The gap between intending and doing, *and what happens to fixes at scale*

This review follows the gap between intending and doing, and the remedies offered for it, from single experiments to a gym megastudy to government trials reaching millions.

MITWA LAB, MUMBAI

IN PLAIN WORDS

People often mean to do something and then do not. This review follows that gap at three sizes. When experiments raised intention on purpose, behaviour moved about half as far as intention did. Among 61,293 gym members, 45% of 53 programmes raised visits while they ran, and 8% did so in the four weeks after they ended. Across 126 government trials reaching over 23 million people, the average nudge moved behaviour by 1.4 percentage points, against 8.7 in published academic papers. A well-known theory of self-control, that willpower runs down with use, found no support in a preregistered test across 36 laboratories, though an exploratory analysis found a small effect. The field studies described here ran in one gym chain, one private gym and two government nudge units, and none of the sources read reports a test of a companion.

How far a change of mind carries into behaviour

An intention, in this literature, is a person's stated commitment to act, and the gap is the distance between that commitment and what the person later does. How large the gap looks depends on what is compared. Conner and Norman (2022) summarise reviews of correlational studies, in which intention is measured and behaviour is then observed, and report that intentions account for 18 to 23% of the variance in behaviour. The share is lower when behaviour is measured objectively. One review found that intention and perceived control accounted for 26% of the variation in self-reported physical activity and 12% in objective measures.

Experiments ask whether behaviour follows when intention is moved on purpose. Webb and Sheeran (2006) pooled 47 experiments in which people were randomly assigned to something that raised the strength of their intention relative to a control condition, and behaviour was compared afterwards. Intention rose by d = 0.66, a medium-to-large step (d is the difference between group averages in standard deviations), and behaviour by d = 0.36, a small-to-medium one. Only the abstract of that paper was read for this review. Conner and Norman (2022) add that reviews of such experiments find correlations (r+) of 0.08 to 0.18, against 0.40 to 0.48 in correlational reviews. The experiments were selected because they did change intention, so they describe what happens when the first step works.

For physical activity, Bieleke et al. (2021), in a review of if-then planning, say that only about half of the people who want to be active do what they intend. They attribute the figure to a 2013 meta-analysis that is cited here only through them, so it is a round number at second hand.

Willpower as a limited resource, tested in 36 laboratories

One well-known theory of self-control holds that it works like a limited resource, so that using it on one task makes later self-control less successful (Vohs et al., 2021). This is ego depletion, which in everyday terms says that willpower runs down with use. It is included here because, if it held, it would be one reason for an intention to go unmet, although the

intention-behaviour sources read here do not link the two.

Inzlicht and Friese (2019) put the number of studies taken to support the theory at about 600. They describe a registered replication report by 23 laboratories and over 2,000 participants, which found no evidence of the effect, and 22 of the 23 laboratories had predicted beforehand that they would replicate it. Several scholars criticised that study, and Inzlicht and Friese say it is hard to know how valid a test it was, since it replicated one specific depletion study.

Vohs et al. (2021) designed a second test to meet that objection. The procedures were chosen as representative of the field instead of copied from one paper, and the analysis plan was fixed in advance. Thirty-six laboratories tested 3,531 people, with a mean age of about 21, but the preregistered exclusions removed nearly a third of the sample and left 2,461 people in the confirmatory analysis. There the difference was d = 0.06, with an interval from -0.02 to 0.14. A Bayesian comparison of the null against an informed alternative centred on 0.30 found the data four times more likely under the null. An exploratory analysis of the full sample, ignoring the preregistered exclusions, did reach significance at d = 0.08, and there the two hypotheses fitted the data about equally well. The effect was also larger among participants who reported more fatigue, though that came from exploratory tests.

Inzlicht and Friese (2019) stop short of dismissing the idea. They describe the evidence as uncertain and cite an analysis that judged it inconclusive. They point to studies in which paying participants, or being kind to them, eased the effect, which is hard to understand if depletion is an incapacity like an empty fuel tank. They also suggest that depletion may be mental fatigue, which has been linked to real-life outcomes in studies whose results are sometimes contradictory, and they say it is unclear how far typical laboratory tasks capture it. The 36-laboratory test used tasks lasting minutes, and no one was followed to a gym.

Planning for when and where, in two kinds of test

One remedy aimed at the gap is the if-then plan, called an implementation intention, which links a named opportunity or obstacle to a planned response. Sheeran et al. (2024), authors of earlier reviews of the technique, pooled 642 tests. The sample-weighted average effect was d = 0.36, and across 301 tests of behaviour it was d = 0.27. For physical activity, 64 tests gave d = 0.14 (interval 0.05 to 0.22). The intervals for vaccination and voting included zero, while effects on cancer screening and medication intake were small to medium. Effects were larger in laboratories than in the field, and larger when people's intentions were already strong. They were also larger for objective measures than for self-report, and did not shrink as the follow-up lengthened.

The review's authors state some limits. It was not preregistered, the median test had about 30 people per condition, and a regression test for small-study effects indicated substantial publication bias. A trim-and-fill correction left the estimate at d = 0.35, but a Bayesian model that averages over bias models gave d = 0.15 (interval 0.08 to 0.22), with wide variation between tests. The authors call that effect small by conventional guidelines. On this evidence planning probably does something, but the size is uncertain, depends on the behaviour and the person, and is small for exercise.

Carrera et al. (2018) tested a simple version in one preregistered field experiment at a private gym in a Midwestern US city. The 877 participants were members who completed an online survey, out of 1,210 invited. The planning group chose the days and times at which they planned to go over the next two weeks, and the control group recorded when they had exercised in the previous two weeks. Visits came from check-in records. The planning group averaged 2.3 visits and the control group 2.6, and the 95% interval excludes an increase of more than 2%. People had engaged with the planning task, since most believed planning helps and the plans they made predicted the days they attended.

The two sources do not agree on vaccination and voting. Carrera et al. (2018) cite large field experiments in which planning raised vaccination, screening and voting, while the pooled intervals for vaccination and voting in Sheeran et al. (2024) include zero. Carrera et al. also suggest reasons that repeated behaviour might differ, namely existing routines, a plan that shifts visits instead of adding them, and weaker cues. They report some related checks but no direct test of each.

Fifty-four programmes in one gym chain

A megastudy compares many interventions in one population, on one objectively measured outcome, over the same period (Milkman et al., 2021). The 61,293 participants were members of an American gym chain who signed up for a free programme with small cash rewards, and 30 scientists from 15 universities designed 54 four-week digital programmes. The placebo group visited 1.48 times a

week while the programmes ran. The authors say they could not determine how representative the final sample is of the chain's members.

Of the 53 experimental conditions, 45% significantly raised weekly visits over placebo, by 9% to 27%. Fifty-two of the 53 conditions added features to a baseline of planning, reminders and microincentives, which alone produced about 0.14 extra weekly visits (9%). All 24 conditions that raised visits, the baseline among them, included planning, reminders and incentives, typically with a further nudge or reward. The top condition, a small bonus for returning after a missed workout, produced about 0.40 extra weekly visits (27%), and the authors warn that top performers' estimates are probably inflated. It follows from the design that the study cannot say whether the planning component did anything separately, because no condition removed it.

After the programmes ended, only 8% of conditions produced a significant rise in visits over the following four weeks, against 2.5% expected by chance. Pooling all 53 conditions, each extra visit during the programme went with 0.28 extra visits over the next ten weeks (interval 0.07 to 0.59). Forecasts by 301 online participants, 156 public health professors and 90 practitioners were not reliably correlated with the observed effects (weighted pooled r = 0.02) and were on average 9.1 times too optimistic. The authors add that larger incentives or in-person coaching would probably have produced larger effects.

How long a habit takes

Habit formation is widely discussed as part of lasting behaviour change, and the question is how long it takes. Singh et al. (2024) reviewed 20 studies of health habits with 2,601 participants, of which 11 were rated at high risk of bias. Four studies reported time to habit, with medians of 59 to 66 days and means of 91 to 154 days (the abstract gives 106 to 154, and the text adds an earlier study's mean of 91). In one healthy-eating study the median was 59 days, individuals ranged from 4 to 335 days, and only 23% reached the study's threshold. The review's authors note that few studies were available and that flossing and diet had the most evidence.

Setting those figures beside the megastudy is our own comparison, and the constructs differ, because habit studies measure self-reported automaticity and the megastudy counted visits. A four-week programme is shorter than every median the review lists, which fits a small carry-over without showing that habit explains it.

Trials that reached 23 million people

Nudges are changes to how a choice is presented that do not forbid options or significantly change economic incentives. The government trials typically combined several, such as simplified letters, personalised messages, planning requests, reminders and social-norm comparisons. DellaVigna and Linos (2022) obtained the records of every trial run by two United States government nudge units, 165 in all, and kept 126 randomised trials with a clear control group, a yes-or-no outcome and no financial incentives, covering over 23 million people. They compared these with 26 published academic trials (505,337 participants) taken from two earlier meta-analyses. The figures here are from the working paper version, which had not been peer reviewed.

Academic papers averaged an 8.7 percentage point rise in take-up, 33.5% above the control group. The government trials averaged 1.4 points, an 8.1% rise, which the authors call sizable and highly significant statistically. The median arm held 484 people in the academic sample and 10,006 in the government trials, giving minimum detectable effects of 6.3 and 0.8 points. A model of selective publication estimated that trials without significant results reach print with probability about 0.1 (standard error 0.1), and corrected the academic average to 3.2 points (standard error 1.9). The authors conclude that publication bias, worsened by low power, can account for the whole gap, and that academic involvement cannot, though differences in nudge features, such as in-person against letter-based contact, also partly explain it. They argue that 1.4 is probably a lower bound, because the units largely rule out default changes and organisations can adopt the best arm.

Maier et al. (2022) reanalysed a published meta-analysis of nudging, which its authors describe in their reply as covering more than 200 studies (Mertens et al., 2022), and corrected for publication bias with a Bayesian model. The uncorrected average was d = 0.43. After correction it was 0.04 (credible interval 0.00 to 0.14), and the Bayes factor for the null in the combined sample was 0.95, so on that number alone the data favoured neither side. The authors nonetheless conclude that once the bias is corrected no evidence remains for nudges as effective tools of behaviour change. A second analysis, using only the most precise estimate per study, gave 0.11 and weak evidence for an effect. Evidence of variation between studies appeared in every category and every domain except finance, which the letter takes to mean that some nudges might work even where the average does not. The original authors had already reported d = 0.08 under severe bias, and their reply agrees that

publication bias and unexplained variation are problems (Mertens et al., 2022). The two analyses use different yardsticks, percentage points of take-up against standardised effects, so their numbers cannot be laid side by side.

What shrinks and what stays

The larger and more fully reported the evidence, the smaller the effect in the nudge and planning literatures, and a well-known theory of self-control, depleted willpower, found no support in a preregistered test. Small positive effects are reported at each scale, some of them uncorrected for publication bias. Behaviour moved by d = 0.36 when intention did, in pooled experiments whose abstract does not say they were checked for publication bias. A bundle of planning, reminders and microincentives raised gym visits by 9%, and government nudge trials raised take-up by 1.4 points. Where correction has been applied the estimates shrink, to d = 0.15 for planning (Sheeran et al., 2024) and 0.04 for nudging (Maier et al., 2022). Whether any of this lasts is where the evidence is weakest, since only 8% of programmes in one gym chain produced a significant rise afterwards, although the pooled estimate across all 53 was a small positive carry-over. This reading is ours and goes beyond any single source. None of the sources read here reports a test of a companion, so this review says nothing about talking with one, and nothing here is a claim about what any Mitwa model does.

What is not yet known

Among the sources read here, Carrera et al. (2018) is the only preregistered field test of planning alone for a repeated behaviour, so it is open whether that null reflects exercise, that gym, or the form of the request.

The 8% persistence figure comes from one chain, one behaviour and four-week programmes. Whether longer programmes or other behaviours keep more of their effect is not measured in the studies read here, apart from the earlier gym studies that Milkman et al. (2021) compare with.

It is unresolved which nudges account for the variation that Maier et al. (2022) report, and whether a near-zero average hides some effective designs. The published version of the government-unit paper was not read, so differences from the working paper are unchecked.

Depletion may survive as mental fatigue. The sources here report mixed real-world evidence that fatigue affects outcomes such as medical decisions and compliance, and they say it is unclear how far typical laboratory tasks capture it.

References

Bieleke, M., Keller, L., and Gollwitzer, P. M. (2021). If-then planning. European Review of Social Psychology, 32(1), 88-122. doi.org/10.1080/10463283.2020.1808936

Carrera, M., Royer, H., Stehr, M., Sydnor, J., and Taubinsky, D. (2018). The limits of simple implementation intentions: Evidence from a field experiment on making plans to exercise. Journal of Health Economics, 62, 95-104. pmc.ncbi.nlm.nih.gov/articles/PMC7872013/

Conner, M., and Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. Frontiers in Psychology, 13, 923464. pmc.ncbi.nlm.nih.gov/articles/PMC9386038/

DellaVigna, S., and Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. Econometrica, 90(1), 81-116. doi.org/10.3982/ECTA18709. Figures read in the NBER Working Paper 27594 version of July 2020: www.nber.org/system/files/working_papers/w27594/w27594.pdf

Inzlicht, M., and Friese, M. (2019). The past, present, and future of ego depletion. Social Psychology, 50(5-6), 370-378. Read as the authors' PsyArXiv preprint: osf.io/q8r4m

Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., and Wagenmakers, E. J. (2022). No evidence for nudging after adjusting for publication bias. Proceedings of the National Academy of Sciences of the United States of America, 119(31), e2200300119. pmc.ncbi.nlm.nih.gov/articles/PMC9351501/

Mertens, S., Herberz, M., Hahnel, U. J. J., and Brosch, T. (2022). Reply to Maier et al., Szaszi et al., and Bakdash and Marusich: The present and future of choice architecture research. Proceedings of the National Academy of Sciences of the United States of America, 119(31), e2202928119. pmc.ncbi.nlm.nih.gov/articles/PMC9351449/

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., et al. (2021). Megastudies improve the impact of applied behavioural science. Nature, 600(7889), 478-483. pmc.ncbi.nlm.nih.gov/articles/PMC8822539/

Sheeran, P., Listrom, O., and Gollwitzer, P. M. (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. kops.uni-konstanz.de/handle/123456789/69905

Singh, B., Murphy, A., Maher, C., and Smith, A. E. (2024). Time to form a habit: A systematic review and meta-analysis of health behaviour habit formation and its determinants. Healthcare, 12(23), 2488. pmc.ncbi.nlm.nih.gov/articles/PMC11641623/

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., et al. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. Psychological Science, 32(10), 1566-1581. pmc.ncbi.nlm.nih.gov/articles/PMC12422598/

Webb, T. L., and Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. Psychological Bulletin, 132(2), 249-268. Abstract only. pubmed.ncbi.nlm.nih.gov/16536643/