Generalizability of choice architecture interventions

Barnabas Szaszi[1†], Daniel G. Goldstein[2], Dilip Soman[3], Susan Michie [4]

[1]Institute of Psychology, Eotvos Lorand University, Budapest, Hungary
[2] Microsoft Research, New York, NY, US
[3] Rotman School of Management, University of Toronto, Toronto, Canada
[4] Centre for Behaviour Change, University College London, London, UK

[†]email: szaszi.barnabas@ppk.elte.hu

Now published as:

Szaszi, Barnabas, Daniel G. Goldstein, Dilip Soman, Susan Michie. (2025). Generalizability of Choice Architecture Interventions. **Nature Reviews Psychology.**

Available at https://rdcu.be/ewD2z

"Note: This version is pre-typesetting, so minor differences may exist compared to the final published version."

Abstract

Although a given choice architecture intervention ('nudge') can be highly effective in some conditions, it can be ineffective or counter-productive in others. Critically, researchers and practitioners cannot reliably predict which of these outcomes will happen based on current knowledge. In this Review, we present evidence that the average effectiveness of choice architecture interventions on behavior is smaller than often reported and that there is substantial heterogeneity in their effect. We outline the obstacles to understanding generalizability such as the complex interaction of moderators and their dynamic change over time, clarify dimensions of generalizability and review research practices (including systematic exploration of moderators and practices designed to enhance generalizability) that could help the field more efficiently accumulate evidence on generalizability. Adopting these practices is essential for advancing nuanced theories and for more accurately predicting the effectiveness of choice architecture interventions across diverse populations, settings, treatments, outputs and analytical approaches.

[H1] Introduction

Choice architecture interventions [1] are a subgroup of behavioral interventions [2] that aim to achieve behavior change by changing the proximal physical, social or psychological environment to prompt or guide behavior. That is, choice architecture interventions do not require the investment of substantial resources, legislation or regulation; do not involve extensive training programs; and preserve 'freedom of choice' by maintaining choice options[3,4]. As an example of the value of this approach, consider that people often fail to appear in court for low-level offenses [5]. Although the typical remedy to this issue might be a policy response of increased fines or threats of forced pretrial detention, choice architecture interventions have effectively complemented these measures by making critical information more salient on summonses or by sending text message reminders [5]. Similarly, reducing household energy consumption is often only approached through increasing the price of electricity, but choice architecture interventions have been shown to substantially reduce the energy consumption of households simply by communicating social norms [6] in some context [7].

To the casual observer, it might seem that choice architecture interventions like these yield economically relevant effects and are easy to implement, with few downsides. Behavioral sciences enthusiasts can readily bring to mind instances of successful interventions, such as a field experiment in collaboration with the United Kingdom tax administration authority in which the statement "Nine out of ten people in the UK pay their tax on time" in official tax reminder letters resulted in a multimillion-pound increase in tax revenue [8]. However, for several potential reasons, such as the intricacies of memory recall [9], the dynamics of scientific publishing [10] and the canonical successful examples used in the early 'nudge' discourse, it might be more challenging to recollect failed interventions or studies that provide a more nuanced picture. However, in reality, the effectiveness of these interventions can vary substantially. For instance, social norm interventions that work in one context can fail when applied to different populations, settings, time periods, or through alternative implementation methods [7,11].

To fulfill the potential of choice architecture interventions, researchers need to be able to predict what treatment works, for whom, when and to what extent across different populations, settings, outcomes, and times [12–14]. However, there is little published evidence about the generalizability of nudges [15–18]. Even with advanced knowledge of key parameters like the population, setting, timing, and implementation details, it remains challenging for experts to accurately predict the effectiveness of these interventions [19]. Without a deeper understanding of the generalizability of choice architecture interventions, their applicability will remain limited.

The ultimate goal of investigating generalizability is not to identify universally generalizable choice architecture interventions—no single behavioral intervention can produce the same outcomes in all situations. Rather, the field needs consistent evidence accumulation on the generalizability of choice architecture interventions, linked to mechanisms that underlie the interventions (instead of simply reporting the effectiveness a given interventions in a specific setting; for a broader discussion see[16,20,21]). Increasing knowledge about the generalizability of the underlying mechanisms could increase understanding of how a study finding applies across different target populations, outcomes, operationalizations and implementations, settings, time and analysis approaches. This definition of generalizability is interlinked with the concept of external validity (although the extent of this overlap varies according to different definitions) [21,26–30].

A deeper understanding of generalizability would benefit both researchers and practitioners. For researchers, it could lead to more developed theories about behavior change, and for practitioners it could increase practical value through better predictions about which intervention works under what conditions. However, these two groups have different goals and incentive structures: to advance a fundamental and broadly applicable understanding of phenomena versus to maximize the impact of an intervention in a specific setting. In the short term, learning about generalizability of human behavior is more aligned with the goals and incentives of the researchers. Thus, we advocate for more focus on the systematic exploration and test of when, where and to what extent choice architecture interventions have an impact rather than on testing the effectiveness of a specific intervention in a particular setting, as most choice architecture research reports have done over the last decades [16,17,31].

In this Review, we address generalizability of choice architecture interventions from the researcher perspective. Our goal is to support knowledge building on how choice architecture interventions can be applied across different target populations, outcomes, operationalizations and implementations, settings, time and analysis approaches [22–24]. We first discuss the average effectiveness of the 'nudge' interventions and the substantial heterogeneity in their impact. Then we consider what is known about the generalizability of these interventions and summarize the main obstacles to generalizability. It is beyond the scope of our paper to define what is 'good enough' generalizability, as what constitutes 'good enough' will always depend on the circumstances and the aim of the specific research (for discussion of evaluative criteria, see [21]). Rather, we go on to review the conceptual dimensions related to generalizability (including analytical variability) and review the research practices that could help the field more efficiently accumulate evidence on it. We conclude by discussing how the adoption of the reviewed research practices – along with additional efforts such as large-scale collaborations and harnessing artificial intelligence could lead to more robust theories of choice architecture mechanisms and to more accurate predictions about their effectiveness.

[H1] The effectiveness and heterogeneity of choice architecture interventions

Understanding the typical effect sizes and heterogeneity of choice architecture interventions sheds light on the importance of understanding their generalizability. If the effects of choice architecture interventions are large with low heterogeneity, the theoretical and practical value of learning about the generalizability of the underlying mechanisms would be limited. In such a scenario, one could reliably expect these interventions to be effective across contexts, regardless of specific features. However, if the average effects are small with considerable heterogeneity, then understanding generalizability becomes crucial. Those applying choice architecture interventions must recognize that in many cases, these interventions may have only a limited impact—and in some instances, they may even be ineffective or counterproductive. We start by summarizing why effect size estimates in meta-analyses of choice architecture interventions are often inaccurate and inflated, then highlight the more reliable estimates and close this section by showing that the heterogeneity in the effect sizes is large.

Our discussion of effect sizes is largely based on bias-corrected meta-analyses that include interventions from a variety of domains (e.g., health, finance etc.) and intervention types. The value of the effect size and heterogeneity estimates we discuss here is to correct researchers' and practitioners' predictions that might stem from well-known success stories or failures. For specific questions about whether a specific choice architecture intervention works in a given setting, researchers should consult the results of more focused meta-analyses within a domain (such as a meta-analysis of digitally delivered interventions in health insurance choices).

[H2] Effect size

To provide informed estimates on the impact of nudging, the field has increasingly turned to systematic review-based meta-analyses to map and combine the results of multiple studies. A meta-analysis of 100 primary publications and 317 effect sizes concluded that nudge interventions produce a median impact of a 21% change in the target behavior [32]. Another study that reanalyzed a combined set of 26 randomized control trials (RCTs) (from [4,32]) found an average 33.4% impact on the target variable [33]Finally, another review of more than 200 primary studies with over 440 effect sizes estimated the average effect of nudging to be a Cohen's d of 0.43 [34]. According to these meta-analyses, the average effect size of choice architecture interventions seems quite large. However, it was often noted that such average effect sizes are implausibly large [33,35,36].

There are multiple reasons to assume that the average effect of choice architecture interventions should be smaller than the estimates presented above. It has been widely argued that in social and behavioral sciences, phenomena are causally dense, meaning that there are always many influencing factors that operate concurrently [31,37,38]. The Piranha theorem suggests that in such systems, the interaction and interference of these factors would overwhelm most single effects leading to predominantly small main effects [39]. Thus, although there may be some large and predictable effects on behavior, such effects are likely not to be numerous.

Another reason that the effect sizes are likely to be inflated so empirical evidence might not accurately represent the average effect [40] is the potential for publication bias, which is widely acknowledged to result in overestimated effect sizes [41,42]. The most robust approach to address this bias when drawing meta-analytic inferences involves obtaining a dataset that encompasses all published and non-published studies, however, this is only feasible in very rare cases. For instance, a meta-analysis of 126 RCTs, including all trials from the two largest nudge research units in the US, found that the average effect of the applied behavioral interventions was an 8% impact on the target behavior [33]. A later study found that the estimate dropped to 6.4% when an invalid paper was excluded from the same meta-analysis [43]. Notably, both of these estimates are much smaller than the estimates that draw on published academic papers (such as 21% [32]). If collecting all papers is not feasible, another way to tackle publication bias is to use analytical methods to adjust for its presence. For instance, a study that applied three different bias-correcting methods found small and varying effects ($d = -0.01$, $SE = 0.02$; $d = 0.07$, $SE = 0.03$; $d = 0.08$, $SE = 0.03$) [36], consistent with the prior estimates when assuming severe publication bias ($d=0.08$) [34], while another study even argued that after adjusting for publication bias there is no substantive evidence for nudging [35] (See more discussion on this topic [44,45].)

Even meta-analyses that rely on an all published and non-published studies [33,46] can be biased, because the set of conducted studies are not chosen randomly from the multiverse of all possible studies. Instead, researchers and practitioners are likely to be influenced by their lay theories about the effectiveness of the interventions, by the costs of the potential studies and by the perceived novelty of studies with higher theoretical or rhetorical value (for similar discussions see [31,47]). These biases should also be noted when interpreting meta-analytic results.

Furthermore, there is another reason to assume that meta-analyses overestimate the average effects of interventions: the existence of common research practices such as p-hacking that can artificially increase the effect sizes estimates and by that the likelihood of obtaining false-positive rates [48,49] Study preregistration offers a potential solution to mitigate the effect of p-hacking and thus the bias in the literature caused by false positive findings [50,51]. In line with this expectation, effects from a random set of non-preregistered psychology studies (median r = 0.36) were considerably larger than effects from preregistered psychology studies (median r = 0.16) [52]. Similarly, across different fields it has been found that large-scale registered replications typically observe much smaller effects than the original studies [53–56] A systematic review and meta-analysis of preregistered choice architecture intervention studies observed that even without correcting for publication bias, the an average effect size is substantially smaller (d=0.23) than in meta-analyses based on predominantly non-preregistered studies (such as d =0.43 from [34]) [57]. However, even eliminating publication bias and p-hacking would leave room for some uncertainty. Fraud and data-fabrication could also inflate the observed results in the literature compromising the validity of the meta-analytic results [58,59]. Yet, there are currently no studies available to estimate or adjust for these effects.

Finally, any conclusions drawn from meta-analytic results should take into account the context and limitations of the underlying studies [47,60,61]. Behavior varies considerably according to context, so the average effect of different studies – for instance, defaulting people in the US into organ donation or reminding people in Kenya about healthy food – does not accurately predict the effect of even the same intervention in a different context. (The default intervention is a strategy that sets a preferred option as the preselected choice that is automatically applied if no alternative is specified by the individual).

In sum, both theory and the average effect of choice architecture interventions in the documented literature suggest that the average effect is smaller than prior research suggested. However, if the effects were small with low heterogeneity, practitioners could still confidently apply them, as small reliable effects at scale could create a large value on the societal level. However, evidence suggests that the effect of choice architecture interventions is extremely heterogeneous and unpredictable.

[H2] Heterogeneity

Domain-general meta-analyses revealed considerable heterogeneity among choice architecture interventions [32–34]. When assuming publication bias, 95% of the effect sizes of nudges should fall between d= -0.92 and 1.08[36]. This wide range from negative to positive effect sizes suggests that in a substantial proportion of cases, new intervention implementation would yield null effects or even be counterproductive. To put heterogeneity into perspective, imagine considering a medication to improve your sleep. If the medication has a similar average effect size as choice architecture interventions, it might increase your sleep on average by 8 minutes per night, which might be a considerable benefit in the long run. However, if the heterogeneity of the medication's effect also matches that of choice architecture interventions, it might increase your sleep by up to 108 minutes per night or reduce it by 92 minutes. Crucially, there is no way of knowing which effect you might experience from this information alone.

The substantial heterogeneity of choice architecture interventions becomes more understandable when one considers the potential moderators of their effectiveness. There have been several [16,17,62–64] attempts to catalog these moderators across different dimensions (see Figure 1). To exemplify the variety of possible moderators, consider the effectiveness of choice architecture interventions using social norms to reduce littering. In this context, researchers can consider the potential moderators that are associated with what to measure (behavior versus attitudes), how the intervention is delivered (such as the timing and salience of reminders), the specific norms tested, and the target population's attentional capacity. Furthermore, many of these moderators could interact, for instance one might consider the moderating effect of attentional capacity on the timing of the intervention or the salience of the reminders.

Even the effect of the most robust choice architecture interventions varies substantially when tested across units, outcomes, or settings [11,32,34,65]. For instance, a default effect study [66] conducted on 11 different participant pools, found significant effect in 10 out of 11, but the effect size varied from about .2 to .9 Cohen's d [67] A meta-analysis revealed that although the vast majority of default studies robustly produced the intended effect, some of the studies didn't find an effect (17%) or even demonstrated a negative effect (3.5%) [68]. For example, in a study on colonoscopy examinations, attendance rates were 22% lower in the default condition — where participants were mailed a prescheduled appointment date and time — compared to when they had to schedule their own appointment (63%) [69].

Replication studies can also provide insights into the heterogeneity of choice architecture intervention outcomes. Although the choice of the population is often less strategic in many single-study replications than in multi-site studies [70,71] , such studies can be used to collect evidence about the boundaries of generalizability. Although direct replications of choice architecture intervention studies are far from mainstream, there are some relevant large-scale replication projects. For example, a reanalysis of 100 replications from the field of psychology[53] showed that replication success was negatively associated with the contextual sensitivity (the extent to which independent coders predicted that a finding would vary depending on time, culture or location)[72], confirming the importance of contextual moderators behind heterogeneous results. In a large-scale endeavor replicating more than 100 studies from the field of judgment and decision-making with participants from Mechanical Turk, the researchers also observed varying results: while some studies had effects larger than in the originally published papers[73], other studies had mixed results [74,75] with some even showing opposite patterns to the original[76].

The optimistic interpretation of the large variance in effect sizes is that the effects of choice architecture interventions are not always small. This interpretation is consistent with the theoretical predictions of the Piranha theorem: if many effects are operating in an uncontrollable way, they can easily combine in unexpected ways, leading to large overall effects. If our field could learn about the generalizability of the choice architecture interventions' effects across contexts, practitioners could select for each situation the most effective version of the most effective nudges and achieve substantial impacts at a minimal cost. However, being unable to generalize choice architecture interventions could lead to negative consequences, as shown by examples in which the application of choice architecture interventions unwittingly resulted in the discouragement of people to take part in cancer screening [69] or diminished drivers' focus on the road[77].

[H1] Obstacles to learning about generalizability

Based on the reviewed evidence, the field currently lacks sufficient knowledge about the generalizability of the effectiveness of interventions. In reviewing the main obstacles to generalizability here, we suggest that the field's inability to reliably generalize arises from both the nature of the problem and from suboptimal research practices. These challenges suggest that the generalizability problem cannot be solved quickly and without a substantial cultural change in choice architecture intervention research.

[H2] Complexity

The core obstacle to learning about generalizability stems from the fact that moderators influencing the effectiveness of behavioral interventions are almost always multiple and a priori unknown. Based on the prior literature, we focus here on five dimensions that each contain an indefinitely large number of potential moderators: unit, treatment, outcome, setting, and analysis (Figure 1). This grouping of moderators is not intended to provide an exhaustive list but rather to provide an indicative list about the main types of moderators that correspond to each of the dimensions. For example, if one wants to know the effect size to expect from a default intervention in a given situation, they should take into account factors such as the topic domain, the importance of the topic for the decision-maker, the motivation of the individuals, the (perceived) ease of change, (perceived) endorsement, (perceived) endowment, and the attentional and emotional capacity of the decision maker at the moment of the decision [68].

As a further complication, moderating factors can interact with one another. For instance, the salience of choice architecture intervention might interact with the attentional capacity of the target audience. As the number of moderators increases, the potential interactions between these moderators grow exponentially challenging the predictability of intervention effectiveness. In the example above, to understand the effects of all eight moderators—which is still very likely an incomplete list— there are also 247 potential interactions between them. Any additional moderator would further increase this number exponentially (for instance, 9 moderators results would result in 502 potential interactions.)

Another challenge is that the constant change of societies including the change in preferences, norms, abilities, and opportunities dynamically impact the effect of moderators. Consequently, the effect of old moderators might disappear and emerging moderators might appear, inducing unexpected heterogeneity. Thus, the effect of an intervention in a new time and setting can deviate from the original findings due to hidden and unexpected moderators that were not relevant (or not considered relevant) when the original study was conducted[49]. For instance, a few decades ago there was negligible discussion about the well-being of animals [78], compared to 2023, when 29% of US vegetarians listed morality as an important factor in their decision not to eat meat [79]. Thus, currently people's beliefs about moral obligations could have an important moderating role for any intervention that aims to increase vegetarian eating, whereas such a moderating factor would not have been relevant in the past. To predict emerging moderators perfectly would require predicting the future change in societies, which is clearly an unrealistic expectation. However, the role of emerging moderators remains important to keep in mind. Empirical evidence on the impact of emerging moderators remains limited and subject to debate. Although some replication initiatives revealed substantial heterogeneity across different settings, [70] which suggests the existence of hidden moderators, other studies have not found such evidence [80].

In sum, the complexity of moderating factors - their sheer number, their dynamic and interacting nature, and their evolution over time - presents a profound obstacle to learning about generalizability. All these makes it extremely challenging to predict when, where, and for whom interventions will be effective.

[H2] Research practices

In addition to the complexity inherent in research studies and their context, certain widespread research practices pose an obstacle to learning about the generalizability of effect sizes from study findings. These research practices arise from a combination of structural constraints—such as limited research funding and time—and incentive systems that reward rapid, publishable, significant results. As a result, decisions about sampling, study design, and data analysis are often guided by short-term impact rather than by goals of optimizing knowledge accumulation and robustness.

When conducting studies, researchers always face constraints due to limited financial resources, which can lead to an oversampling of populations who are less costly to recruit, are easily accessible or reachable even without direct financial compensation [81–83]. A predominant trend is the overrepresentation of participants from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies, from university settings and people who are can be reached via online survey platforms [84] Consequently, the participant populations of many behavioral science studies are not representative of the broader populations to which they aim to generalize their findings [84,85]. This lack of representativeness raises concerns because individuals from underrepresented cultural, demographic, or psychological groups may respond differently to interventions than those from the populations typically studied. Thus sampling participants with specific characteristics may impact the generalizability of study outcomes. Variation exists in results between WEIRD and non-WEIRD populations suggesting that cultural characteristics play a big role in behavioral sciences [84]. It has been shown in the health and medical decision-making literature that digitally collected and traditionally collected data may lead to opposing findings [86]. Classic social psychology effects have also been found to be varied between undergraduate and general population samples [87]. Finally, certain foundational effects in behavioral science, such as the presence of decision biases were also shown to be stronger within populations of participants that closely resemble the characteristics of the undergraduate-student samples of the initial studies [87]. Together, these findings underscore that heavy reliance on WEIRD or convenience samples limits our ability to draw accurate conclusions about how interventions work across broader, more diverse populations. As long as participant samples remain narrow and unrepresentative, our understanding of generalizability will remain fragmented and potentially misleading.

Researcher decisions about the operationalization and implementation of the treatments and outputs of a study can also influence the impact of treatments, which in turn affects generalizability—because the findings may not hold when those design choices differ across studies or real-world applications. For example, in a mega-study in which different research groups devised designs to test five different hypotheses related to moral judgments, negotiations, and implicit cognition, the resulting direction of effects were contradicting each other in several cases[88]. In a meta-analysis of 45 different tests of competitions and moral behavior, there was substantial design heterogeneity, which was approximately 1.6 times larger than the standard error of the effect sizes[89]. These results suggest that even when studying similar hypotheses, variation in researcher design decisions can lead to divergent findings, making it difficult to determine whether observed effects are robust or simply artifacts of specific methodological choices—thus posing a significant obstacle to generalizability. Heterogeneity observed in intervention-specific meta-analyses [68] further underscores the critical role of specific implementations of interventions on outcomes. For instance, stimulus variation [90–92]—including the linguistic content [93,94]—can substantially impact the effectiveness of interventions. For example the stimuli that the experimenter chooses to test led to opposing results and therefore opposing conclusions about people's biases in consequential domains like saving [95]. These results underscore the necessity for systematically exploring stimuli variation in pursuit of generalizable results (for an extended discussion, see [31]). The challenge of generalizability also extends to other aspects of research design[96]. For instance, results can be altered by the mode of intervention delivery - the same letter emphasizing the importance of paying taxes and possibility of a tax audits has smaller effect on tax paying behavior when received by post versus when delivered in person by a tax officer [97,98] . Finally, the selection of outcome measures can also influence the estimated effectiveness of an intervention. For example, an intervention might improve financial attitudes in the short term but fail to produce meaningful changes in long-term behaviors such as saving or borrowing [99–101].

Decisions about data analysis, such as data-processing, model specifications or inference criteria can also lead to qualitatively opposing inferences from the same dataset [102–105]. Using the same dataset to answer the same research questions, researchers came to different conclusions across a wide range of topics from neural basis of risk-taking [106] to religiosity[107]. In an initiative to test the impact of analysis decisions, 100 behavioral and social sciences studies were independently re-analyzed; all re-analyses resulted in the same conclusion as described in the original study for only 34% of the studies (Box 1)[108]. However, a typical scientific journal article only contains the results from one or a few analysis pipelines [109], leaving unexplored the generalizability across alternative analytical choices [110,111].

Taken together, these factors reveal a pattern: methodological choices—whether about who is studied, how interventions and impact measurements are operationalized and implemented, or how data are analyzed—systematically shape what we learn from behavioral science. Without broader and more deliberate efforts to address these constraints and biases, our understanding of generalizability will remain limited, potentially distorting the real-world implications of behavioral interventions.

[H1] Enhancing evidence accumulation about generalizability]

Although the obstacles to understanding generalizability are not trivial, there are research practices that can help overcome them. In this section, we review research practices across different stages of choice architecture research projects that can improve knowledge accumulation regarding the generalizability of interventions. These practices complement prevalent methods of testing the effectiveness of a choice architecture intervention within a specific context using an experimental setup. These practices are not all relevant or feasible for all choice architecture research projects, but their adoption could substantially increase knowledge of choice architecture intervention generalizability.

[H2] Mechanisms

To reveal how the mechanisms underlying choice architecture interventions generalize, researchers need to first define the mechanisms that underlie the specific choice architecture intervention under investigation [14,16,21,112–114]. In its most common form, 'mechanism' refers to a causal chain between the treatment and the outcome [20,21] (for alternative definitions see [115–117]). So defining the mechanism means identifying the sequence of psychological or behavioral processes triggered by the intervention that ultimately lead to the observed effect. [14,118–120]. For example, in a default intervention promoting organ donation, the mechanism might involve ease (the reduced the effort required to make the choice), which leads to higher consent rates. Assuming so, to understand the generalizability of defaults, researchers should focus on understanding when and why easing the effort required promotes a given choice. For instance, the effectiveness of default effects has been argued to be driven not only by ease, but also endorsement (the perception that the default option is recommended or approved by the choice architect.), and endowment (the extent to which the pre-selected option is viewed as the status quo) [67,69,122]. In such cases, the description of the mechanism and the exploration of generalizability should entail all potential routes.

[H2] Moderators

The systematic exploration of moderators is notably absent in published reports of choice architecture interventions [16,122]. Depending on available resources and the specific settings, various methods can be flexibly employed to uncover potential moderators such as reviewing published papers and meta-analyses, consulting domain experts or engaging the target population through surveys, interviews, and focus groups [123] [122].

To support the researchers' endeavor to the reveal the boundaries of generalizability and to explore of moderators, prior research proposed several approaches [21,111,124,125]. Theory-focused classifications highlight the crucial role of auxiliary, statistical and inferential assumptions that may moderate the effectiveness of findings: theoretical claims may or may not hold when auxiliary assumptions change, the same applies to empirical hypotheses under different statistical assumptions, and to statistical hypotheses depending on the inferential framework used [110,111]. Others emphasized the importance of the sample, research design and analysis path as key dimensions behind the heterogeneity of research findings and limiting generalizability [125]. The widely used UTOS framework proposed units, treatments, outcomes, and settings as the core dimensions central to generalizability[124]; while the more recent M-STOUT framework complemented this by adding mechanism and time as important dimensions to consider[21].

Building on these works, we outline five conceptual dimensions of potential moderators that choice architecture researchers should consider when aiming to understand generalizability: units (i.e. population), treatment, outcome, setting, and analysis. (Figure 1, Box 1). The type of moderators corresponding to units refers to the characteristics, such as capabilities or preferences, of the treated population The categories of treatment and outcome pertain to the ways the independent and dependent variables are operationalized and implemented. Setting encompasses the broadly defined environment where data is collected. Each conceptual dimension can contain an indefinitely large number of potential moderators that need to be explored for specific use cases. Importantly, cutting across these five dimensions, interactions among moderators and change over time should also be considered.

Compared to prior studies, this list suggest several changes rooted in methodological development that happened after the replication crisis in psychology [26,48]. As researchers tried to understand why prior research failed to replicate, they had to realize that even minor changes in the way an intervention is implemented[31,88,89,125] and analytic decisions are made [103,111,125] can lead to differing results and conclusions. Furthermore, all the moderators can change over time – for example, emerging moral values around animal welfare may now influence meat consumption, whereas they did not several decades ago. However, *time* is an overall effect expressed across all dimensions, but not an independent dimension itself as it is only expressed by the change of specific moderators within the highlighted five dimensions.

[H2] Stages of research

When accumulating evidence about generalizability, the identification of potential moderators can guide the research design including decisions about sampling, whether to measure or manipulate specific moderators, and how to make analytic decision (Figure 2). Researchers can strategically sample from a population that enables them to explore the moderators, measure or manipulate the actual levels of moderators, and analyze the data with heterogeneity in mind. Finally, considerations for reporting ensure that each research study is as informative as possible.

For example, consider a research team studying the effect of default interventions on energy tariff choice. Because considering moderators from the outset helps researchers decide how to design their study to learn about generalizability, they firs review and reflect on potential moderators and their potential interactions – across the five conceptual dimensions outlined above. Next, based on their available resources and the perceived importance of the moderators, they strategically recruit participants with varying levels of attentional capacity, decide to manipulate the ease of switching tariffs experimentally and measure the perceived endorsement of the default option. They also decide to perform a multiverse analysis to assess how arbitrary analytical decisions influence their estimation of the treatment effect.

[H3] Sampling

The characteristics of the research sample determine and limit the ability to explore the effects of the potential moderators and to explore the generalizability of the interventions [126,127]. When researchers test the effectiveness of behavioral interventions in convenience samples [128] they also constrain how much they can learn about the generalizability of their findings. Collecting representative data using random sampling from the target population enables the exploration of generalizability across different dimensions. However, this sampling is rarely an option because it is always costly and in several situation researchers cannot randomly sample people from the whole population. Although collecting quasi representative samples along variables such as gender, age, education or employment—demographic characteristics usually available for probability sampling by data collection providers—can improve researchers' ability to explore generalizability[129], they cannot know to what extent these standard variables - usually available for probability sampling by data collection providers - overlap with the relevant moderators of the target intervention. Purposive sampling, or selecting participants based on specific characteristics of interest to the researcher (here the moderators) [130], can potentially provide an alternative solution. Through this approach, sampling becomes an integral part of the study design, aligning with the goal of revealing the generalizability of the results along the moderators [17,131]. For instance, researchers who hypothesized that school achievement would be a key moderator of an online mindfulness intervention employed a purposive sampling strategy to select schools stratified by this dimension[132]. Using this purposive sampling approach provided the authors with enough data variation and power to test and confirm the moderating effect of baseline achievement on the intervention's effectiveness.

[H3] Measurement

Measuring the level of the potential moderators – such as the perceived endorsement of the default option - along the applied intervention is one of the most easily applicable methods to learn about generalizability[16,17,64,133]. Even when assessment of the moderators' impact is not the primary focus of a study, learning about their impact could be essential building blocks of knowledge as other researchers conducting secondary analyses and meta-analyses could leverage this information to construct more nuanced models to predict the generalizability of choice architecture interventions. In the past, limited sample sizes often hindered researchers from analyzing the effect of all potential moderators and interactions in a model but advancements in machine learning and the increasing feasibility of large-scale testing [e.g., 134–136], enable researchers to effectively analyze a broad range of moderators. Although, conducting more extensive measurement of moderators corresponding to units, treatment, outcome, and setting, beyond basic demographics could be a challenge especially when working with field partners. Researchers should be prepared to allocate resources to help their partners to build up the necessary capabilities.

[H3] Systematic variation

When resources permit, behavioral intervention researchers should systematically vary (randomize) the levels of the chosen potential moderators. Integrative experiment design[31] promotes systematically mapping the relevant dimensions of the design space—in this case the moderators—and then iteratively testing the effect of these dimensions in a commensurable way. This strategy, also known as radical randomization[137] can include varying elements such as the stimulus [92,95,138], modality [98], or other specifics of design [89], including features relating to the setting [132] or the population [87,139,140]. Large scale collaborations including multi-site many labs studies [70,71] are examples of how researchers can systematically varying moderators in their research. Although it is not possible with any single experiment to map all the potential moderators of a behavioral phenomenon, there are some ongoing large-scale efforts that vary multiple dimensions of generalizability, For instance, in the Global Happiness Megastudy[141], researchers vary the population, the research design and the analytical choices to map generalizability across these dimensions.

[H3] Analysis

Analysis decisions, including how to process the data, construct the statistical model, select algorithms and software for model estimation, define the inference criteria (frequentist, Bayesian, or likelihood),and use of machine learning versus computational modeling provide a wealth of potential influences on the outcomes of a study. Although several studies suggest that the arbitrary analysis decisions of researchers can lead to qualitatively different conclusions [106,108,142], a typical research project reports one primary analysis approach, maybe with a few robustness checks [103,143]. If researchers want to know whether their results generalize beyond their specific analysis choices, and see how their choices moderate their results, they need to systematically conduct and report all justifiable analysis paths[103]. Researchers can follow two possible types of solutions (extensively discussed elsewhere [109]): multiple investigators can independently follow a single analysis approach [102] or a research team can perform numerous analyses across the set of reasonable pipelines to reveal how the results generalize (the multiverse approach) [103,104].

[H3] Reporting

When reporting behavioral intervention results, researchers should assume that intervention effect sizes are largely heterogeneous [17,64] Researchers should systematically report which moderators they have data on and which they do not and detail how the measured and varied moderators and their interactions influence the main treatment effect. If multiple analyses or a multiverse analysis is conducted, researchers should report how the analytical choices impact the results.

Returning to the five dimensions discussed above (unit, treatment, outcome, setting, analysis), researchers should clearly state which dimensions of generalizability and specific moderators were tested. Ideally, they should also report the variables identified as potential moderators during the exploration phase, even if these were not tested later. For example, consider a study using a default intervention. If researchers measure perceived ease of change but cannot assess other relevant moderators—such as perceived endorsement or perceived endowment—they should report which measured moderators were significant and which were not, while also acknowledging any blind spots (i.e., potentially important moderators that were not measured). This transparency enables readers to make more informed judgments about the boundaries of our knowledge regarding the findings' generalizability (see also[145]). When considering the moderators regarding the units, and how to report them, researchers could turn to the 'constraints on generality' statement[146] that provides a structured framework for articulating constraints on generalizability with a focus on identifying and justifying populations they aim the results to generalize.

In summary, practices enhancing the knowledge accumulation about generalizability are numerous, but we do not pretend that they are easily or trivially implemented. Some of these practices—such as reviewing the potential moderators or conducting a multiverse analysis — can be implemented by researchers themselves. Other practices—like creating more extensive reporting and more exploratory studies —might face backlash from research journals. Finally, other practices such as measuring additional moderators might be hard to implement when collaborating with field partners who have limited flexibility. Designing studies that allow us to draw conclusions about generalizability often requires significantly more resources than studies that do not address this issue .

We acknowledge that the implementation of these practices puts extra burden on researchers. However, the integration of these practices into the research process is worth the effort. Without accelerating the understanding of generalizability, researchers risk continuing to produce results that are incommensurable, lead to limited theoretical advancement and limited practical applicability.

[H1] Summary and future directions

Contrary to the suggestion of prior meta-analyses, the average effectiveness of published choice architecture interventions is smaller than typically reported. Furthermore, there is substantial heterogeneity in their observed effect sizes and our field has insufficient evidence to predict the effectiveness of choice architecture interventions across different settings and implementations. The limitations of generalizability arise from the inherent complexity of these interventions and dynamically interacting moderators that shape the outcomes. Additionally, suboptimal research practices contribute to limited knowledge about when and why choice architecture interventions work.

We also reviewed concrete practices that could improve evidence accumulation about generalizability within research at different steps in the research pipeline. To enable these practices, research funders must recognize the need for additional resources to incorporate the study of generalizability into research. Last but not least, intervention researchers should embrace the rigor-enhancing open science practices to maximize the value of ongoing research efforts [57,147].

Beyond these recommendations, two emerging trends have the potential to catalyze the process of learning about the effectiveness of behavioral interventions: large-scale collaborations and technological advancements in artificial intelligence and machine learning.] Collaborative approaches can be a viable way to collect large datasets, enabling the exploration of systematic variance along a large number of potential moderating factors that individual behavioral science teams cannot do on their own. Researchers might volunteer to pool their samples [71,148] or financial resources [149] to collect data from a more diverse or representative population. However, simply collecting larger datasets does not fix the generalizability challenge. If sampling methods are biased or do not vary along the key moderators, even large-scale RCTs that are shown to be effective in one setting will not necessarily be effective in others [150,151]. Yet, large and strategically diverse datasets can be leveraged to discover nuanced relationships between the multitude of variables influencing the effectiveness of interventions [152].

In other collaborative approaches, some parts of the research process might be crowdsourced, such as the design of various interventions [19,153] or different versions of the same interventions, [89] which provides data about the generalizability of effectiveness across implementations. Alternatively, the moderator exploration can be conducted by the involvement of an expert panel [129.] or the analysis can be conducted by a group of researchers to ensure the analytical robustness of results[102]. Large-scale crowd-sourced collaborations can be a viable way to overcome resource constraints, although they involve different challenges (including coordination and collective action problems [71], misaligned incentives [154], and the pressure to produce rapid results [155]).

Advancements in artificial intelligence models hold substantial potential to accelerate evidence accumulation on the generalizability of interventions. For instance, large language models (LLMs) can in some cases simulate human-like responses and behavior potentially mimicking the response of different groups of people [159,160], and could potentially be used to explore the generalizability of interventions' effectiveness across different populations. Furthermore, LLMs could generate and test hypotheses[161,162] at an unprecedented speed, supporting the exploration of sources of heterogeneity. However, results from such simulations should be applied cautiously, as LLMs' ability to predict human behavior remains limited [163] and the sociocultural biases that their outputs exhibit are not fully understood [164]. These biases could lead to the perpetuation of existing generalizability problems [165,166] and worsen existing inequalities and biases of the literature[167].

Another potential usage of advanced analytical techniques is evidence synthesis across scientific disciplines and studies. Projects like the Human-Behavior-Change Project[157,158] or the Nudge Cartography Project [156] exemplify this trend. They aim to extract information from the effectiveness of behavioral interventions along hundreds of potential dimensions and leverage advanced machine learning techniques to answer what interventions work, compared with what, why and how well [157,158]. As the field of generative artificial intelligence is undergoing a rapid improvement, its current capabilities may not reflect its potential to support evidence accumulation—even in the near future.

Researchers should embrace the generalizability challenge of choice architecture interventions, which are characterized by small average effect sizes, large heterogeneity and limited ability to predict which of the outcomes will happen based on current knowledge. We hope that our Review helps usher in an era with more focus on generalizability at every step of the research process, increased large-scale collaborations, and a greater reliance on methodological and technological advancements. **[Au:OK? Ok]** Until good theories and models are developed that can predict the generalizability of interventions, there remains no strong substitute for testing interventions in new contexts before deploying large-scale interventions.

Display items

Figure captions

Figure 1. Dimensions of generalizability The figure lists the dimensions of the framework across which choice architecture researchers should explore the generalizability of mechanism. Each dimension of the framework (vertical text, left) can have an indefinitely large number of corresponding potential moderators (coloured boxes, left). The moderators listed for each dimension are intended as examples rather than a complete list. Interactions between moderators and change over time as moderators emerge and fade should also be considered.

Figure 2. Learning about generalizability across the research process Scientific practices enhance the accumulation of evidence on the generalizability of choice architecture interventions.

Box 1. Examples of different type of moderators in choice architecture interventions

These examples showcase instances in which the impact of specific moderators, corresponding to different conceptual dimensions, was investigated in large samples.

[H2] Unit

In a sequence of Randomized Controlled Trials (RCTs) with over 580,000 households, providing information about neighbors' consumption was associated with a 2% reduction in energy consumption[6]. Subsequently, an analysis of a scaled version of this intervention (111 randomized control trials with 8.6 million households across the United States), revealed that the original data overestimated the effect, likely attributable to varied population characteristics[7]. For instance, the households in the initial experiments were, on average, more environmentally friendly, wealthier, and possessed larger houses than those in the scaled intervention, which afforded more room for the observed decrease.

[H2] Treatment

Two megastudies (N = 47,206 and N = 689,693) tested text-based choice architecture interventions in promoting vaccination in the United States [168,169]. The specific implementations of the reminder messages such as the number of sent messages or the timing of the messages were varied across conditions. There was considerable variance in the vaccination rates across the different conditions based on the number and timing messages.

[H2] Outcome

A study tested the effects of financial education interventions on household financial decision making in India (N=1,328) found that the effect of financial education depended on the measured outcomes[101]. The treatment improved financial awareness and attitudes but was not found to improve longer term savings and borrowing.

[H2] Setting

A study tested the effects of a text alert after each instance of using a credit card on credit card overspending in South Korea [170]. In contrast to earlier studies where reminders about recent transactions were displayed prior to purchasing (on the same screen where the payment was made), this new intervention showed little to no reduction in spending for high spenders but led to an increase in spending among light and medium spenders. The authors hypothesized that the difference in results was due to the settings in which the information was provided; presented prior to purchasing versus presented separately on a mobile device (and also available later).

[H2] Analysis

At least five independent analysts were invited to reanalyze the original data for 100 studies (including 8 behavioral intervention studies) [108]. In 5 of the 8 behavioral intervention studies, all re-analysts reached the same conclusion as the original authors (finding a significant effect). However, in 3 studies, differences in analytical approaches led some analysts to draw conclusions that differed from those of the original authors.

References

1. Thaler, R. H. & Sunstein, C. R. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. (Yale University Press, New Haven, CT, 2008).
2. Michie, S., Van Stralen, M. M. & West, R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement. Sci.* **6**, 1–12 (2011).
3. Halpern, D. *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. (Random House, 2015).
4. Benartzi, S. *et al.* Should governments invest more in nudging? *Psychol. Sci.* **28**, 1041–1055 (2017).
5. Fishbane, A., Ouss, A. & Shah, A. K. Behavioral nudges reduce failure to appear for court. *Science* **370**, (2020).
6. Allcott, H. & Rogers, T. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *Am. Econ. Rev.* **104**, 3003–37 (2014).
7. Allcott, H. Site selection bias in program evaluation. *Q. J. Econ.* **130**, 1117–1165 (2015).
8. Hallsworth, M., List, J. A., Metcalfe, R. D. & Vlaev, I. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *J. Public Econ.* **148**, 14–31 (2017).
9. Frost, P. *et al.* The influence of confirmation bias on memory and source monitoring. *J. Gen. Psychol.* **142**, 238–252 (2015).
10. Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* **54**, 30–34 (1959).
11. Jachimowicz, J. M., Hauser, O. P., O'Brien, J. D., Sherman, E. & Galinsky, A. D. The critical role of second-order normative beliefs in predicting energy conservation. *Nat. Hum. Behav.* **2**, 757–764 (2018).
12. Hallsworth, M. A manifesto for applying behavioural science. *Nat. Hum. Behav.* **7**, 310–322 (2023).
13. Straßheim, H. The rise and spread of behavioral public policy: An opportunity for critical research and self-reflection. *Int. Rev. Public Policy* **2**, 115–128 (2020).
14. Cartwright, N. & Hardie, J. *Evidence-Based Policy: A Practical Guide to Doing It Better*. (Oxford University Press, 2012).
15. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, (2022).
16. Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A. & Aczel, B. A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work. *J. Behav. Decis. Mak.* **31**, 355–366 (2018).
17. 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).
18. Osman, M. *et al.* Learning from behavioural changes that fail. *Trends Cogn. Sci.* **24**, 969–980 (2020).
19. Milkman, K. L. *et al.* Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
20. Cartwright, N. Middle-range theory - (Teoría del rango medio: Without it what could anyone do? *Theor. Int. J. Theory Hist. Found. Sci.* **35**, 269–323 (2020).
21. Findley, M. G., Kikuta, K. & Denly, M. External Validity. *Annu. Rev. Polit. Sci.* **24**, 365–393 (2021).
22. Moeller, J. *et al.* Generalizability crisis meets heterogeneity revolution: Determining under which boundary conditions findings replicate and generalize. (2022).

23. Schloss, P. D. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio* **9**, 10.1128/mbio. 00525-18 (2018).

24. Whitaker, K. Publishing a reproducible paper. *Figshare Httpsdoi Org106084m9 Figshare* **5440621**, v2 (2017).

25. Pearl, J. & Bareinboim, E. External Validity: From Do-Calculus to Transportability Across Populations. in *Probabilistic and Causal Inference: The Works of Judea Pearl* vol. 36 451–482 (Association for Computing Machinery, New York, NY, USA, 2022).

26. Vazire, S., Schiavone, S. R. & Bottesini, J. G. Credibility Beyond Replicability: Improving the Four Validities in Psychological Science. *Curr. Dir. Psychol. Sci.* **31**, 162–168 (2022).

27. Lesko, C. R., Ackerman, B., Webster-Clark, M. & Edwards, J. K. Target Validity: Bringing Treatment of External Validity in Line with Internal Validity. *Curr. Epidemiol. Rep.* **7**, 117–124 (2020).

28. Marcellesi, A. External validity: Is there still a problem? *Philos. Sci.* **82**, 1308–1317 (2015).

29. Bareinboim, E. & Pearl, J. A General Algorithm for Deciding Transportability of Experimental Results. *J. Causal Inference* **1**, 107–134 (2013).

30. Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Sociol. Methods* 243–263 (1957).

31. Almaatouq, A. *et al.* Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* **47**, e33 (2024).

32. Hummel, D. & Maedche, A. How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *J. Behav. Exp. Econ.* **80**, 47–58 (2019).

33. DellaVigna, S. & Linos, E. RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* **90**, 81–116 (2022).

34. Mertens, S., Herberz, M., Hahnel, U. J. & Brosch, T. The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proc. Natl. Acad. Sci.* **119**, (2022).

35. Maier, M. *et al.* No evidence for nudging after adjusting for publication bias. *Proc. Natl. Acad. Sci.* **119**, e2200300119 (2022).

36. Szaszi, B. *et al.* No reason to expect large and consistent effects of nudge interventions. *Proc. Natl. Acad. Sci.* **119**, e2200732119 (2022).

37. Meehl, P. E. Theory-testing in psychology and physics: A methodological paradox. *Philos. Sci.* **34**, 103–115 (1967).

38. Gelman, A. Causality and Statistical Learning. *Am. J. Sociol.* **117**, 955–966 (2011).

39. Tosh, C., Greengard, P., Goodrich, B., Gelman, A. & Hsu, D. The piranha problem: Large effects swimming in a small pond. *Not. Am. Math. Soc.* (2025).

40. Bartoš, F. *et al.* Meta-analyses in psychology often overestimate evidence for and size of effects. *R. Soc. Open Sci.* **10**, 230224 (2023).

41. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. *Introduction to Meta-Analysis*. (John Wiley & Sons, 2011).

42. Rosenthal, R. & Gaito, J. Further evidence for the cliff effect in interpretation of levels of significance. *Psychol. Rep.* (1964).

43. Simonsohn, U., Simmons, J. & Nelson, L. Meaningless Means #2: The Average Effect of Nudging in Academic Publications is 8.7%. www.datacolada.ort/106 (2022).

44. Mertens, S., Herberz, M., Hahnel, U. J. J. & Brosch, T. Reply to Maier et al., Szaszi et al., and Bakdash and Marusich: The present and future of choice architecture research. *Proc. Natl. Acad. Sci.* **119**, e2202928119 (2022).

45. Bakdash, J. Z. & Marusich, L. R. Left-truncated effects and overestimated meta-analytic means. *Proc. Natl. Acad. Sci.* **119**, e2203616119 (2022).

46. Banerjee, A. & Urminsky, O. The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments. *Mark. Sci.* (2024) doi:10.1287/mksc.2021.0018.

47. Simonsohn, U. 'The' Effect Size Does Not Exist. https://datacolada.org/33 (2015).

48. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).

49. Parsons, S. *et al.* A community-sourced glossary of open scholarship terms. *Nat. Hum. Behav.* **6**, 312–318 (2022).

50. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl. Acad. Sci.* **115**, 2600–2606 (2018).

51. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* **7**, 632–638 (2012).

52. Schäfer, T. & Schwarz, M. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Front. Psychol.* **10**, 813 (2019).

53. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

54. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).

55. Camerer, C. F. *et al.* Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).

56. Errington, T. M. *et al.* Investigating the replicability of preclinical cancer biology. *Elife* **10**, e71601 (2021).

57. Szaszi, B., Venczel, F., Szecsi, P. & Borbala, G. A systematic review and meta-analysis of the preregistered choice architecture interventions. Preprint at https://osf.io/preprints/psyarxiv (2024).

58. Shu, L. L., Mazar, N., Gino, F., Ariely, D. & Bazerman, M. H. Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proc. Natl. Acad. Sci.* **109**, 15197–15200 (2012).

59. Stricker, J. & Günther, A. Scientific Misconduct in Psychology. *Z. Für Psychol.* **227**, 53–63 (2019).

60. Simmons, J., Nelson, L. & Simonsohn, U. Meaningless Means #1: The Average Effect of Nudging Is d = .43. https://datacolada.org/105 (2022).

61. Simonsohn, U., Simmons, J. & Nelson, L. D. Above averaging in literature reviews. *Nat. Rev. Psychol.* **1**, 551–552 (2022).

62. Atkins, L. *et al.* A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. *Implement. Sci.* **12**, 1–18 (2017).

63. Yang, S. *et al.* The elements of context. *Res. Rep. Ser. Behav. Inf. Organ. Partnersh. Behav. Econ. Action Rotman* (2023).

64. Mažar, N. & Soman, D. *Behavioral Science in the Wild*. (University of Toronto Press, 2022).

65. Ning, Z., Xin, L. I. U., Shu, L. I. & Rui, Z. Nudging effect of default options: A meta-analysis. *Adv. Psychol. Sci.* **30**, 1230 (2022).

66. Johnson, E. J. & Goldstein, D. G. Do defaults save lives? *Science* **302**, 1338–1339 (2003).

67. Krefeld-Schwalb, A., Sugerman, E. R. & Johnson, E. J. Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. *Proc. Natl. Acad. Sci.* **121**, e2306281121 (2024).

68. Jachimowicz, J. M., Duncan, S., Weber, E. U. & Johnson, E. J. When and why defaults influence decisions: A meta-analysis of default effects. *Behav. Public Policy* **3**, 159–186 (2019).

69. Narula, T., Ramprasad, C., Ruggs, E. N. & Hebl, M. R. Increasing colonoscopies? A psychological perspective on opting in versus opting out. *Health Psychol.* **33**, 1426 (2014).

70. Klein, R. A. *et al.* Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).

71. Moshontz, H. *et al.* The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Adv. Methods Pract. Psychol. Sci.* **1**, 501–515 (2018).

72. Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J. & Reinero, D. A. Contextual sensitivity in scientific reproducibility. *Proc. Natl. Acad. Sci.* **113**, 6454–6459 (2016).

73. Vonasch, A. J. *et al.* "Less Is Better" in Separate Evaluations Versus "More Is Better" in Joint Evaluations: Mostly Successful Close Replication and Extension of Hsee (1998). *Collabra Psychol.* **9**, (2023).

74. Imada, H. *et al.* Rewarding more is better for soliciting help, yet more so for cash than for goods: Revisiting and reframing the Tale of Two Markets with replications and extensions of Heyman and Ariely (2004). *Collabra Psychol.* **8**, 32572 (2022).

75. Ziano, I. *et al.* Numbing or Sensitization? Replications and Extensions of Fetherstonhaugh et al.(1997)'s "Insensitivity to the Value of Human Life". *J. Exp. Soc. Psychol.* **97**, 104222 (2021).

76. CORE Team. Collaborative Open-science and meta Research. (2025) doi:DOI 10.17605/OSF.IO/5Z4A8.

77. Hall, J. D. & Madsen, J. M. Can behavioral interventions be too salient? Evidence from traffic safety messages. *Science* **376**, eabm3427 (2022).

78. Singer, P. *Animal Liberation*. (Ecco, Harper Collins Publishers, New York, 1975).

79. Orth, T. The ethics of eating animals: Which factors influence Americans' views? *YouGov* https://today.yougov.com/health/articles/45577-ethics-eating-animals-which-factors-matter-poll (2023).

80. Delios, A. *et al.* Examining the generalizability of research findings from archival data. *Proc. Natl. Acad. Sci.* **119**, e2120377119 (2022).

81. Tipton, E. How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* **39**, 478–501 (2014).

82. Camerer, C. The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. *Available SSRN 1977749* (2011).

83. Pearl, J. Generalizing experimental findings. *J. Causal Inference* **3**, 259–266 (2015).

84. Henrich, J., Heine, S. J. & Norenzayan, A. Beyond WEIRD: Towards a broad-based behavioral science. *Behav. Brain Sci.* **33**, 111–135 (2010).

85. Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D. & Mortenson, E. Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspect. Psychol. Sci.* **15**, 1295–1309 (2020).

86. Mortensen, K. & Hughes, T. L. Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature. *J. Gen. Intern. Med.* **33**, 533–538 (2018).

87. Yeager, D. S., Krosnick, J. A., Visser, P. S., Holbrook, A. L. & Tahk, A. M. Moderation of classic social psychological effects by demographics in the US adult population: New opportunities for theoretical advancement. *J. Pers. Soc. Psychol.* **117**, e84 (2019).

88. Landy, J. F. *et al.* Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol. Bull.* **146**, 451 (2020).

89. Huber, C. *et al.* Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proc. Natl. Acad. Sci.* **120**, e2215572120 (2023).

90. Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *J. Verbal Learn. Verbal Behav.* **12**, 335–359 (1973).

91. Wells, G. L. & Windschitl, P. D. Stimulus sampling and social psychological experimentation. *Pers. Soc. Psychol. Bull.* **25**, 1115–1125 (1999).

92. Judd, C. M., Westfall, J. & Kenny, D. A. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *J. Pers. Soc. Psychol.* **103**, 54 (2012).

93. Capraro, V., Di Paolo, R., Perc, M. & Pizziol, V. Language-based game theory in the age of artificial intelligence. *J. R. Soc. Interface* **21**, 20230720 (2024).

94. Capraro, V., Halpern, J. Y. & Perc, M. From Outcome-Based to Language-Based Preferences. *J. Econ. Lit.* **62**, 115–154 (2024).

95. Goldstein, D. G. Leveling up applied behavioral economics. *Behav. Econ. Guide Introd. Dan Goldstein Pp VI–XI* (2022).

96. Lejarraga, T. & Hertwig, R. How experimental methods shaped views on human competence and rationality. *Psychol. Bull.* **147**, 535 (2021).

97. Doerrenberg, P. & Schmitz, J. *Tax Compliance and Information Provision: A Field Experiment with Small Firms*. https://www.econstor.eu/handle/10419/110751 (2015).

98. Carey, R. N. *et al.* Describing the 'how'of behaviour change interventions: a taxonomy of modes of delivery. in *UK Society for Behavioural Medicine Conference* 1–2 (2016).

99. Webb, T. L. & Sheeran, P. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* **132**, 249 (2006).

100. Osman, M. Backfiring, reactance, boomerang, spillovers, and rebound effects: Can we learn anything from examples where nudges do the opposite of what they intended? (2020).

101. Carpena, F., Cole, S., Shapiro, J. & Zia, B. The ABCs of Financial Education: Experimental Evidence on Attitudes, Behavior, and Cognitive Biases. *Manag. Sci.* **65**, 346–369 (2019).

102. Aczel, B. *et al.* Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* **10**, e72185 (2021).

103. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).

104. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208–1214 (2020).

105. Szaszi, B. *et al.* Does alleviating poverty increase cognitive performance? Short- and long-term evidence from a randomized controlled trial. *Cortex* **169**, 81–94 (2023).

106. Botvinik-Nezer, R. *et al. Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams*. 843193 https://www.biorxiv.org/content/10.1101/843193v1 (2019) doi:10.1101/843193.

107. Hoogeveen, S. *et al.* A many-analysts approach to the relation between religiosity and well-being. *Relig. Brain Behav.* 1–47 (2022) doi:10.1080/2153599X.2022.2070255.

108. Aczel, B. *et al.* Investigating the analytical robustness of the social and behavioural sciences. *Manuscr. Submitt. Publ.* (2024).

109. Wagenmakers, E.-J., Sarafoglou, A. & Aczel, B. One statistical analysis must not rule them all. *Nature* **605**, 423–425 (2022).

110. Trafimow, D. A New Way to Think About Internal and External Validity. *Perspect. Psychol. Sci.* **18**, 1028–1046 (2023).

111. Trafimow, D. Generalizing across auxiliary, statistical, and inferential assumptions. *J. Theory Soc. Behav.* **52**, 37–48 (2022).

112. Michie, S. & Johnston, M. Theories and techniques of behaviour change: Developing a cumulative science of behaviour change. *Health Psychol. Rev.* **6**, 1–6 (2012).

113. IJzerman, H. *et al.* Use caution when applying behavioural science to policy. *Nat. Hum. Behav.* **4**, 1092–1094 (2020).

114. Davis, R., Campbell, R., Hildon, Z., Hobbs, L. & Michie, S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychol. Rev.* **9**, 323–344 (2015).
115. Imai, K., Keele, L., Tingley, D. & Yamamoto, T. Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies. *Am. Polit. Sci. Rev.* **105**, 765–789 (2011).
116. Weller, N. & Barnes, J. *Finding Pathways: Mixed-Method Research for Studying Causal Mechanisms*. (Cambridge University Press, 2014).
117. Goertz, G. *Multimethod Research, Causal Mechanisms, and Case Studies: An Integrated Approach*. (Princeton University Press, 2017).
118. Sartori, G. Concept Misformation in Comparative Politics. *Am. Polit. Sci. Rev.* **64**, 1033–1053 (1970).
119. Martel Garcia, F. & Wantchekon, L. Theory, External Validity, and Experimental Inference: Some Conjectures. *Ann. Am. Acad. Pol. Soc. Sci.* **628**, 132–147 (2010).
120. Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. (Basic books, 2018).
121. Dinner, I., Johnson, E. J., Goldstein, D. G. & Liu, K. Partitioning default effects: why people choose not to choose. *J. Exp. Psychol. Appl.* **17**, 332 (2011).
122. Hajdu, N., Szaszi, B. & Aczel, B. Extending the Choice Architecture Toolbox: The Choice Context Mapping. *Sage Open* (2024).
123. Gallagher, R., Gyani, A., Tan, C. & Tindall, K. Explore: Four simple ways to map and unpack behaviour. **Handbook of the Behvioral Insights Team**, (2022).
124. Cronbach, L. J. & Shapiro, K. *Designing Evaluations of Educational and Social Programs*. (Jossey-Bass, 1982).
125. Holzmeister, F. *et al.* Heterogeneity in effect size estimates. *Proc. Natl. Acad. Sci.* **121**, e2403490121 (2024).
126. Rosenbaum, P. R. & Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **79**, 516–524 (1984).
127. Tipton, E. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38**, 239–266 (2013).
128. Sedgwick, P. Convenience sampling. *Bmj* **347**, (2013).
129. Lohr, S. L. *Sampling: Design and Analysis*. (Chapman and Hall/CRC, 2021).
130. Campbell, S. *et al.* Purposive sampling: complex or simple? Research case examples. *J. Res. Nurs.* **25**, 652–661 (2020).
131. Punch, K. *Developing Effective Research Proposals*. (Sage, 2000).
132. Yeager, D. S. *et al.* A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).
133. Czibor, E., Jimenez-Gomez, D. & List, J. A. The dozen things experimental economists should do (more of). *South. Econ. J.* **86**, 371–432 (2019).
134. Vlasceanu, M., Doell, K., Bak-Coleman, J. & Van Bavel, J. J. Addressing Climate Change with Behavioral Science: A Global Intervention Tournament in 63 Countries. (2023).
135. Većkalov, B. *et al.* A 27-country test of communicating the scientific consensus on climate change. (2023).
136. West, R. *et al.* Using machine learning to extract information and predict outcomes from reports of randomised trials of smoking cessation interventions in the Human Behaviour-Change Project. *Wellcome Open Res.* **8**, 452 (2024).
137. Baribault, B. *et al.* Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci.* **115**, 2607–2612 (2018).

138. Bahník, Š. & Vranka, M. A. If it’s difficult to pronounce, it might not be risky: The effect of fluency on judgment of risk does not generalize to new stimuli. *Psychol. Sci.* **28**, 427–436 (2017).

139. Awad, E. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018).

140. Michie, S. *et al.* Representation of behaviour change interventions and their evaluation: Development of the Upper Level of the Behaviour Change Intervention Ontology. *Wellcome Open Res.* **5**, 123 (2021).

141. Szaszi, B., Harry Clelland, Folk, D., ... & Dunn, E. Global Happiness Megastudy. *manuscript* (2025).

142. Menkveld, A. J. *et al.* Nonstandard Errors. *J. Finance* **79**, 2339–2390 (2024).

143. Silberzahn, R. *et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).

144. Szaszi, B. & Venczel, F. The analytical robustness of behavioral interventions. *Manuscript* (2024).

145. Clarke, B., Schiavone, S. & Vazire, S. What limitations are reported in short articles in social and personality psychology? *J. Pers. Soc. Psychol.* (2023).

146. Simons, D. J., Shoda, Y. & Lindsay, D. S. Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect. Psychol. Sci.* **12**, 1123–1128 (2017).

147. Maier, M. *et al.* Exploring Open Science Practices in Behavioural Public Policy Research. (2023).

148. Wang, K. *et al.* A multi-country test of brief reappraisal interventions on emotions during the COVID-19 pandemic. *Nat. Hum. Behav.* **5**, 1089–1110 (2021).

149. Cologna, V., Niels, M. & Trust in Science Collaboration. Trust in scientists and their role in society: a global assessment in 66 countries. *Nat. Hum. Behav.* (2024).

150. Dai, H. *et al.* Behavioural nudges increase COVID-19 vaccinations. *Nature* **597**, 404–409 (2021).

151. Rabb, N. *et al.* Evidence from a statewide vaccination RCT shows the limits of nudges. *Nature* **604**, E1–E7 (2022).

152. Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).

153. Milkman, K. L., Beshears, J., Choi, J. J., Laibson, D. & Madrian, B. C. Using implementation intentions prompts to enhance influenza vaccination rates. *Proc. Natl. Acad. Sci.* **108**, 10415–10420 (2011).

154. Mischel, W. The Toothbrush Problem. *APS Obs.* **21**, (2008).

155. Frith, U. Fast Lane to Slow Science. *Trends Cogn. Sci.* **24**, 1–2 (2020).

156. Linnea, G. Nudge cartography: building a map to navigate behaivoral research. (2023).

157. Michie, S. *et al.* The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implement. Sci.* **12**, 1–12 (2017).

158. Michie, S. *et al.* The Human Behaviour-Change Project: An artificial intelligence system to answer questions about changing behaviour. *Wellcome Open Res.* **5**, (2020).

159. Grossmann, I. *et al.* AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).

160. Hämäläinen, P., Tavast, M. & Kunnari, A. Evaluating large language models in generating synthetic hci research data: a case study. in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* 1–19 (2023).

161. Batista, R. M. & Ross, J. Words that Work: Using Language to Generate Hypotheses. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4926398 (2024).

162. Manning, B. S., Zhu, K. & Horton, J. J. Automated Social Science: Language Models as Scientist and Subjects. Preprint at https://doi.org/10.48550/arXiv.2404.11794 (2024).

163. Capraro, V., Paolo, R. D. & Pizziol, V. Assessing Large Language Models' ability to predict how humans balance self-interest and the interest of others. Preprint at https://doi.org/10.48550/arXiv.2307.12776 (2024).

164. Arzaghi, M., Carichon, F. & Farnadi, G. Understanding Intrinsic Socioeconomic Biases in Large Language Models. in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* vol. 7 49–60 (2024).

165. Atari, M., Xue, M. J., Park, P. S., Blasi, D. & Henrich, J. Which humans? (2023).

166. Peng, C. *et al.* A Study of Generative Large Language Model for Medical Research and Healthcare. *ArXiv Prepr. ArXiv230513523* (2023).

167. Capraro, V. *et al.* The impact of generative artificial intelligence on socioeconomic inequalities and policy making. Preprint at https://doi.org/10.48550/arXiv.2401.05377 (2024).

168. Milkman, K. L. *et al.* A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proc. Natl. Acad. Sci.* **118**, e2101165118 (2021).

169. Milkman, K. L. *et al.* A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proc. Natl. Acad. Sci.* **119**, e2115126119 (2022).

170. Kim, J., Yoon, Y., Choi, J., Dong, H. & Soman, D. Surprising Consequences of Innocuous Mobile Transaction Reminders of Credit Card Use. *J. Interact. Mark.* 10949968231189505 (2023).

Highlighted references

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A. & Aczel, B. A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work. *J. Behav. Decis. Mak.* 31, 355–366 (2018).

A comprehensive systematic review examining the choice architecture movement, summarizing how to improve the evidence synthesis of nudge interventions.

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).

A critical review arguing that behavioral science can only create meaningful change by rigorously addressing heterogeneity in research findings.

Findley, M. G., Kikuta, K. & Denly, M. External Validity. *Annu. Rev. Polit. Sci.* 24, 365–393 (2021).

A thorough review of external validity in research, providing a framework for understanding how research findings can be applied beyond specific study contexts.

Almaatouq, A. *et al.* Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* 47, e33 (2024).

A review study proposing a new, more integrative experimental design in social and behavioral sciences, moving beyond simplistic hypothesis testing.

DellaVigna, S. & Linos, E. RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* 90, 81–116 (2022).

A comprehensive analysis of randomized controlled trials from two nudge units, providing large-scale evidence about intervention effectiveness.

Landy, J. F. *et al.* Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol. Bull.* 146, 451 (2020).

A methodological study demonstrating how research design choices fundamentally shape research results.

Silberzahn, R. *et al*. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv. Methods Pract. Psychol. Sci.* 1, 337–356 (2018).

A landmark study showing how variations in analytical approaches can dramatically alter conclusions when applied to the same dataset.

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

The study shows how variations in population sampling, study design, and analytical approaches create substantial heterogeneity that significantly reduces the generalizability of scientific findings.

Acknowledgements

ChatGPT was used to check grammar and improve the language of the manuscript. The work of Barnabas Szaszi was supported by the Eötvös Loránd University Excellence Fund (EKA) by the unwavering support and patience of Melinda Szrenka.

Author contributions

All authors contributed substantially to discussion of the content. B.S. wrote the first draft of the manuscript, while D.G.G., D.S., and S.M. reviewed, edited, and provided critical revisions.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Psychology thanks [Referee#1 name], [Referee#2 name] and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note