Kaiser, Tim; Kloidt, Juliane; Mata, Jutta; Hertwig, Ralph

Working Paper

A Meta-Meta-Analysis of Behavior Change Interventions: Two Tales of Behavior Change

CESifo Working Paper, No. 11863

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Kaiser, Tim; Kloidt, Juliane; Mata, Jutta; Hertwig, Ralph (2025) : A Meta-Meta-Analysis of Behavior Change Interventions: Two Tales of Behavior Change, CESifo Working Paper, No. 11863, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/320084

CESifo WORKING PAPERS

11863
2025

May 2025

A Meta-Meta-Analysis of Behavior Change Interventions: Two Tales of Behavior Change

Tim Kaiser, Juliane Kloidt, Jutta Mata, Ralph Hertwig

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.ifo.de/en/cesifo/publications/cesifo-working-papers
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.ifo.de/en/cesifo/publications/cesifo-working-papers

A meta-meta-analysis of behavior change interventions: Two tales of behavior change

Authors: Tim Kaiser[1,2,3,4*], Juliane Kloidt[5], Jutta Mata[6,7], Ralph Hertwig[7]

Affiliations:

[1]Department of Business and Economics, University of Kaiserslautern-Landau (RPTU); Landau, D-76829, Germany.

[2]Kiel Institute for the World Economy (IfW Kiel); Kiel, D-24105, Germany.

[3]CESifo; Munich, D-81679, Germany.

[4]IZA Institute of Labor Economics; Bonn, D-53113, Germany.

[5]School of Psychology and Neuroscience, University of Glasgow; Glasgow, G128QB, United Kingdom.

[6]School of Social Sciences, University of Mannheim; Mannheim, D-68161, Germany.

[7]Center for Adaptive Rationality, Max Planck Institute for Human Development; Berlin, D-14195, Germany.

*Corresponding author. Email: tim.kaiser@rptu.de

Abstract: Behavioral science interventions like incentives, nudges, and boosts are increasingly used in public policy, but their effectiveness remains debated. We conducted a meta-meta-analysis on behavior change interventions across health, finance, and sustainability outcomes. Our analysis covers 838 effects from 269 meta-analyses, encompassing 6,327 randomized controlled trials and over 9 million individuals from non-clinical populations of all ages in both developed and developing economies. Our findings tell two stories: First, extracted treatment effects are generally positive but highly variable (M = .173; SD = .195), indicating some interventions impact behavior. However, after adjusting for publication bias, the mean posterior effect pooling domains and interventions is .063 (95% credible interval .044 to .08, BF_{10} = 139.8) with substantial unexplained heterogeneity ($\hat{\tau}$ = .129). Future research requires improved reporting and deeper contextual analysis to address this heterogeneity. Even small effect sizes can yield significant impacts when scaled across populations and sustained over time.

Human behavior is at the heart of both the causes of and the solutions to many global crises—from the climate and biodiversity crisis, antimicrobial resistance, pandemics, and the obesity crisis to misinformation and democratic backsliding. For instance, over 50% of the carbon emission reductions needed to reach net zero are estimated to depend on behavior change [1]. Similarly, reducing meat consumption, which is currently too high to be environmentally sustainable [2], is only possible with behavior change—no technical approach can make food production efficient enough to sufficiently reduce its environmental impact at current consumption levels (e.g., [3]).

The critical role of human behavior in both creating and solving global problems has raised an important question: How can individual behavior be harnessed to help solve these problems? Several categories of policy interventions aim at changing individual behavior, including legal and ethical interventions (e.g., regulations or ethical guidelines), educational interventions (e.g., compulsory schooling), and targeted behavior change interventions, which use behavioral science to change individual behavior to benefit individuals and societies. Each of these policy approaches to behavior change has advantages and disadvantages. Regulations, for instance, are relatively slow and subject to lobbying interest and "blood-sport strategies" [4]. Ethical guidelines lack teeth unless they are enforced whereas compulsory schooling can be perceived as coercive and restrictive of independent thought. Finally, behavioral science-based interventions may ultimately be much ado about nothing (we return to this possibility shortly). One response to this mixed scorecard is to advise policy makers to systematically integrate classes of interventions rather than betting on just one. But even then, a realistic assessment of the interventions' likely impact (i.e., causal effects) and robustness across diverse environmental and social contexts is essential.

Behavioral science-based interventions have become widely employed in public policy to guide individual behaviors and decisions in a variety of domains, such as health, household finance, and environmentally sustainable consumer choice [5,6]. One classic approach relies on *economic incentives* [7], which can be direct or indirect price incentives (as either monetary or non-monetary rewards) for target behaviors [8,9]. A second approach implements changes to the *choice architecture* that harnesses individuals' behavioral tendencies and biases without limiting their freedom of choice or changing economic incentives. These "nudges" [10–12] might bet on individuals displaying a status quo bias or loss aversion to set defaults that people would not bother to change. A third class of interventions focuses on *empowering* individuals with the competences to overcome cognitive biases and motivational hurdles and make informed choices [13–16]. "Boosts" provide individuals, teams, or institutions with relevant and actionable knowledge, with human capital in the form of decision-making competences, or with metacognitive strategies to help them make autonomous decisions that are in line with their goals, without changing monetary incentives.

We conducted a meta-meta-analysis—that is, a meta-analysis of existing meta-analyses of the success of behavior change interventions. We focused on behavior change interventions for two reasons. First, they have attracted considerable attention from policy makers over the past two decades due to their promise of cost effectiveness (i.e., small changes resulting in big changes; [11,17]) and liberty preservation [18]. Second, the success of behavioral science-based interventions can be empirically evaluated. Recent evaluations appear to suggest that the empirical record is mixed, offering two distinct tales of behavior change: one encouraging, one grim.

The encouraging tale can be found in a widely cited recent meta-analysis of nudging interventions that concluded that "choice architecture interventions successfully promote behavior change across key behavioral domains, populations, and locations" (p. 1; [19]). Overall behavior change was estimated to have a small to medium effect size (Cohen's d = .43, 95% *CI* [.38, .48]). However, re-analyses [20–22] of this meta-analysis suggest no evidence for average effects of choice architecture interventions after adjusting for publication bias. Additionally, heterogeneity in treatment effects was found to be high, suggesting little reason to expect globally consistent effects. This is in line with a recent meta-analysis comparing the soberingly small effect sizes in trials run by nudge units in governments with the larger effect sizes obtained in nudge trials reported in academic publications [23]. Overall, the effects of interventions in pilot contexts differ from those of interventions at scale, leading to uncertainty and disagreement around the practical implications of behavior change interventions (e.g., [7,23–25]). This is the grim tale.

At least partly in response to the large variety of contexts that behavior change interventions are applied in, a growing strand of literature in psychology and the behavioral sciences deviates from the traditional approach of conceptualizing causal treatment effects as a single parameter that represents the true effect of a class of behavioral interventions [26]. Instead, this literature conceptualizes causal treatment effects in terms of their heterogeneity and a distribution of possible treatment effects that vary with context and individual characteristics. Discrepancies between academic and policy interventions may suggest that meta-analytical estimates of effects are inflated due to questionable research practices such as p-hacking and publication bias [19–22,27–36], which occurs when studies that do not find the desired outcome are not published. Publication bias can stem from a perceived or real lack of interest in the "failed" studies and nil results on the part of editors, reviewers, or colleagues; it can also be due to conflicts of interest that lead to the suppression of results that do not fit a particular research agenda [37].

The reasons behind heterogeneity, publication bias, and p-hacking exist in literatures beyond recent meta-analytical examinations of choice architecture interventions. We therefore sought to empirically inform the discourse on individual behavior change interventions more generally [7,15]. To this end, we aggregated the meta-analytical evidence on behavior change interventions across domains and intervention classes and analyzed the impact of publication bias and the extent of heterogeneity in treatment effects. This is, to our knowledge, the first meta-meta-analysis of behavior change interventions. Our goal was to estimate realistic effect sizes of three classes of behavior change interventions—choice architecture, infopowerment (a word we coined to capture interventions that often provide both information and empowerment strategies), and incentives—across a range of key domains, including health (e.g., obesity, physical activity, or substance use), financial behaviors, and sustainable consumption. Would we find again two different tales, one encouraging and one grim, reflecting the distorting effects of publication bias? Or would we find a more nuanced picture, with varied effects for different classes of behavioral science interventions and across domains?

Results

The search process for this meta-meta-analysis of behavior change interventions retrieved 41,406 initial records (Fig. 1). The final sample contained 269 meta-analyses with 838 unique treatment effect estimates that met our inclusion criteria. We included meta-analyses published across three decades, yet most fell within the most recent eight years of our literature search. This emphasis

on recent publications reflects a growing interest in behavioral science interventions as well as the increased availability of high-quality randomized controlled trials (RCTs) to synthesize evidence (Fig. 2a). In total, the included meta-analyses synthesized evidence from 6,327 RCTs, involving more than nine million female and male participants of all ages from non-clinical populations.

Although the diversity of targeted behaviors increased over time, most meta-analyses pertained to the domain of health (n = 260). Of the meta-analyses that met the inclusion criteria, very few covered finance (n = 5), sustainability (n = 3), or a combination of health and sustainability (n = 1). We therefore split extracted estimates from the health domain into nine distinct groups, arriving at a total of 11 outcome behaviors that were used for all subgroup analyses: (1) overweight and obesity, (2) substance use, (3) physical activity and sedentary behavior, (4) general health, (5) sexual health, (6) eating and nutrition, (7) medical procedures, (8) feeding practices, (9) oral health, (10) financial behaviors, and (11) sustainable consumption (see Table 1 for definitions and examples). On average, each outcome behavior included 63 extracted estimates, ranging from 9 (sustainable consumption) to 195 (obesity).

We further categorized treatment effects into three intervention classes: choice architecture, infopowerment, and incentives (Table 1). This process, however, was often hampered by unspecific descriptions of interventions within meta-analyses, mixes of different interventions within one meta-analysis, and the requirement to categorize treatment effects into only one intervention class (see Supplementary Materials). As a result, we could only categorize about 20% (k = 167) of treatment effects, with most of those effects representing infopowerment interventions (k = 127), followed by incentives (k = 39). Within the time period analyzed, only one treatment effect was derived from pure choice architecture interventions (d = .352). As a consequence of this highly skewed distribution, we dropped the class of choice architecture interventions from further analyses.

For the sake of comparability, we converted treatment effects into standardized mean differences d, with positive effects depicting intended outcomes (e.g., increased savings, weight loss; see Supplementary Materials). We then took a bird eye's view to investigate the interventions: We first described the raw effects distributions from the included meta-analyses, then analyzed whether interventions produced the intended outcomes, considering meta-analytical estimates of the means of distributions of true effects (i.e., random-effects meta-analysis) and robust Bayesian meta-analysis (RoBMA) accounting for publication bias.

Description of extracted effects

The unweighted mean of the 838 extracted effects was .173 (Mdn = .138), indicating that, on average, meta-analyses of behavior change interventions reported the intended positive directional changes in behavior (Fig. 2b). The extracted estimates varied considerably (SD = .195), suggesting that some studies reported much larger effects, whereas others produced zero effects or unintended outcomes. Many of the larger effects were accompanied by small inverse standard errors, indicating that these estimates were less precise. To investigate whether the large spread of the extracted estimates reflected distinct subgroups, we split the estimates by intervention class, outcome behavior, and combinations of both.

For the modest subset of effects that we categorized by intervention class, the unweighted mean effect was .176 (*SD* = .204) for infopowerment and .241 (*SD* = .205; Fig. 2c) for incentives. Although mean infopowerment effects were smaller than mean incentive effects, the accompanying standard deviations were large, indicating considerable spread within and substantial overlap between the distributions. Infopowerment interventions were also applied to all outcome behaviors, whereas incentive interventions were only applied to some (e.g., substance use, physical activity, general health).

Across outcome behaviors, samples with widely spread estimates had larger unweighted mean effects than samples with more tightly clustered raw estimates (Fig. 2d). For outcome behaviors that only included few estimates (e.g., oral health, feeding practices), larger mean estimates may therefore be driven by outliers. It is noteworthy that the relatively modestly studied outcome of financial behaviors showed the smallest mean effect and the least variance (*d* = .047; *SD* = .025; *k* = 13), perhaps describing a more homogenous conceptual and interventional focus in the field.

Due to the large variability found for most outcome behaviors, we also studied effect size distributions for unique combinations of intervention class and outcome behavior where at least two estimates were available (see Table S1). The largest effects were reported for incentives × substance use (*d* = .315, *SD* = .168; *k* = 14) and the smallest effects were reported for incentives × physical activity (*d* = .070, *SD* = .157; *k* = 4). In contrast, the largest effects for infopowerment interventions were reported for physical activity (*d* = .313, *SD* = .199; *k* = 18) and the smallest effects for obesity (*d* = .099, *SD* = .137; *k* = 19).

Overall, the descriptive examination of the extracted effects may support a cautious interpretation of the encouraging tale of behavioral science-based interventions: Extracted effects of behavior change interventions showed desired outcomes across our full sample, all intervention classes, and all outcome behaviors. Importantly, however, the extracted effect sizes showed similarly large standard deviations across all subsamples split by intervention, outcome behavior, and both. This persistent variability suggests additional sources of heterogeneity that should be taken into account.

Sources of heterogeneity in extracted effects

The substantial variance in the extracted effect size distributions suggests systematic heterogeneity that could reflect a variety of analytical decisions and/or study sampling procedures. Using a specification curve (Fig. 3), we investigated whether effect sizes of similar magnitude shared selected characteristics. The top panel of Fig. 3 depicts all extracted effects organized by size. The panels below categorize each effect according to seven factors (statistical power, outcome domain, outcome behavior, intervention class, delivery mode, control group, and intervention setting). Each factor includes several mutually exclusive characteristics. A vertical tile suggests the presence of the characteristic; a lack of a vertical tile suggests the absence of either the characteristic or conclusive information.

We first determined which estimates were derived from studies with sufficient power to detect effects of .14 (median raw effect), .10, and .05. Most meta-analyses at the upper tail of the empirical distribution (i.e., larger effect sizes) were not sufficiently powered to detect effects smaller than .10 or .05, suggesting that low power and publication bias may be a problem in this

literature. Considering only estimates from well-powered studies would truncate the empirical distribution of uncorrected treatment effects around very small, yet positive, estimates.

We next evaluated whether empirical distributions differed across various study sampling procedures. Despite our efforts to include a diverse range of behavioral outcomes, interventions, and populations, the resulting sample primarily focused on health behavior change using complex interventions and various types of control groups in individuals from high-income economies.

We did not observe systematic sampling trends across domains or outcomes since most treatment effects targeted health behaviors. Relatedly, we did not observe systematic similar trends across different interventions (likely because most treatment effects could not be classified) or using one versus multiple modes of intervention delivery (e.g., in-person, online, written). Similarly to intervention classifications, heterogenous control groups were routinely combined at the meta-analytic level. We therefore did not observe patterns in our broad categorization of control groups, including no treatment (k = 128), standard care (k = 72), attenuated care (k = 39), and other treatment (k = 73). Lastly, splitting our sample by economic setting did not reveal systematic trends likely because most participants were from advanced economies, with more than 50% being from the United States (based on information available from 158 meta-analyses).

Overall, Fig. 3 illustrates that effect sizes differed according to various analytical decisions and sampling characteristics; however, other than statistical power, the presence or absence of a single characteristic was not systematically associated with higher or lower effectiveness. A bird's eye meta-meta-analytic view may not be the optimal approach for detecting systematic differences, as the true sources of heterogeneity were aggregated twice before entering our analyses (i.e., in the primary study and in the meta-analysis).

Naïve inference from random-effects models

Considering widely spread distributions of extracted effects (Fig. 2) and differing combinations of study characteristics (Fig. 3), our naïve model assumed heterogeneity in true effects at the population level. Because we extracted up to 23 effects from the same meta-analysis, we also assumed within-study dependencies in our dataset. To account for both, we implemented random-effects meta-analytic models using robust variance estimation (RVE; [38]) and multivariate random-effects (MRE; [39]). Here, we focus on RVE and present results from both procedures in Table S2.

The pooled RVE effect was .171 (SE = .008), which was similar to the pooled raw estimate (d = .173; Fig. 4). The between-study heterogeneity was τ = .102, suggesting a widely spread distribution of true effects in the RVE model. Indeed, 90.95% of the model's variance was attributed to heterogeneity in true effects (I^2), suggesting that less than 10% of the variance was due to random sampling error. Compared to the raw mean effects, the general RVE effect for infopowerment interventions decreased slightly (M = .159; SE = .018), whereas the mean RVE effect for incentive interventions increased (M = .316; SE = .052). The ratio of true heterogeneity to total variance (I^2) remained high at 93.97% and 85.50%, respectively, suggesting that most variance in true effects was still due to unobserved between-study differences.

Across the 11 outcome behaviors, mean RVE effects differed slightly from raw mean effects, translating into minor changes in rank order of effectiveness (Fig. 4, Table S2). Heterogeneity ratios (I^2) were highest for oral health (95.89%) and medical procedures (95.27%) and lowest for sustainable consumption (61.74%), sexual health (74.56%), and financial behavior (74.63%). For some outcome behaviors, heterogeneity ratios may reflect the variety of included behaviors (e.g., many behaviors categorized as medical procedures, few behaviors categorized as sexual health or financial behaviors). For other outcome behaviors (e.g., oral health, sustainable consumption), heterogeneity ratios may instead reflect unique study characteristics and/or analytical decisions for each included effect.

Mean RVE effects for combinations of intervention class and outcome behavior also presented with minor differences compared to descriptive mean effects. Whereas heterogeneity ratios (I^2) were substantially lower for incentives × general health (32.54%) and incentives × substance use (35.77%), heterogeneity ratios remained substantial for most infopowerment combinations reflecting the broad inclusion criteria for this intervention class.

Overall, naïve RVE random-effects models revealed mean effects of a similar magnitude to the means from our descriptive data analyses. Because of substantial heterogeneity, however, the results do not support claims of one true effect size (a fixed parameter), even in small subsamples of intervention–outcome combinations. The RVE results may therefore support a version of the encouraging tale, but these results should be interpreted with caution given the large variance in effect distributions.

Adjusting for publication bias

While RVE random-effects models account for differences in the precision of the underlying estimates and model their residual heterogeneity, true effects of behavior change interventions in the population may be masked by publication bias. Only 62% of included meta-analyses (168 out of 269) addressed publication bias, using various assumptions and correction methods. We therefore extracted uncorrected treatment effects and adjusted all meta-meta-estimates ex post.

We applied various methods, including selection models [27,40], weighted average of the adequately powered (WAAP; [41]), and Robust Bayesian Meta-Analysis (RoBMA) [42]. Here we focus on the current state-of-the-art publication bias correction method RoBMA to estimate publication bias-corrected mean posterior effect and to quantify how much the adjusted effects are deflated. Please note that compared to the other publication bias correction methods, the RoBMA-corrected effects were by far the smallest estimates (see Table S3).

For all estimates combined, the mean RoBMA-corrected posterior effect was .063 (95% credible intervals [CI] = [.043, .082]), suggesting a much smaller estimate than the uncorrected effect distributions (d = .173) or the random-effects model (d = .171; Fig. 4). Still, the RoBMA-corrected effect remained positive, the credible intervals ruled out zero effects, and the corresponding Bayes factor suggested strong evidence for the presence of an effect (BF_{10} = 71.99). The heterogeneity estimate (τ = .129; 95% CI = [.120, .138]) and the corresponding Bayes factor ($BF^{rt} \rightarrow \infty$), however, confirm substantial residual heterogeneity from the naïve RVE model. The RoBMA model also returned extreme evidence for publication bias (BF_{pb} = 1.02^{27}), likely aggregating publication bias from the included meta-analyses and their primary studies.

For infopowerment interventions, RoBMA results showed very weak evidence against the presence of an effect, extreme heterogeneity, and extreme publication bias (d = .025; 95% CI = [.000, .114]). For incentive interventions, RoBMA results suggested strong evidence of an effect, extreme heterogeneity, and weak evidence against publication bias (d = .215; 95% CI = [.000; .282]). The RoBMA posterior mean for incentive interventions was by far the largest meta-analytic mean for all analyzed sub-samples, even exceeding the posterior means of all depicted incentive-outcome combinations (Fig. 4). This is because we only performed publication bias correction for intervention-outcome subgroups with at least 5 extracted effects but included isolated effects (e.g., on physical activity) in the analysis on all incentive interventions. When performing RoBMA only on incentive effects that were included in an intervention-outcome subgroup (i.e., incentives × substance use, incentives × general health), the posterior mean would approximately be half in size (d = .117; 95% CI = [.000, .260]) with weak evidence for the presence of an effect, extreme heterogeneity, and moderate evidence for publication bias.

For all 11 outcome behaviors, RoBMA showed posterior means of $d < .10$. Estimates of posterior mean effects were statistically insignificant for obesity, medical procedures, and oral health. Substance use and general health showed weak evidence for the presence of an effect; all other outcome behaviors showed weak to strong evidence against the presence of an effect. All outcome behaviors showed strong or extreme evidence for heterogeneity except sustainable consumption (weak evidence against). Lastly, all outcome behaviors had strong to extreme evidence for publication bias except general health (moderate evidence).

Combinations of intervention class and outcome behavior depicted weak to strong evidence against the presence of an effect except for incentives × general health showing weak evidence in favor of an effect. Evidence for the presence of heterogeneity and publication bias was lower, yet still considerable, for intervention-outcome subgroups than for subgroups split by interventions or outcomes only. It follows that, for publication bias correction, analyses of smaller, more homogenous effect size samples can lead to more precise estimates. Bayes factors on evidence for the presence of an effect, heterogeneity, and publication bias, however, indicate the need for careful interpretation of effects of behavior change interventions in the population.

Discussion

Our synthesis of 269 meta-analyses encompassing 838 treatment effect estimates provides an overview of the effectiveness of behavior change interventions. The majority of these meta-analyses of RCTs focused on health behaviors; much less evidence was available for finance and sustainability. Compared to the conventional benchmarks used in behavioral science, the extracted estimates were generally small and exhibited high heterogeneity. This was the case for both pooled estimates and estimates, which were separated according to intervention type and behavioral outcomes. Substantial heterogeneity persisted even after accounting for between- and within-study differences through subgroup analyses. This was not surprising given that unique specifications (e.g., analytical decisions and sampling procedures) were aggregated on the primary study level *and* the meta-analytic level before entering our analyses.

Correcting for publication bias using RoBMA revealed strong evidence for the general effectiveness of behavior change interventions across all extracted effects. Although the corrected mean effect was substantially smaller than both the uncorrected and random-effects estimates, it was still statistically significant at conventional levels, with credible intervals ruling

out zero. Our subgroup analyses only revealed strong evidence for the effectiveness of incentive interventions. The evidence of an effect for specific outcome behaviors was weak in some cases and contradicting in others. The RoBMA results supported findings on heterogeneity from the random-effects models and also strongly suggested publication bias across most of the sampled literatures on behavior change interventions. Whereas our findings from descriptive analyses of extracted effects and naïve random-effects models may therefore support the encouraging tale about behavioral science-based interventions, the results from publication bias correction paint a less optimistic picture.

The encouraging tale emphasizes the positive general effects and high heterogeneity of behavior change interventions, suggesting that, in many cases, these interventions may have significant impacts on individual behaviors across a range of domains, including health, finance, and sustainability. The variation in effect sizes between studies may reflect the potential for large, context-specific successes, hinting at the prospect of well-targeted and tailored interventions delivering substantial benefits. A key task is to better understand the sources of this heterogeneity. It may stem primarily from research design [43]; it may also pertain to intervention, outcomes, populations, and context.

The grim tale emerging from our meta-meta-analysis of behavior change interventions is one of inflated expectations and misleading conclusions. While extracted effects and random-effects models cautiously suggest positive and promising effects of behavior change interventions, effects diminish substantially once publication bias is accounted for. The estimated naïve effect sizes, initially around 0.20 standard deviations, are deflated to approximately 0.06, leaving substantial unexplained heterogeneity in treatment effects across meta-analyses. This residual variation raises questions about the reliability of behavioral interventions, in particular around their consistency and generalizability. The lack of systematic documentation of contextual factors and heterogeneity in meta-analyses further exacerbates this issue, suggesting that what works in one setting may not work in others. This points to three profound challenges within the field. First, there is a need for more rigorous standards that can address and incorporate the heterogeneity of effects, including learning from failed interventions (e.g., review by [44]). Second, it is crucial to consider contexts rather than simply relying on evidence from one specific situation. This point, which has recently been emphasized in fields such as self-control research, highlights the need to consider the interaction between the individual and environment and calls for collective action in examining the limitations of individual behavior change [45]). Third, the empirical record on human behavior and performance that is invoked to derive and justify behavioral interventions is often equally heterogenous [46].

Both narratives emerging from our meta-meta-analysis on behavior change interventions are a call to action: to implement better reporting standards in research on behavioral interventions, to gain a clearer picture of the effects of behavior change interventions in specific contexts, and to carefully interpret the meaning of small statistical effect sizes.

We identified intervention types as one source of systematic variation in the effectiveness of behavior change interventions. We categorized these interventions into three broad classes. The largest and conceptually most heterogeneous class, infopowerment, included interventions ranging from simple information provision to complex individualized counseling. The wide variation in treatment effect estimates for infopowerment interventions likely reflects the diversity of its components. Due to poor documentation and a lack of clear conceptual

definitions, we had to merge interventions that provide information and interventions geared at empowerment into one class, leaving us unable to unpack infopowerment interventions further. The other two classes, choice architecture and incentives, were conceptually more focused and thus included fewer estimates. Indeed, only one of the 838 estimates could be classified as a choice architecture intervention, indicating limited meta-analytic evidence from RCTs of architectural nudges in the time period we analyzed. Incentive interventions demonstrated larger and more homogeneous effects than infopowerment interventions. However, these two classes are likely to differ in their target domains, their required resources, and the sustainability of their effects post-intervention [47].

In our sample of meta-analyses, incentive interventions primarily targeted health behaviors (e.g., substance use, physical activity, general health), whereas infopowerment interventions spanned the domains of health, finance, and sustainability. While incentives were particularly present in health behaviors their acceptability generally varies depending on the behavioral outcome. For instance, people accepted incentives for physical activity, weight loss, and self-management more readily than incentives for breastfeeding, medication adherence, or vaccination [48]. Moreover, while incentives can motivate people to take up a new behavior, it is usually abandoned once the incentive is removed unless the incentive is combined with other psychological means, such as engaging intrinsic motivation, forming habits, or establishing social norms (*47*).

Due to insufficient reporting details in many meta-analyses, we could not formally analyze these moderators. Although our analysis of the evidence suggests a more optimistic view for the role of incentives (strong evidence for an effect and heterogeneity), we therefore caution against definitive judgments about the relative effectiveness of infopowerment versus incentive (and choice architecture) interventions in changing individual behavior.

Similar to intervention classes, behavioral outcomes are an important source of systematic variation in the effectiveness of behavior change interventions. For this meta-meta-analysis, we grouped relatively similar behaviors into distinct domains and then differentiated domains into more specific outcomes that varied in conceptual scope, from specific behavioral outcomes targeting household finance to more complex behavioral outcomes targeting obesity or physical activity (Fig. 2d, Fig. 4). Complex behaviors are more challenging to change than simpler ones. For example, sustaining weight loss requires permanent lifestyle changes and is often more difficult than behaviors such as opening a retirement savings account and setting an automatic monthly deposit.

Additionally, consensually desirable behaviors such as recycling or quitting smoking may be easier to achieve because they are supported through social structures (e.g., norms or guidelines) and legal regulations (e.g., single-use plastic ban, public smoke free zones). In contrast, behaviors such as weight loss or increased physical activity may not be universally desirable or accessible for everyone in any context, and commercial and public environments may even promote counterproductive behaviors (e.g., limited access to healthy foods, pedestrian-unfriendly neighborhoods). The success of behavior change interventions in these cases may therefore depend more heavily on the individual's motivation and resources.

In total, the present meta-meta-analysis adds to the debate on how effective interventions are and highlights the risk of inflated benchmarks. Behavior change interventions may have effect sizes

of 0.05 to 1 SD units, on average, with high variability, depending on the unique specifications of the study. Notwithstanding small effect sizes, behavioral interventions can produce meaningful real-world outcomes (see [32,49–51]). Common examples of highly consequential, yet, small effects include correlations between anti-inflammatory drugs and pain relief (r = .14) and calcium intake and bone mass in premenopausal women (r = .08; (*51*)).

In addition, even small statistical effect sizes can be economically significant. For example, an effect of .05 SD on retirement savings at the extensive margin in the U.S. Survey of Consumer Finance 2022 [53] of individuals younger than 35 years corresponds to an annual increase in savings of about $3,000 (2022 purchasing power parity) relative to mean savings of about $24,000 (2022 purchasing power parity)—that is, an annual increase in savings of about 12 percent. Furthermore, small changes in financial behaviors can accumulate over time. An additional annual savings of $3,000 can translate into approximately $122,000 for individuals who invest over a 30-year window and receive an annual interest rate of 2%. Similarly, small improvements in public health behaviors may lead to significant cost savings at the system level. Classifying treatment effects as "negligible" based solely on the analysis of scale-free statistical effect sizes therefore risks overlooking the fact that an intervention with small statistical effect sizes can still have tangible benefits in the real world.

Methods

Literature search

We conducted systematic literature searches for behavior change interventions in health, finance, and sustainability up to 29 May 2020. We implemented Boolean operators for all target domains in Web of Science and EBSCO (Econlit and PsychInfo) and searched PubMed for health behavior change interventions only (see Table S4). We examined the comprehensiveness of our systematic search through manual searches and identified additional meta-analyses meeting the inclusion criteria. We limited our search to meta-analyses that were available as full-text and published in English or German. RH, TK, and JM conducted the literature searches; JK and one intensively trained research assistant reviewed the full texts of the articles. Their inter-rater reliability was high (Cohen's kappa = .883); any discrepancies were resolved in consensus between JK, TK, and the research assistant.

Inclusion and exclusion criteria

We selected meta-analyses according to the following participant–intervention–comparison–outcome–study (PICOS) criteria: (1) Female and male participants of all ages from the general population. We excluded meta-analyses with clinical populations (e.g., people with HIV) but included high-risk populations and populations with common health conditions (e.g., obesity). We included meta-analyses with clinical and non-clinical populations if effects could be extracted separately for non-clinical populations. (2) Behavior change interventions using infopowerment, incentives, and choice architecture. Infopowerment interventions offered semantic or procedural knowledge, skills, or decision tools, and/or changed the external environment (for details and examples, see Table 1). Incentive interventions changed monetary or non-monetary incentivization. Choice architecture interventions harnessed cognitive or motivational biases without changing them. We excluded effects from non-behavioral interventions (e.g., pharmacological). (3) Passive or active control groups that met the same

participant inclusion criteria as treatment groups. (4) Behavioral outcomes related to health, finance, and sustainability. Examples for eligible health behaviors include objective lifestyle-related outcomes, physical activity, nutrition, and alcohol consumption (for details, see Table 1). Examples for financial behaviors include saving behaviors, business profits, and education investments. Examples for sustainable consumption behaviors include energy usage, towel reuse, and meat consumption. (5) Meta-analyses of random control trials (RCTs) with randomization at individual or cluster-level. We included meta-analyses with primary studies other than RCTs (e.g., quasi-experimental, cross-sectional) if effects could be extracted for RCTs only.

Data extraction

We extracted information for each included meta-analyses and for each included effect size. We did not consult primary studies as this was outside the scope of this project. On the study level, we extracted general information about the meta-analysis (e.g., study title, list of authors, publication year) and included primary studies (e.g., number of RCTs, sample size, setting and cost of intervention, delivery mode and target, publication bias). On the effect-size level, we extracted uncorrected effects with uncertainty measures (i.e., standard deviation, standard error, or confidence intervals), the applied meta-analytic model, and descriptions of the intervention, control group, and outcome.

We classified interventions by re-examining the included meta-analyses. We only classified effect sizes that met three criteria: specific descriptions of applied intervention tools existed for each effect, applied intervention tools were homogenous across primary studies, and interventions could be classified as infopowerment, incentives, or choice architecture (for details, see Supplementary Materials).

If sufficient information was available, we categorized control groups as no treatment, other treatment, standard care, attenuated treatment, or mixed controls. We lastly split health-related outcomes into nine distinct groups, resulting in a total of 11 outcome behaviors: (1) overweight and obesity, (2) substance use, (3) physical activity and sedentary behavior, (4) general health, (5) sexual health, (6) eating and nutrition, (7) medical procedures, (8) feeding practices, (9) oral health, (10) financial behaviors, and (11) sustainable consumption.

Initial data extraction collected 913 treatment effects from 293 meta-analyses. We excluded entries that could not be standardized ($k = 54$) or did not report uncertainty measures ($k = 12$). We transformed all uncertainty measures into standard errors (SEs) and then converted estimates and SEs into standardized mean difference (d) units where necessary (for details, see Supplemental Materials). We coded treatment effects so that a positive estimate indicated the intended behavior change. Finally, we examined the distribution of the standardized estimates and excluded nine estimates that were outliers above the 99th percentile ($d > 1.60$).

In total, these preparatory steps led to the exclusion of 75 effects and 24 meta-analyses. Our final sample therefore consisted of 269 meta-analyses with 838 unique treatment effects. Considering the diversity of included behavior change meta-analyses, we decided to assess the influence of intervention class, outcome behavior, and combinations of both.

Statistical analyses

We conducted statistical analyses using R (v4.3.3; [54]) and JAGS (v4.3.2; [55]) with the packages *tidyverse* [56], *metafor* [57], *robumeta* [58] and *RoBMA* [59]. A reproduction manual including the complete dataset and code is available in the Supplementary Materials.

Random-effects models

The included meta-analyses (and their primary studies) differed considerably across included populations, interventions, control groups, outcome behaviors, and study design. Drawing general conclusions about the effectiveness of behavior change interventions across domains therefore required us to account for apparent heterogeneity. Meta-analyses commonly assume between-study heterogeneity, allowing true effects to vary across studies with the same within-study measurement error. The average effect estimate therefore represents the mean of the distribution of true effects rather than a single true effect:

$$y_{ij} = \beta_0 + \upsilon_j + \epsilon_{ij}, \tag{1}$$

where y_{ij} represents the ith estimated effect for each study j. β_0 marks the mean of the true effects distribution, υ_j is the study-level random effect with $\upsilon_j \sim N(0, \tau^2)$, τ^2 is the between-study variance in true effects, and $\epsilon_{ij} \sim N\left(0, \sigma_{ij}^2\right)$ depicts the residual of the ith estimated effect for each study j.

In addition to between-study heterogeneity, our meta-analysis assumed within-study dependencies because we extracted several treatment effects from the same study where applicable. We implemented two procedures to address multiple, potentially correlated, treatment effects within studies. First, multivariate meta-analysis included random effects for each study and nested random effects for each treatment effect within studies [39]. Multivariate approaches, however, make assumptions (i.e., multivariate normality, known covariance structure) that, if not met, reduce model accuracy.

In contrast to multivariate meta-analysis, [38] suggest robust variance estimation (RVE) that does not require additional model assumptions. In RVE, inverse variance weights adjust the between-study variance τ^2 depending on within-study sampling variances σ_{ij}^2 and the assumed common within-study correlation of treatment effects ρ. We estimated the RVE model with $\rho = .80$ as the default within-study correlation of estimates (see [38]).

Publication bias

Publication bias, or the tendency to preferentially publish statistically significant treatment effects in primary studies, translates into inflated meta-analytic estimates that compromise meaningful inference on true effects and their distribution in the population. Yet only a subset of 168 of 269 included meta-analyses addressed publication bias; those that did employed various approaches with differing robustness. To allow for meaningful inferences, we therefore extracted uncorrected meta-analytic estimates and adjusted them for publication bias using several frequentist and Bayesian procedures.

In the first frequentist method we employed, selection models use step-functions on *p*-value distributions to correct for selective publication for positive outcomes by increasing the weight

of studies in *p*-value intervals with lower publication probability to match the weight of studies with highest publication probabilities [40]. Step functions were computed across three *p*-value intervals for positive effects that were significant ($p_{one-tailed} < .025$), marginally significant ($.025 \leq p_{one-tailed} < .05$), and non-significant ($p_{one-tailed} \geq .05$). If the three-step selection model failed to converge due to an insufficient number of *p*-values within a specified interval, a two-step-function distinguished between significant and non-significant *p*-values only.

The second frequentist method also used step functions for *p*-value intervals but allowed for clustered effect sizes, accounting for within-study dependencies [27]. We allowed the resulting conditional publication probabilities (β_p) to be asymmetric around zero.

The third frequentist method, weighted average of the adequately powered (WAAP; [41]), proposes that studies must be adequately powered to estimate true effects. When assuming conventional levels of statistical significance ($\alpha = .05$) and adequate power ($1 - \beta = .80$), the true effect must be at least 2.8 standard errors away from zero to reject the null (cf. [60], p. 441). This means that the standard error of an estimate must be smaller than the absolute value of the underlying effect divided by 2.8. As the underlying true effect was unknown, we chose the median uncorrected effect ($d = .14$) from our sample as possible effect. The WAAP procedure therefore only considered effect sizes with standard errors smaller than $0.14 \div 2.8 = 0.05$ (at $\alpha = .05$ and $1 - \beta = .80$) as estimates of underlying true effects.

We also applied RoBMA [42], which combines various frequentist methods into a Bayesian model-averaging framework. RoBMA establishes estimates of true effects by weighting included models against their performance. We implemented RoBMA using its default prior settings (i.e., standard normal distribution for treatment effects, inverse gamma distribution with $\alpha = 1$ and $\beta = .15$ for heterogeneity, and PET-PEESE and six step functions for publication bias adjustment; [61]). We report the estimate of the corrected true effect, the heterogeneity estimate, and Bayes factors that quantify the strength of evidence for the presence of an effect (BF_{10}), the presence of heterogeneity (BF^{rf}), and the presence of publication bias (BF_{pb}).

References and Notes

1. International Energy Agence (IEA). Net Zero by 2050 - A Roadmap for the Global Energy Sector. (2021).
2. Willett, W. *et al.* Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems. *The Lancet* **393**, 447–492 (2019).
3. Poore, J. & Nemecek, T. Reducing food's environmental impacts through producers and consumers. *Science* **360**, 987–992 (2018).
4. McGarity, T. O. Administrative Law as Blood Sport: Policy Erosion in a Highly Partisan Age. *Duke Law Journal* **61**, 1671–1762 (2012).
5. Chater, N. & Loewenstein, G. The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behav Brain Sci* 1–60 (2022) doi:10.1017/S0140525X22002023.
6. Hallsworth, M. A manifesto for applying behavioural science. *Nat Hum Behav* **7**, 310–322 (2023).

7. List, J. A., Rodemeier, M., Roy, S. & Sun, G. K. Judging Nudging: Understanding the Welfare Effects of Nudges Versus Taxes. *National Bureau of Economic Research* **No. w31152**, (2023).
8. Charness, G. & Gneezy, U. Incentives to Exercise. *Econometrica* **77**, 909–931 (2009).
9. Gneezy, U., Meier, S. & Rey-Biel, P. When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives* **25**, 191–210 (2011).
10. Benartzi, S. *et al.* Should Governments Invest More in Nudging? *Psychological Science* **28**, 1041–1055 (2017).
11. Thaler, R. H. & Sunstein, C. R. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. (Yale University Press, New Haven, CT, 2008).
12. Thaler, R. H. & Sunstein, C. R. *Nudge: The Final Edition*. (Penguin, London, 2021).
13. Hertwig, R. When to consider boosting: some rules for policy-makers. *Behav. Public Policy* **1**, 143–161 (2017).
14. Hertwig, R. & Grüne-Yanoff, T. Nudging and Boosting: Steering or Empowering Good Decisions. *Perspect Psychol Sci* **12**, 973–986 (2017).
15. Hertwig, R. & Ryall, M. D. Nudge Versus Boost: Agency Dynamics Under Libertarian Paternalism. *The Economic Journal* **130**, 1384–1415 (2020).
16. Herzog, S. M. & Hertwig, R. Boosting: Empowering citizens with behavioral science. *Annual Review of Psychology* (in press).
17. Halpern, J. Y. *Actual Causality*. (The MIT Press, Cambridge, Massachusetts, 2016).
18. Thaler, R. H. & Sunstein, C. R. Libertarian Paternalism. *The American Economic Review* **93**, 175–179 (2003).
19. Mertens, S., Herberz, M., Hahnel, U. J. J. & Brosch, T. The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2107346118 (2022).
20. Bakdash, J. Z. & Marusich, L. R. Left-truncated effects and overestimated meta-analytic means. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2203616119 (2022).
21. Maier, M. *et al.* No evidence for nudging after adjusting for publication bias. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2200300119 (2022).
22. Szaszi, B. *et al.* No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences of the United States of America* **119**, e2200732119 (2022).
23. DellaVigna, S. & Linos, E. RCTs to Scale: Comprehensive Evidence From Two Nudge Units. *ECTA* **90**, 81–116 (2022).

24. Al-Ubaydli, O., Lai, C. Y. & List, J. A. A Simple Rational Expectations Model of the Voltage Effect. Preprint at https://ideas.repec.org/p/nbr/nberwo/30850.html (2023).

25. Allcott, H. Site Selection Bias in Program Evaluation*. *The Quarterly Journal of Economics* **130**, 1117–1165 (2015).

26. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat Hum Behav* **5**, 980–989 (2021).

27. Andrews, I. & Kasy, M. Identification of and Correction for Publication Bias. *American Economic Review* **109**, 2766–2794 (2019).

28. Brodeur, A., Cook, N. & Heyes, A. Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* **110**, 3634–3660 (2020).

29. Elliott, G., Kudrin, N. & Wüthrich, K. Detecting *p* -Hacking. *ECTA* **90**, 887–906 (2022).

30. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: Unlocking the file drawer. *Science* **345**, 1502–1505 (2014).

31. Friese, M. & Frankenbach, J. p-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates. *Psychological Methods* **25**, 456–471 (2020).

32. Götz, F. M., Gosling, S. D. & Rentfrow, P. J. Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science* **17**, 205–215 (2022).

33. Ioannidis, J. P. A., Stanley, T. D. & Doucouliagos, H. The Power of Bias in Economics Research. *The Economic Journal* **127**, F236–F265 (2017).

34. Stanley, T. D., Carter, E. C. & Doucouliagos, H. What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin* **144**, 1325–1346 (2018).

35. Vivalt, E. Specification Searching and Significance Inflation Across Time, Methods and Disciplines. *Oxf Bull Econ Stat* **81**, 797–816 (2019).

36. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological Bulletin* **86**, 638–641 (1979).

37. DeVito, N. J. & Goldacre, B. Catalogue of bias: publication bias. *BMJ evidence-based medicine* **24**, 53–54 (2019).

38. Tanner-Smith, E. E. & Tipton, E. Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS: Robust variance estimation. *Res. Syn. Meth.* **5**, 13–30 (2014).

39. Jackson, D., Riley, R. & White, I. R. Multivariate meta-analysis: Potential and promise. *Statist. Med.* **30**, 2481–2498 (2011).

40. Vevea, J. L. & Woods, C. M. Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions. *Psychological Methods* **10**, 428–443 (2005).

41. Stanley, T. D., Doucouliagos, H. & Ioannidis, J. P. A. Finding the power to reduce publication bias. *Statistics in Medicine* **36**, 1580–1598 (2017).

42. Bartoš, F., Maier, M., Wagenmakers, E., Doucouliagos, H. & Stanley, T. D. Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods* **14**, 99–116 (2021).

43. Holzmeister, F. *et al.* Heterogeneity in effect size estimates. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2403490121 (2024).

44. Osman, M. *et al.* Learning from Behavioural Changes That Fail. *Trends in Cognitive Sciences* **24**, 969–980 (2020).

45. Hofmann, W. Going beyond the individual level in self-control research. *Nat Rev Psychol* **3**, 56–66 (2023).

46. Lejarraga, T. & Hertwig, R. How experimental methods shaped views on human competence and rationality. *Psychological Bulletin* **147**, 535–564 (2021).

47. Winkler-Schor, S. & Brauer, M. What Happens When Payments End? Fostering Long-Term Behavior Change With Financial Incentives. *Perspect Psychol Sci* 17456916241247152 (2024) doi:10.1177/17456916241247152.

48. Hoskins, K., Ulrich, C. M., Shinnick, J. & Buttenheim, A. M. Acceptability of financial incentives for health-related behavior change: An updated systematic review. *Preventive Medicine* **126**, 105762 (2019).

49. Evans, D. K. & Yuan, F. How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis* **44**, 532–540 (2022).

50. Kaiser, T., Lusardi, A., Menkhoff, L. & Urban, C. Financial education affects financial knowledge and downstream behaviors. *Journal of Financial Economics* **145**, 255–272 (2022).

51. Funder, D. C. & Ozer, D. J. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science* **2**, 156–168 (2019).

52. Meyer, G. J. *et al.* Psychological Testing and Psychological Assessment. *American Psychologist* **56**, 128–165 (2001).

53. Board of Governors of the Federal Reserve System. *2022 Survey of Consumer Finances*. https://www.federalreserve.gov/econres/scfindex.htm (2023).

54. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria* (2023) doi:https://www.R-project.org/.

55. Depaoli, S., Clifton, J. P. & Cobb, P. R. Just Another Gibbs Sampler (JAGS): Flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics* **41**, 628–649 (2016).

56. Wickham, H. *et al.* Welcome to the Tidyverse. *JOSS* **4**, 1686 (2019).

57. Viechtbauer, W. Conducting Meta-Analyses in *R* with the **metafor** Package. *J. Stat. Soft.* **36**, (2010).

58. Fisher, Z., Tipton, E. & Zhipeng, H. robumeta: Robust Variance Meta-Regression. 2.1 https://doi.org/10.32614/CRAN.package.robumeta (2014).

59. Bartoš, F. & Maier, M. RoBMA: An R Package for Robust Bayesian Meta-Analyses. (2020).

60. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (Cambridge University Press, Cambridge, UK, 2006).

61. Maier, M., Bartoš, F. & Wagenmakers, E.-J. Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods* **28**, 107–122 (2023).

Acknowledgments: We thank René Buschong, Mattea Dallacker, Ana Sofia Morais, Leonie Strickling, Yannic Sander, and Florian Wittemann for helpful comments and excellent research assistance. We also thank Frank Renkewitz for providing constructive feedback on the manuscript. We are grateful to Deb Ain for her meticulous editing of the manuscript.

Funding:

UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1 (JK)

Author contributions:

Conceptualization: TK, JM, RH

Methodology: TK, JK, JM, RH

Investigation: TK, JK, JM, RH

Formal analysis: JK

Visualization: JK

Writing – original draft: TK, JK, JM, RH

Writing – review & editing: TK, JK, JM, RH

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data, code, and materials used in the analysis are available at https://osf.io/7aq56/.

Supplementary Materials

Intervention classification procedure

Effect size conversion

Reproduction manual

Fig. S1

Tables S1 to S5

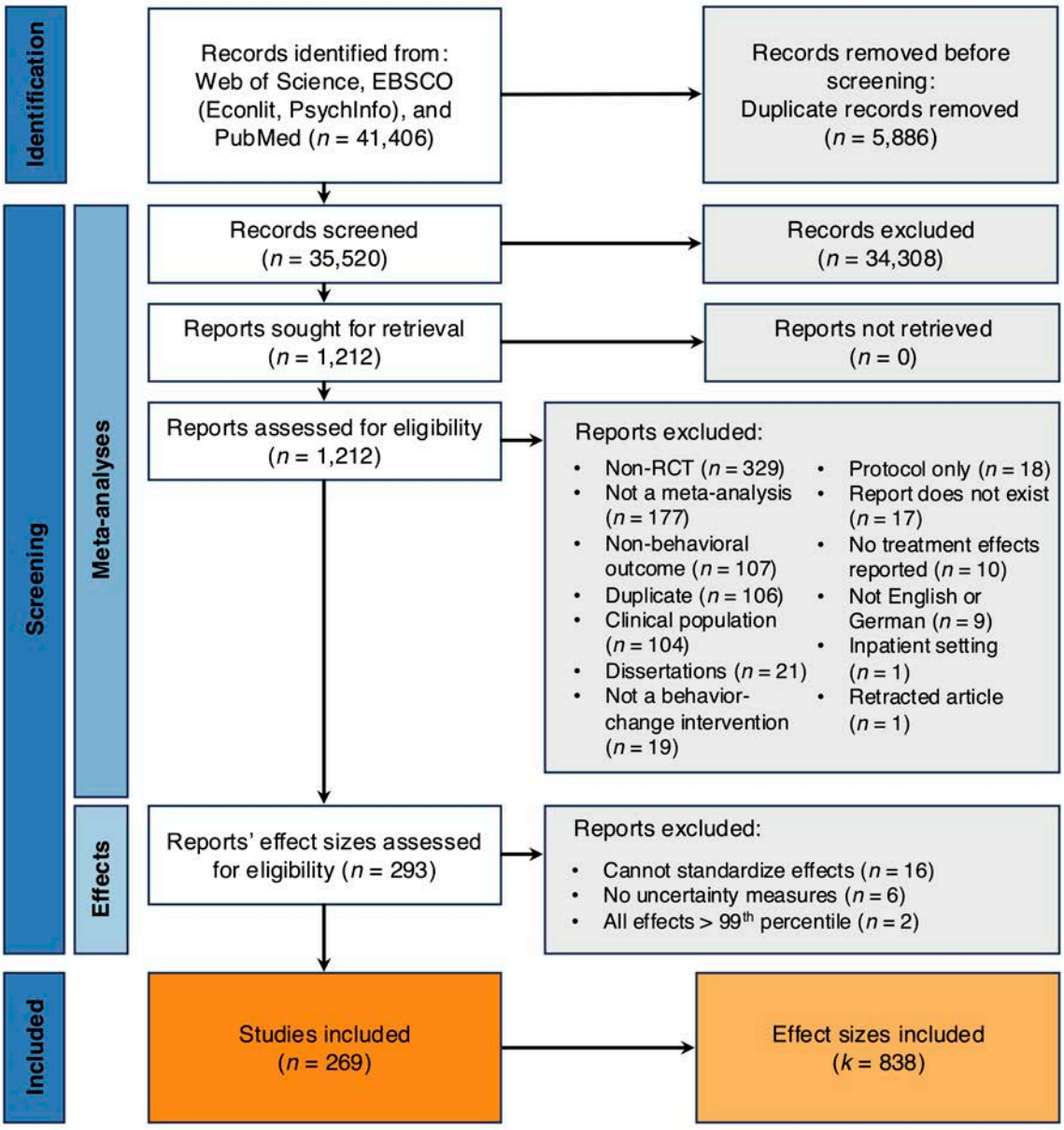

Fig. 1. Preferred reporting items for systematic reviews and meta-analyses flowchart for study inclusion. The initial search identified 41,406 records. After duplicate removal, 35,520 records entered the screening process, which consisted of screening title and abstract, retrieving the full text, assessing the eligibility of reports, and assessing the eligibility of reported effect sizes. The final sample included 269 meta-analyses and 838 treatment effects that met the inclusion criteria.

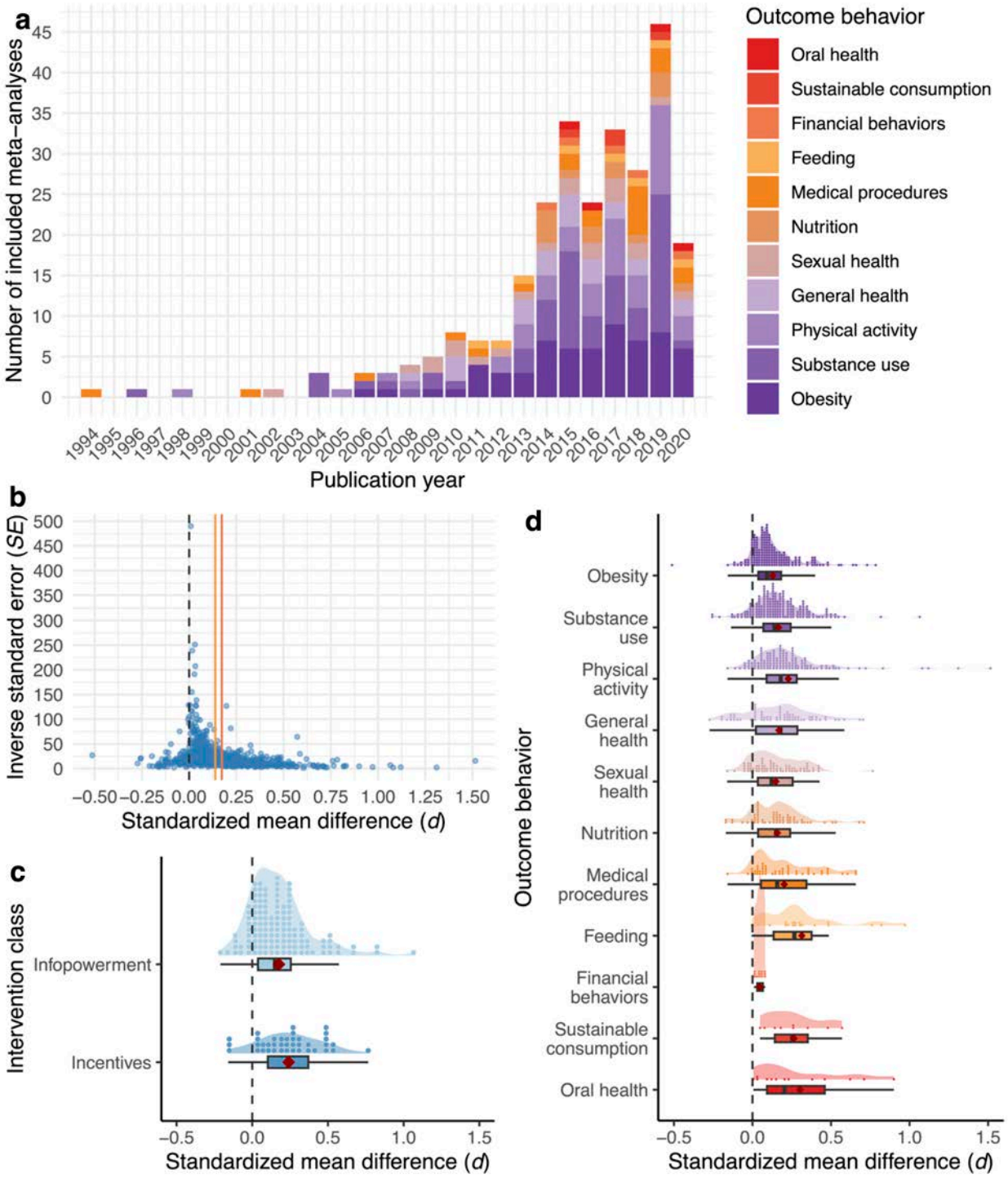

Fig. 2. Overview of included meta-analyses and effect sizes. (a) Included meta-analyses (n = 269) split into 11 distinct outcome behaviors. Over the analyzed time period, the number of included meta-analyses grew exponentially, as did the diversity of assessed outcomes. **(b)** Precision of uncorrected effects (n = 838). Smaller inverse standard errors suggest lower precision; larger standardized mean differences suggest increased intended behaviors. The mean uncorrected effect was .173 (red intercept); the median was .138 (orange intercept). **(c, d)** Distribution of uncorrected meta-analytic effects split by (c) intervention class (n = 2) and (d)

outcome behavior (n = 11). Dot plots show absolute distributions; density plots show relative distributions; boxplots show central tendency measures: interquartile range, median, and superimposed means (red diamonds).

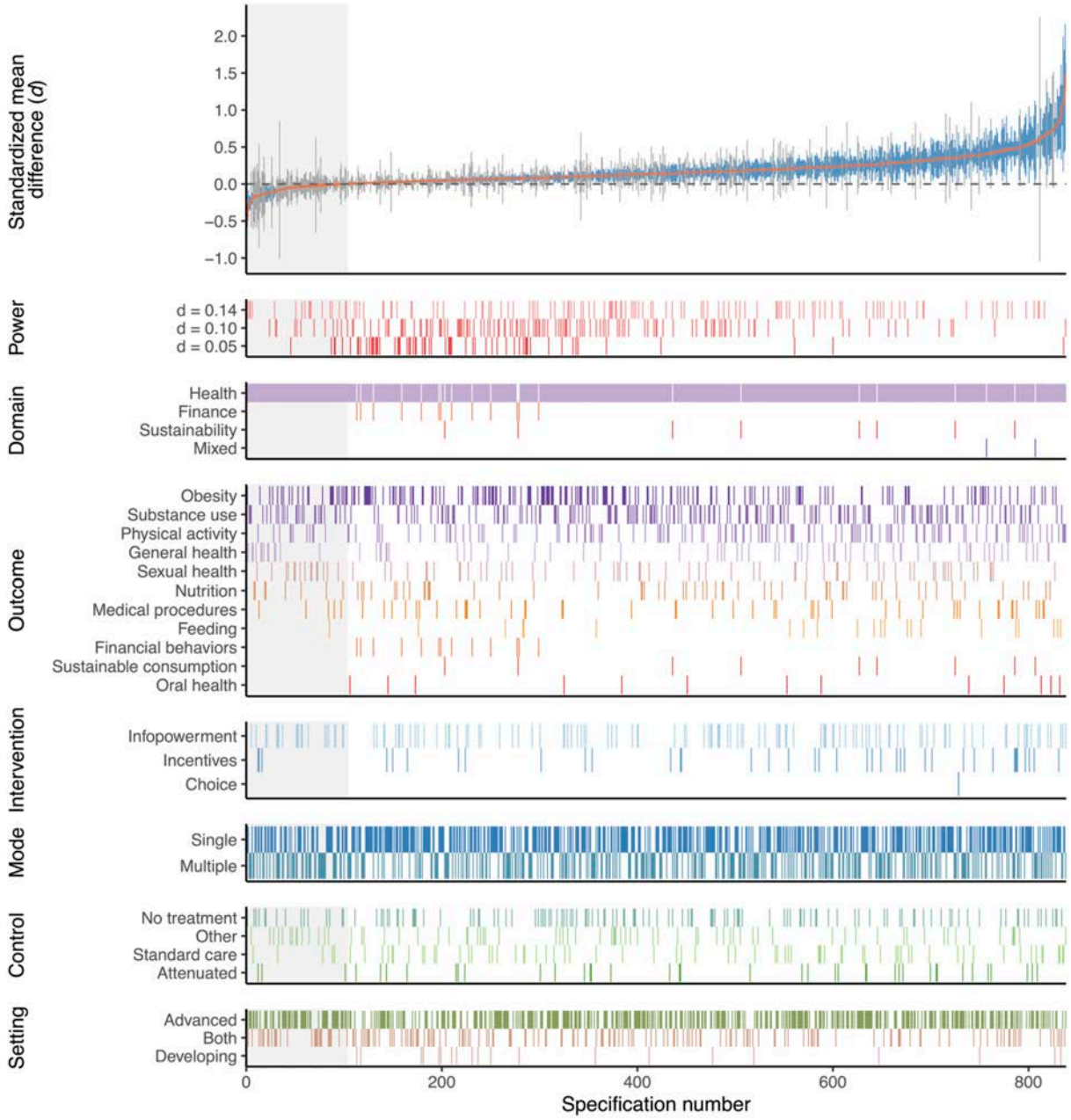

Fig. 3. Specification curve of meta-analytic model. Each line represents a specific effect size (and 95% confidence intervals), with details shown as tiles directly below the lines: power of the effect size (with d = .14 representing the median effect in our sample), research domain, outcome behavior, intervention class, one or multiple modes of intervention delivery (e.g., in person, online, and/or in writing), control group, and economic setting. The lines illustrate that, for instance, studies with lower statistical power were less likely to reach statistical significance (i.e., most effects with sufficient power at d = .05 are non-significant, marked as gray lines; blue lines denote significant results). Light gray areas indicate negative effect sizes. The specification curve simultaneously highlights the impact of analytic decisions and missing reported information on variables of interest.

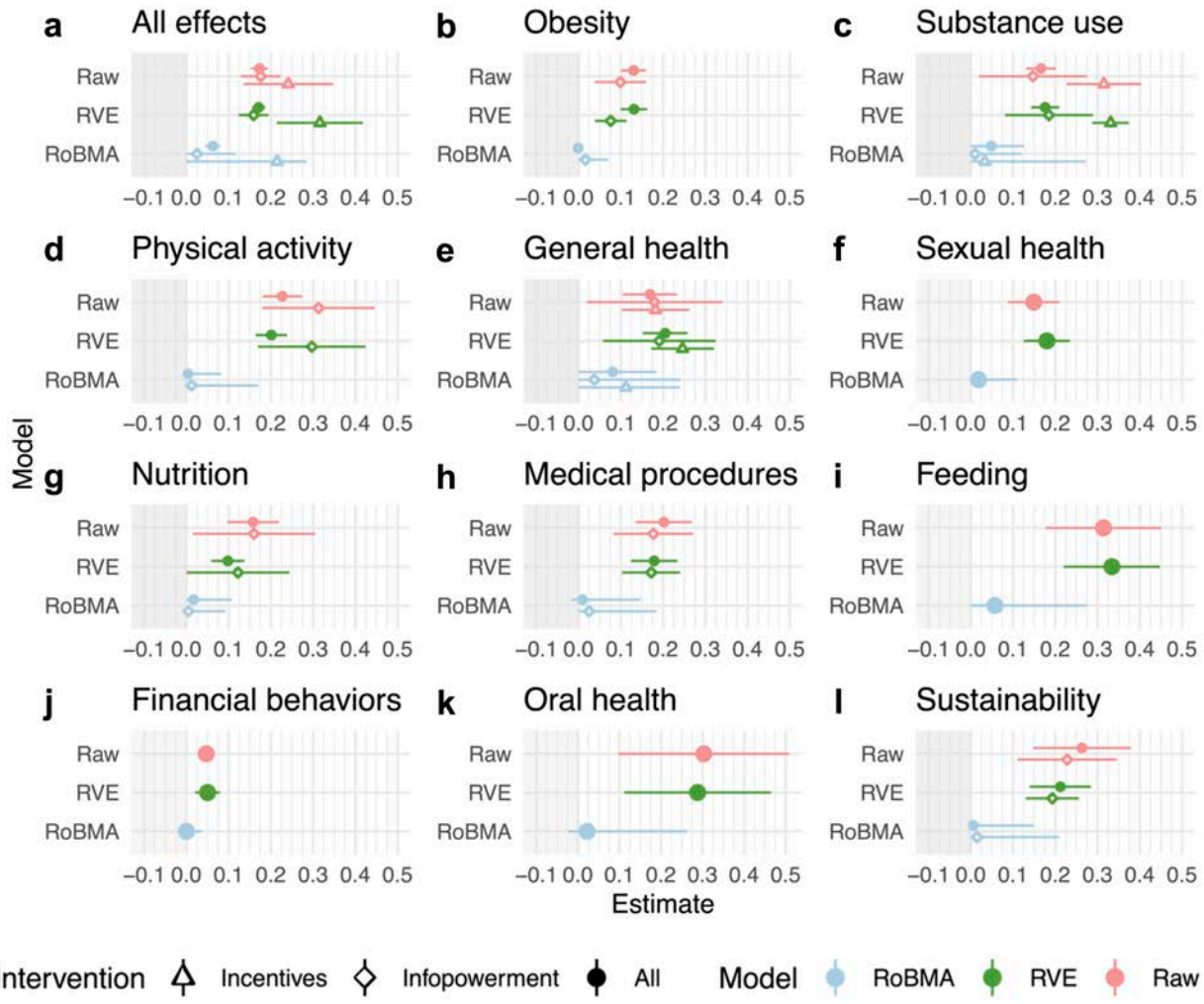

Fig. 4. Assessment of meta-analytic models and publication bias. Mean effects with 95% credible intervals were computed using raw effects, robust variance estimation (RVE), and robust Bayesian meta-analysis (RoBMA). Filled dots depict effects from all included interventions; diamonds depict effects from infopowerment interventions; triangles depict incentive interventions. Light gray areas indicate negative estimates. **(a)** All included effects (n = 838), infopowerment interventions (n = 127), incentive interventions (n = 39). **(b–l)** Effects split by outcome behavior (n = 11) are presented in descending order of included effects. Intervention–outcome category estimates are available for combinations with > 5 effects that converged.

Term	Definition	Examples
Panel A: Intervention class (*n* = 3)		
Infopowerment	Changes in semantic knowledge ("knowing that") or procedural knowledge ("knowing how"), skills (e.g., Bayesian reasoning, goal implementation, self-control strategies), decision tools (heuristics, routines, decision trees), personal commitment (adherence, goal setting), external environment (information representation or physical environment), or any combinations thereof	Decision aids, counseling, feedback, motivational interviewing, physician advice
Incentives	Changes in incentivization (e.g., reward, punishment) combined with procedural information on how to access the incentive	Household cash transfers, free influenza vaccine, financial reward for smoking abstinence
Choice architecture	Changes in the choice architecture that harness cognitive or motivational biases without correcting them. A nudge is any aspect of the choice architecture that alters behavior in a predictable way without forbidding any options or significantly changing their economic incentives	Default rules, order of items on a menu/website, cafeteria design
Panel B: Outcome behavior (*n* = 11)		
Obesity	Changes in body weight and BMI (or zBMI for children), weight-related health markers	Weight loss/gain, percentage of body fat, blood pressure, waist circumference
Substance use	Preventing, reducing, and/or quitting substance use, including tobacco, alcohol, and illicit drugs	Smoking cessation and/or abstinence, drinks per week, binge drinking, cannabis use
Physical activity	Increase in physical activity and/or decrease in sedentary behaviors, activity-related health markers	Moderate to vigorous physical activity, daily steps, cardiorespiratory fitness, skeletal muscle mass, bone mineral density, reduction in sedentary behavior
General health	Changes in general health-related behaviors, combination of multiple related health behaviors (i.e., lifestyle outcomes)	Hygiene practices, sleep time, sun protection, early neonatal mortality, injury prevention
Sexual health	Reduction of sexual risk behaviors, rates of sexually transmitted infections (STIs), human immunodeficiency virus (HIV) risk, unintended pregnancies	Sexual activity, sexual risk behaviors, use of condoms and/or birth control
Nutrition	Intake of healthy or unhealthy foods, levels of glucose/triglycerides/cholesterol, energy consumed, at-risk behaviors for eating disorders	Fruit and vegetable intake, sweetened beverage intake, emotional eating

Medical procedures	Vaccination uptake/completion, medical appointment attendance, uptake of medical screenings	Influenza vaccination uptake, cancer screening attendance, healthcare appointment attendance
Feeding	Breastfeeding and feeding practices for children	Duration of exclusive breastfeeding, age at introduction of complementary foods, feeding frequency
Financial behaviors	Changes in financial behaviors	Saving behaviors, total savings, business profits, education investment
Sustainable consumption	Changes in environmentally sustainable behaviors	Energy usage, increased recycling, towel reuse, meat consumption
Oral health	Reduction in caries, improved markers for dental hygiene	Dental caries, gingival index, plaque index

Table 1. Definitions and examples of included intervention classes and outcome behaviors. Intervention classes were determined using a three-step procedure that only considered interventions with sufficiently specific descriptions and homogenous intervention tools across primary studies that could be categorized into one of the three intervention classes (for details, see Supplementary Materials). Overlap and/or ambiguities between outcome behaviors were resolved through selecting the outcome behavior that most closely represented the research aims of the corresponding meta-analysis.

Supplementary Materials for

A meta-meta-analysis of behavior change interventions: Two tales of behavior change

Tim Kaiser, Juliane Kloidt, Jutta Mata, Ralph Hertwig

The PDF file includes:

Intervention classification procedure
Effect size conversion
Reproduction manual
Fig. S1
Tables S1 to S5

Intervention classification procedure

Treatment effect estimates included in this study were extracted from meta-analyses of diverse interventions. To achieve more precise intervention classes, we categorized the interventions using three increasingly restrictive classification steps: (1) specific descriptions of applied intervention tools were available for each effect, (2) applied intervention tools were homogenous across primary studies, and (3) interventions could be classified as infopowerment, incentives, or choice architecture. From our total sample of 838 treatment effect estimates, we excluded 486 estimates in the first step, 141 estimates in the second step, and 44 estimates in the final step. The final sample therefore comprised 167 treatment effect estimates.

One researcher (JK) completed the screening procedure (steps 1 and 2) and all researchers classified interventions (step 3). To test step 3, all researchers completed a subsample of 20 treatment effects together. Of the remaining effect sizes (k = 191), 167 treatment effects were coded by two researchers and 24 treatment effects were coded by three researchers. This method resulted in 73% agreement across raters. JK classified all effect sizes (k = 191); TK, JM, and RH classified 54 to 64 effect sizes each. Divergent classifications were discussed and resolved by all researchers.

Step 1: Specificity

Treatment effects passed step 1 if it was possible to extract a specific description of the applied intervention tools underlying the extracted treatment effect estimate. Descriptions were sufficiently specific if they reported how an intervention induced behavior change. We excluded estimates that provided only generic labels (e.g., "lifestyle intervention") or described the intervention context (e.g., "school-based intervention") but did not explain the intervention procedure. Estimates with descriptions that may refer to complex procedures (e.g., "weight loss program," "culturally appropriate behavioral interventions") were retained for subsequent steps. Out of 838 effects, we excluded 486 effect sizes with insufficiently specific descriptions about the applied intervention tools and entered 352 effect sizes into step 2.

Step 2: Homogeneity

To pass step 2, treatment effect estimates had to be based on the same intervention tools across primary studies. In other words, estimates must be derived from the same single intervention, or the same combination of interventions. We included combinations of various intervention tools to reflect current practices. Treatment effect estimates were excluded if meta-analyses did not provide information about individual intervention components. We did not consult primary studies as this exceeded the scope of the project. Out of 352 effects, we excluded 141 effect sizes with heterogenous intervention components. We thus entered 211 effect sizes into the third and final step of the refinement procedure.

Step 3: Classification

In step 3, we categorized treatment effect estimates into one of three categories: infopowerment, incentives, or choice architecture (see Table 1). We therefore excluded estimates that were derived from interventions with components reflecting different categories (e.g., infopowerment and incentives). We could not determine more specific intervention classes (e.g., differentiating between information provision and competence development) because many meta-analyses reported only sparse details about included interventions.

Classification ambiguities were resolved through discussions between raters. For example, we decided to categorize interventions targeting individuals' commitment to behavior change as infopowerment interventions rather than choice architecture interventions because changes in decision making were consciously driven by the individual rather than the environment. In another instance, we excluded exercise interventions that lacked any psychological component but instead treated exercise as medicine. Out of 211 effects, we classified the interventions for 127 effect sizes as infopowerment, 39 as incentives, and one effect size as choice architecture. We thus excluded 44 effect sizes at the final step.

Effect size conversion

Included effect sizes were presented in various metrics that could be continuous, correlational, or binary. Treatment effects were therefore converted into standardized mean differences prior to data analysis.

Converting continuous effect sizes into *d*

Converting from MD to d

Some studies only reported point estimates instead of effect sizes. We converted the difference between point estimates—that is, the mean difference (MD)—into the standardized mean difference (d), following [1]:

$$d = \frac{MD}{S_{pooled}}, \tag{1}$$

where S_{pooled} is the pooled standard deviation for the treatment and control groups combined. When S_{pooled} was not directly reported, it was calculated using

$$S_{pooled} = \sqrt{n \times SE_{MD}}\,, \tag{2}$$

where n is the sample size and SE_{MD} is the reported standard error of the mean difference. We converted the standard error of the mean difference (MD) into the standardized mean difference following [2]:

$$SE_d = \sqrt{\frac{n}{n^2} + \frac{d^2}{2n}}\,, \tag{3}$$

where n is the sample size and d is the effect as standardized mean difference. The same procedure was applied to effect sizes and standard errors for weighted mean differences and average treatment effects (ATEs).

Converting from g to d

We converted effect sizes from a Hedge's g (g) into a standardized mean difference (d), using

$$d = \frac{1}{J(df)} \times g\,, \tag{4}$$

where J is the sample correction factor. As argued elsewhere [3], the fraction on the left is approximately 1 if sample sizes are large enough. We therefore approximated

$$d \approx g\,. \tag{5}$$

Converting correlational effect sizes into *d*

Converting from r to d

We converted effect sizes from a correlation (r) into the standardized mean difference (d), following [1]:

$$d = \frac{2r}{\sqrt{1-r^2}}\,.$$

(6)

We converted the standard error of the correlation (r) into the standardized mean difference following [4]:

$$SE_d = \sqrt{\frac{4SE_r^2}{(1-r^2)^3}}\,.$$

(7)

Converting binary effect sizes into *d*

Converting from the log odds ratio (OR) to d

We converted effect sizes from a log odds ratio ($LogOddsRatio$) into the standardized mean difference (d), following [1]:

$$d = LogOddsRatio \times \frac{\sqrt{3}}{\pi},$$

(8)

where π is the mathematical constant (approximately 3.14159). Following [5], we applied the same formula for transforming log odds ratio standard errors into standardized mean difference standard errors.

Converting from the risk ratio (RR) to d

We first transformed effect sizes from a risk ratio ($RiskRatio$) into a log odds ratio ($LogOddsRatio$), then converted them into the standardized mean difference (d):

$$LogOddsRatio = \frac{1-p_{treatment}}{1-p_{control}} \times RiskRatio\,.$$

(9)

As argued elsewhere [3], the fraction on the left is approximately 1 if the probabilities of behavior change are small, or if group differences are small. We therefore approximated

$$LogOddsRatio \approx RiskRatio$$

(10)

and applied Equation (9) to transform the risk ratio ($RiskRatio$) into the standardized mean difference (d). The same procedure was applied to effect sizes, referred to as rate ratios.

Reproduction manual

This manual introduces the OSF project containing the data and code to reproduce this meta-meta-analysis (https://osf.io/7aq56/). The OSF project includes a *read-me* file, three folders (*data*, *RoBMA*, *02b-meta-meta-gmm*), and nine Rmd-files (*01-meta-meta-descriptives*, *02-meta-meta-models*, *02a-meta-meta-robma*, *03-meta-meta-fig2*, *04-meta-meta-fig3*, *05-meta-meta-fig4*, *06-meta-meta-tabSM1*, *07-meta-meta-tabSM2*, *08-meta-meta-tabSM3*).

The *data* folder includes the complete dataset (meta-meta-data.csv), the output from the clustered selection models (meta-meta-gmm-outputs.csv), and the output from all random-effects models and publication bias correction methods used (meta-meta-effects-models.csv). The *RoBMA* folder includes model outputs and models summaries from all robust Bayesian meta-analyses (RoBMA; [6]) that were conducted. As RoBMA requires the software JAGS and increased computational power, we saved the outputs to be used in further analyses. The *02b-meta-meta-gmm* folder includes adjusted code from an R-project, developed and published elsewhere [7], to run clustered selection models. As the pre-existing code restricted outputs to tex-files, we summarized them into a csv-file (meta-meta-gmm-outputs.csv) to be used for further analyses (see *data* folder).

The flowchart in fig. S1 illustrates the dependencies between the complete dataset and the coding scripts to reproduce all performed analyses. Detailed information for each analysis script, including a description, required R-packages, input files, and output files, are presented in table S5.

The complete dataset meta-meta-data.csv is required to run the code in *01-meta-meta-descriptives*, *02a-meta-meta-robma*, *02b-meta-meta-gmm*, *03-meta-meta-fig2*, *04-meta-meta-fig3*, and *06-meta-meta-tabSM1*. Running the code in *02-meta-meta-models* requires the datasets meta-meta-data.csv and meta-meta-gmm-outputs.csv (see *data* folder or *02b-meta-meta-gmm* folder) as well as the RoBMA model summaries (see *RoBMA* folder or *02a-meta-meta-robma*). The combined model output, meta-meta-effects-models.csv (see, *02-meta-meta-models* or *data* folder), is required for running the code in *05-meta-meta-fig4*, *07-meta-meta-tabSM2*, and *08-meta-meta-tabSM3.*

References and Notes

1. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. Converting among effect sizes. in *Introduction to meta-analysis* (John Wiley & Sons, Ltd, 2009).
2. Hedges, L. V. & Olkin, I. *Statistical Methods for Meta-Analysis*. (Academic Press, Cambridge, Massachusetts, 2014).
3. Fusar-Poli, P. & Radua, J. Ten simple rules for conducting umbrella reviews. *Evid Based Mental Health* **21**, 95–100 (2018).
4. Lipsey, M. W. & Wilson, D. B. *Practical Meta-Analysis*. (Sage Publications, Inc, London, 2001).
5. Chinn, S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statist. Med.* **19**, 3127–3131 (2000).
6. Bartoš, F., Maier, M., Quintana, D. S. & Wagenmakers, E.-J. Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis. *Advances in Methods and Practices in Psychological Science* **5**, 251524592211092 (2022).
7. Andrews, I. & Kasy, M. Identification of and Correction for Publication Bias. *American Economic Review* **109**, 2766–2794 (2019).
8. Jackson, D., Riley, R. & White, I. R. Multivariate meta-analysis: Potential and promise. *Statist. Med.* **30**, 2481–2498 (2011).
9. Tanner-Smith, E. E. & Tipton, E. Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS: Robust variance estimation. *Res. Syn. Meth.* **5**, 13–30 (2014).
10. Vevea, J. L. & Woods, C. M. Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions. *Psychological Methods* **10**, 428–443 (2005).
11. Stanley, T. D., Doucouliagos, H. & Ioannidis, J. P. A. Finding the power to reduce publication bias. *Statistics in Medicine* **36**, 1580–1598 (2017).
12. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (Cambridge University Press, Cambridge, UK, 2006).
13. Bartoš, F., Maier, M., Wagenmakers, E., Doucouliagos, H. & Stanley, T. D. Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods* **14**, 99–116 (2021).

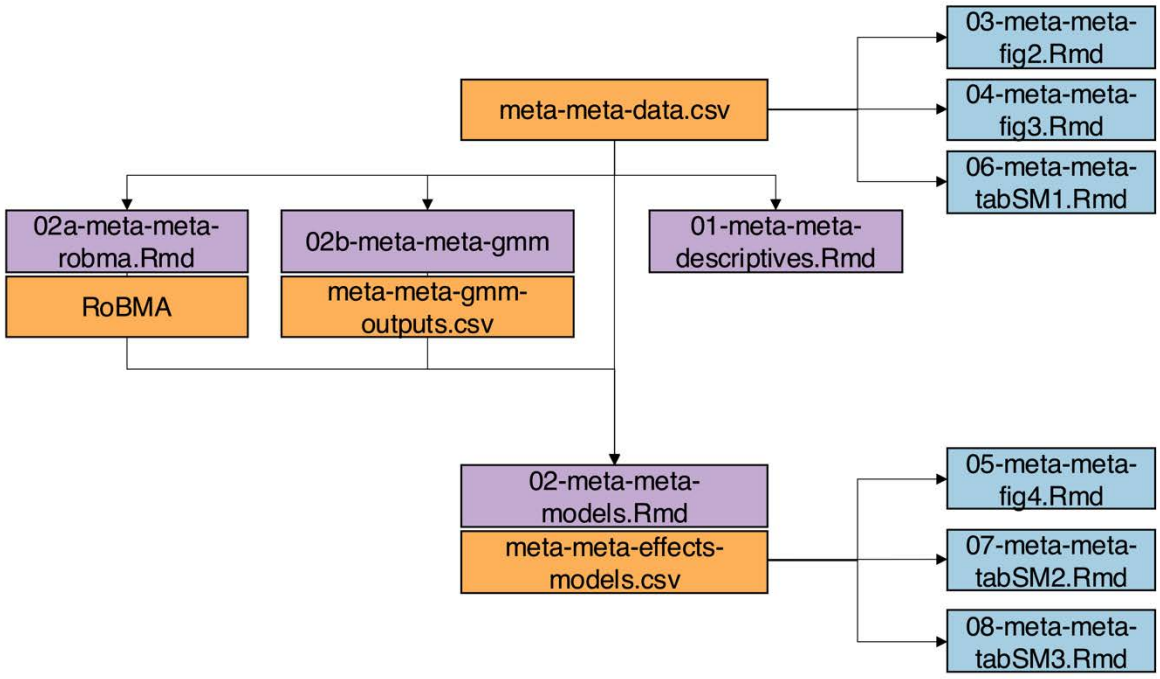

Fig. S1. Flowchart for performed analyses. Datasets and folders are in orange boxes, analysis files in purple boxes, and code to recreate reported tables and figures in blue boxes. Arrows specify dependencies to perform analyses.

5

	M	*SD*	P1	P10	P20	P30	P40	P50	P60	P70	P80	P90	P99	*k*	*n*
All effects	.173	.195	−.167	−.006	.034	.067	.096	.138	.178	.226	.293	.400	.820	838	269
Panel A: Intervention class (*n* = 2)															
Infopowerment	.176	.204	−.153	−.028	.028	.057	.096	.150	.186	.230	.280	.460	.827	127	51
Incentives	.241	.205	−.159	.026	.072	.144	.194	.235	.277	.321	.437	.484	.683	39	13
Panel B: Outcome behavior (*n* = 11)															
Obesity	.130	.158	−.123	.000	.027	.057	.077	.092	.116	.153	.210	.349	.664	195	75
Substance use	.166	.170	−.214	−.011	.046	.081	.119	.140	.178	.220	.283	.357	.607	192	73
Physical activity	.227	.246	−.100	.014	.063	.105	.141	.180	.210	.258	.330	.470	1.240	136	60
General health	.169	.213	−.220	−.123	.020	.058	.110	.178	.220	.256	.329	.430	.689	71	30
Sexual health	.149	.161	−.127	−.026	.003	.054	.074	.122	.160	.231	.290	.350	.519	74	26
Nutrition	.158	.177	−.170	−.005	.028	.040	.126	.150	.172	.230	.262	.346	.691	63	27
Medical procedures	.202	.197	−.095	.010	.037	.058	.086	.157	.207	.277	.367	.475	.658	50	23
Feeding	.314	.257	.003	.071	.084	.211	.235	.267	.289	.312	.463	.733	.940	22	9
Financial behaviors	.047	.025	.009	.012	.023	.035	.043	.045	.053	.062	.073	.077	.081	13	5
Oral health	.303	.290	.008	.026	.059	.109	.144	.204	.260	.412	.557	.692	.877	13	4
Sustainable consumption	.263	.177	.050	.071	.116	.157	.195	.253	.263	.318	.403	.498	.563	9	4
Panel C: Intervention–outcome combination (*n* = 14)															
Infopowerment × substance use	.147	.242	−.178	−.076	−.022	.031	.050	.108	.136	.189	.254	.345	.973	38	12
Infopowerment × obesity	.099	.137	−.084	−.029	.013	.040	.065	.093	.099	.123	.158	.178	.499	19	9
Infopowerment × physical activity	.313	.199	.109	.157	.173	.191	.208	.244	.286	.319	.442	.575	.804	18	8
Infopowerment × medical procedures	.177	.196	−.005	.013	.054	.058	.067	.091	.155	.203	.263	.464	.644	17	11
Infopowerment × nutrition	.160	.189	−.114	−.024	−.002	.025	.059	.165	.218	.242	.280	.415	.503	14	5
Incentives × general health	.182	.152	−.123	.031	.056	.096	.156	.208	.252	.268	.285	.354	.408	14	2
Incentives × substance use	.315	.168	.039	.102	.192	.224	.264	.299	.323	.475	.488	.516	.543	14	6
Infopowerment × general health	.180	.200	−.151	−.053	.050	.140	.194	.230	.242	.258	.283	.359	.450	7	5
Infopowerment × sustainable consumption	.228	.147	.053	.095	.142	.161	.180	.216	.253	.260	.266	.373	.469	6	3
Infopowerment × medical procedures	.235	.287	−.152	−.085	−.010	.094	.227	.360	.406	.452	.474	.474	.474	5	5
Infopowerment × financial behaviors	.051	.032	.018	.022	.025	.029	.039	.053	.068	.077	.079	.080	.081	4	4
Incentives × physical activity	.070	.157	−.151	−.080	.000	.079	.112	.123	.134	.146	.161	.176	.190	4	2
Infopowerment × oral health	.305	.106	.232	.245	.260	.275	.290	.305	.320	.335	.350	.365	.378	2	1
Infopowerment × sexual health	.300	.087	.239	.250	.263	.275	.287	.300	.312	.324	.337	.349	.360	2	1

Table S1. Empirical distributions of raw treatment effect estimates. Distributions of raw treatment effect estimates split by intervention class, outcome behavior, and combination of intervention class and outcome behavior. *M* = mean; *SD* = standard deviation; *PX* = percentile of distribution; *k* = number of estimates; *n* = number of meta-analyses.

	n	*k*	Raw estimates	Multivariate random-effects		Robust variance estimation	
			M (SE)	*M (SE)*	I^2	*M (SE)*	I^2
All effects	269	838	.173 (.010)	.167 (.008)	91.47	.171 (.008)	90.95
Panel A: Intervention class (*n* = 2)							
Infopowerment	51	127	.176 (.024)	.164 (.019)	94.31	.159 (.018)	93.97
Incentives	13	39	.241 (.054)	.295 (.047)	80.13	.316 (.052)	85.80
Panel B: Outcome behavior (*n* = 11)							
Obesity	75	195	.130 (.015)	.132 (.015)	89.00	.131 (.016)	89.18
Substance use	73	192	.166 (.018)	.167 (.016)	84.95	.176 (.017)	85.87
Physical activity	60	136	.227 (.024)	.201 (.019)	87.78	.201 (.019)	84.21
General health	30	71	.169 (.033)	.191 (.027)	91.79	.205 (.027)	82.85
Sexual health	26	74	.149 (.031)	.156 (.026)	79.57	.180 (.028)	74.56
Nutrition	27	63	.158 (.031)	.128 (.020)	83.18	.098 (.020)	79.48
Medical procedures	23	50	.202 (.034)	.183 (.030)	97.30	.179 (.028)	95.27
Feeding	9	22	.314 (.070)	.276 (.049)	94.00	.334 (.058)	90.08
Financial behaviors	5	13	.047 (.007)	.044 (.008)	82.24	.050 (.015)	74.63
Oral health	4	13	.303 (.105)	.274 (.094)	92.77	.288 (.090)	95.89
Sustainable consumption	4	9	.263 (.059)	.214 (.048)	86.74	.212 (.037)	61.74
Panel C: Intervention–outcome combination (*n* = 9)							
Infopowerment × substance use	12	38	.147 (.065)	.182 (.055)	82.85	.185 (.053)	88.19
Infopowerment × obesity	9	19	.099 (.031)	.068 (.020)	43.29	.076 (.019)	44.48
Infopowerment × physical activity	8	18	.313 (.068)	.278 (.053)	64.94	.297 (.065)	77.76
Infopowerment × medical procedures	11	17	.177 (.048)	.161 (.042)	98.73	.172 (.035)	96.47
Infopowerment × nutrition	5	14	.160 (.074)	.126 (.061)	85.10	.122 (.062)	83.37
Incentives × general health	2	14	.182 (.041)	.173 (.035)	68.63	.246 (.038)	32.54
Incentives × substance use	6	14	.315 (.045)	.278 (.044)	72.58	.331 (.022)	35.77
Infopowerment × general health	5	7	.180 (.082)	.181 (.061)	80.20	.191 (.068)	79.11
Infopowerment × sustainable consumption	3	6	.228 (.060)	.204 (.054)	88.24	.193 (.032)	61.73

Table S2. Random-effects models. Raw treatment effect estimates versus clustered random-effects estimates. Analyses were conducted on all treatment effect estimates, and on estimates split by intervention class, outcome behavior, and combination of intervention class and outcome behavior. Clustered random-effects models were computed for intervention–outcome combinations with > 5 effect sizes. Raw treatment effect estimates clustered standard errors at the study level. The multivariate model included random effects for each study and for each estimate nested within a study [8]. The robust variance estimation model (RVE; [9]) added random effects through inverse variance weights. *n* = number of meta-analyses; *k* = number of estimates; *M* = mean; *SE* = standard error; I^2 = ratio of true heterogeneity to total variance across observed estimates, presented in %.

	n	*k*	Selection models	Clustered selection models		WAAP		RoBMA			
			M (SE)	*M (SE)*	β_p *(SE)*	*M (SE)*	$k_{d=0.14}$	*M* (95% *CI*)	BF_{10}	BF^{rf}	BF_{pb}
All effects	269	838	.103 (.009)	.084 (.011)	.105 (.025)	.099 (.007)	316	.063 (.044, .080)	139.72	Inf.	2.01E+27
Panel A: Intervention class (*n* = 2)											
Infopowerment	51	127	.088 (.024)	–	–	.111 (.018)	42	.025 (.000, .115)	0.45	Inf.	3.50E+06
Incentives	13	39	.242 (.039)	.110 (.029)	.120 (.066)	.207 (.062)	9	.215 (.000, .284)	33.98	3.45E+20	0.50
Panel B: Outcome behavior (*n* = 11)											
Obesity	75	195	.093 (.016)	–	–	.081 (.012)	92	−.001 (.000, .000)	0.04	2.80E+189	1.08E+11
Substance use	73	192	.105 (.018)	.105 (.021)	.238 (.126)	.113 (.016)	62	.048 (.000, .127)	1.41	1.06E+143	1.84E+03
Physical activity	60	136	.135 (.030)	.160 (.029)	.200 (.107)	.116 (.017)	36	.004 (.000, .082)	0.09	1.03E+129	2.32E+05
General health	30	71	.080 (.038)	−.017 (.056)	.026 (.024)	.123 (.031)	27	.081 (.000, .184)	1.79	4.67E+127	4.34
Sexual health	26	74	.087 (.024)	.063 (.027)	.086 (.041)	.076 (.016)	27	.018 (.000, .109)	0.39	7.33E+27	112.61
Nutrition	27	63	.070 (.024)	.083 (.041)	.088 (.055)	.098 (.016)	20	.017 (.000, .108)	0.33	1.85E+15	209.51
Medical procedures	23	50	.130 (.046)	.026 (.048)	.034 (.021)	.135 (.030)	28	.009 (−.017, .147)	0.16	Inf.	115.84
Feeding	9	22	.102 (.110)	–	–	.090 (.032)	7	.057 (.000, .276)	0.58	6.20E+23	45.04
Financial behaviors	5	13	.032 (.013)	–	–	.046 (.007)	12	.000 (.000, .036)	0.08	17.32	48.05
Oral health	4	13	.374 (.113)	–	–	.021 (.016)	2	.022 (−.024, .263)	0.25	7.41E+16	29.39
Sustainable consumption	4	9	–	–	–	.126 (.064)	3	.006 (.000, .149)	0.10	0.55	292.83
Panel C: Intervention–outcome combination (*n* = 9)											
Infopowerment × substance use	12	38	.123 (.053)	.056 (.050)	.240 (.323)	.147 (.035)	9	.010 (.000, .120)	0.17	3.28E+20	37.39
Infopowerment × obesity	9	19	.044 (.020)	.085 (.034)	.575 (.852)	.051 (.025)	7	.016 (.000, .070)	0.46	3.82	5.82
Infopowerment × physical activity	8	18	.104 (.139)	–	–	.165 (.065)	2	.012 (.000, .170)	0.15	0.36	861.57
Infopowerment × medical procedures	11	17	.107 (.073)	.024 (.015)	.010 (.012)	.119 (.043)	13	.025 (.000, .184)	0.33	2.68E+256	19.29
Infopowerment × nutrition	5	14	.007 (.067)	–	–	.020 (.010)	3	.005 (.000, .092)	0.10	4.30E+04	76.64
Incentives × general health	2	14	.172 (.050)	–	–	.159 (.071)	4	.112 (.000, .239)	2.37	2.20E+03	1.12
Incentives × substance use	6	14	.218 (.067)	–	–	.126 (.094)	2	.033 (.000, .272)	0.27	10.12	32.59
Infopowerment × general health	5	7	.043 (.126)	–	–	.260 (.030)	2	.037 (.000, .241)	0.39	1.26E+03	2.91
Infopowerment × sustainable consumption	3	6	–	–	–	.150 (.102)	2	.015 (.000, .209)	0.17	1.04	34.99

Table S3. Publication bias corrected effects. Results from frequentist and Bayesian publication bias correction procedures. Analyses were conducted on all effect sizes, and on effects split by intervention class, outcome behavior, and combination of intervention class and outcome behavior. Clustered random-effects models were computed for intervention–outcome combinations with > 5 effect sizes. Selection models applied step functions on *p*-value distributions to correct for selective publication for positive outcomes [10]. Step functions were computed across three *p*-value intervals ($p_{one\text{-}tailed} < .025$ vs. $.025 \leq p_{one\text{-}tailed} < .05$ vs. $p_{one\text{-}tailed} \geq .05$). If the model failed to converge, *p*-values were computed across only two *p*-value intervals ($p_{one\text{-}tailed} < .05$ vs. $p_{one\text{-}tailed} \geq .05$). Selection models did not converge for either selection model for the outcome category of sustainable consumption and for the combination of infopowerment × sustainable consumption. Clustered selection models applied the same step functions but allowed for clustered effect sizes [7]. The resulting conditional publication probabilities (β_p) were allowed to be asymmetric around zero. Model results were implausible for the outcome category of feeding and for the combinations of infopowerment × nutrition and infopowerment × general health. Clustered selection models did not converge for the intervention class of infopowerment;

for the outcome categories obesity, financial behaviors, oral health, and sustainable consumption; and for the combinations of infopowerment × physical activity, incentives × general health, incentives × substance use, and infopowerment × sustainable consumption. The weighted average of the adequately powered (WAAP; [11]) considered only adequately powered studies. With conventional levels of statistical significance ($\alpha = 0.05$) and power ($1 - \beta = 0.8$), the true effect must be at least 2.8 standard errors away from zero to reject the null (cf. [12], p. 441). The standard error of an estimate must therefore be smaller than the absolute value of the underlying effect divided by 2.8. As the true effect is unknown, we chose the median uncorrected effect ($d = 0.14$) from our sample as possible effect. We therefore only considered effect sizes with standard errors smaller than 0.14/ 2.8 = .05 (at $\alpha = 0.05$ and $1 - \beta = 0.8$). Robust Bayesian meta-analysis (RoBMA; [13]) determined estimates of true effects through weighting various publication bias correction methods by their performance. Bayes factors (BFs) indicate the strength of presented evidence with $1 < BF < 3$ suggesting weak evidence, $3 < BF < 10$ moderate evidence, and $BF > 10$ strong evidence. We considered $BFs < 1$ as inconclusive. n = number of meta-analyses; k = number of estimates; WAAP = weighted average of the adequately powered; RoBMA = robust Bayesian model averaging. M = mean; SE = standard error, β_p = publication probability; 95% CI = 95% credible intervals; BF_{10} = Bayes factor for an effect; BF^{rf} = Bayes factor for heterogeneity; BF_{pb} = Bayes factor for publication bias; Inf. = infinity.

Database	Boolean operator
Panel A: Health	
Web of Science	(TS=(Meta-analy* AND (eat* OR diet* OR nutrition OR "physical activity" OR exercise OR health) AND (intervention OR nudg* OR prevention OR "behavior change" OR "BCT" OR "promotion"))) NOT (TI= (patient OR disease OR diabetes OR disorder OR depression OR cancer OR "fall prevention" OR cardiovascular OR hospital))
EBSCO (Econlit, PsychInfo)	((Meta-analy* AND (eat* OR diet* OR nutrition OR "physical activity" OR exercise OR health) AND (intervention OR nudg* OR prevention OR "behavior change" OR "BCT" OR "promotion"))) NOT (TI= (patient OR disease OR diabetes OR disorder OR depression OR cancer OR "fall prevention" OR cardiovascular OR hospital))
PubMed	"Meta-Analysis"[Publication Type] AND ("Clinical Studies as Topic"[Mesh] OR "Health Promotion"[Mesh] OR "Preventive Medicine"[Mesh] OR "prevention and control"[Subheading]) AND ("Exercise"[Mesh] OR "Diet"[Mesh] OR "Eating"[Mesh] OR "Diet, Healthy"[Mesh])
Panel B: Finance	
Web of Science	(TS=(meta-analy*) AND TS=(financ* literacy* OR financ* behav* OR financ* educ*))
EBSCO (Econlit, PsychInfo)	((meta-analy*) AND financ* literacy* OR financ* behav* OR financ* educ*))
Panel C: Sustainability	
Web of Science	(TS= (meta-analy*) AND TS = (pro-environment* OR proenvironment* OR "environmental behavio*" OR "ecological behavio*" OR "green behavio*" OR recycling OR "water conservation" OR "water consumption" OR "energy) conservation" OR "energy consumption" OR "energy use" OR "resource conservation")
EBSCO (Econlit, PsychInfo)	((meta-analy*) AND (pro-environment* OR proenvironment* OR "environmental behavio*" OR "ecological behavio*" OR "green behavio*" OR recycling OR "water conservation" OR "water consumption" OR "energy) conservation" OR "energy consumption" OR "energy use" OR "resource conservation"))

Table S4. Boolean operators for systematic literature searches. Final searches for all panels were performed across all databases up until May 29, 2020. The initial search identified 41,406 records. After duplicate removal, record retrieval, and title and abstract screening, 1,212 reports entered full text screening. Assessing the eligibility of full-text reports resulted in a final sample of 269 meta-analyses and 838 unique treatment effect estimates.

Coding script	Description	Dependencies	Input files	Output files
Panel A: Performed analyses				
01-meta-meta-descriptives.Rmd	Descriptive analyses on included studies, effect sizes, and PICOS criteria	tidyverse	meta-meta-data.csv	–
02a-meta-meta-robma.Rmd	Code for robust Bayesian model averaging (RoBMA). Analyses require additional software (JAGS) and time to run.	tidyverse metafor rjags RoBMA	meta-meta-data.csv	X_RoBMA.RDS (RoBMA model) X_sum.RDS (summary) for each assessed combination of effects
02b-meta-meta-gmm	R-project for performing clustered selection model analyses; project and code adjusted from [7]	tidyverse RColorBrewer latex2exp xtable here	@functions_publication_bias.R meta-meta-data.csv	GMMEstimates AllSelectionModel.tex for each performed analysis tex-output was summarized as meta-meta-gmm-outputs.csv
02-meta-meta-models.Rmd	Code for performing random-effects models (multivariate random-effects, robust variance estimation) and publication bias analyses (selection models, weighted average of adequately powered). Code for combining outputs with RoBMA and clustered selection model results.	tidyverse metafor robumeta	meta-meta-data.csv meta-meta-gmm-outputs.csv X_sum.RDS for each assessed combination	meta-meta-effects-models.csv
Panel B: Reported output				
03-meta-meta-fig2.Rmd	Stacked barplot, scatterplot, density-dot-plots with boxplots	tidyverse ggdist RColorBrewer	meta-meta-data.csv	fig2-panelX.pdf for panels A to D
04-meta-meta-fig3.Rmd	Specification curve consisting of a forest plot and tile plots	tidyverse RColorBrewer patchwork	meta-meta-data.csv	fig3.pdf
05-meta-meta-fig4.Rmd	Forest plots on random-effects and publication bias corrected effects	tidyverse RColorBrewer scales patchwork	meta-meta-effects-models.csv	fig4.pdf
06-meta-meta-tabSM1.Rmd	Raw effect size distributions	tidyverse	meta-meta-data.csv	sm1-table.csv
07-meta-meta-tabSM2.Rmd	Random-effects table	tidyverse	meta-meta-effects-models.csv	sm2-table.csv
08-meta-meta-tabSM3.Rmd	Publication bias table	tidyverse	meta-meta-effects-models.csv	sm3-table.csv

Table S5. Overview of analysis scripts.