessment, source references, and access limits in the technical appendix.
20. Milkman, K. L.; Gromet, D.; Ho, H.; et al. (2021). Megastudies improve the impact of applied behavioural science . Nature . DOI: 10.1038/s41586-021-04128-4 Primary source / version examined. Detailed assessment, source references, and access limits in the technical appendix.
21. Energy-report effects decay after discontinuation. Detailed assessment, source references, and access limits in the technical appendix.
22. Practical use, limitations, and safeguards. Detailed assessment, source references, and access limits in the technical appendix.
23. How the ratings were determined. Detailed assessment, source references, and access limits in the technical appendix.

Disclosures and corrections

How we research and score these reviews.

Relevant interests: Hreha wrote the commercially available behavior-change book *Real Change* and offers behavioral-science services. These are potentially competing commercial and intellectual interests. Additional relationships, funding, review copies, sponsorships, and affiliate arrangements have not been comprehensively established; no blanket absence-of-interests claim is made.

Publisher: The Behavioral Scientist / Haystack Group LLC, identified in the site's terms. Editorial contact: Jason Hreha. This is not an official Red Pen Reviews review and is not affiliated with or endorsed by Red Pen Reviews.

Corrections: Send the passage, proposed correction, and sources through the contact form. Substantiated factual errors will be corrected; material changes after first publication will be dated. See the book-specific evidence assessment and read the full editorial notice for scope, score interpretation, professional-advice limitations, attribution, and correction policy.

Book reviewed: *How to Change*, Katy Milkman. 2021 Portfolio / Penguin ebook; ISBN 9780593083765. Evidence assessed through 13 September 2026. The assessment distinguishes evidence available before the book from research published afterward.

Housekeeper evidence assessment: The housekeeper discussion considers the 2011 replication attempt alongside the original study. See the evidence assessment and score explanation.

Evidence assessment: This review does not rely on the retracted deadline paper as supporting evidence. It assesses the voter, clinician-poster, norm, affirmation, fresh-start, and checklist findings, including later research. The scoring record explains the overall 72% rating.