Received: 19 October 2022 | Revised: 6 May 2023 | Accepted: 22 May 2023

DOI: 10.1111/padm.12941

ORIGINAL ARTICLE

public administration WILEY

Seeing the nudge from the trees: The 4S framework for evaluating nudges

Stuart Mills | **Richard Whittle**

Department of Economics, University of Leeds, Leeds, UK

Correspondence
Stuart Mills, Department of Economics, University of Leeds, Leeds, UK.
Email: s.mills1@leeds.ac.uk

Abstract

Nudging is a popular and influential approach in policymaking. Yet, it has faced substantial criticism from several policy perspectives, with growing concern raised about the efficacy of some nudge interventions. This article offers an evaluative framework for nudging which captures these various perspectives. Our 4S framework highlights the importance of nudges being *sufficient*, *scalable*, and *subjective*, in addition to being *statistically significant*, to be an effective policy response. We review various nudge interventions, coupled with various methodological critiques, to demonstrate the need for a more expansive evaluative framework. The 4S framework synthesizes these sizeable literatures and numerous critiques to meet this need, serving as an important contribution to behavioral policymakers. We argue that the 4S framework complements existing frameworks for designing behavioral interventions as an evaluative framework. By adopting the 4S framework, policymakers will be better placed to design interventions which are effective in relation to the wider policy environment.

1 | INTRODUCTION

In a 2019 paper, published in *Nature Communications*, Claudia Nisa and colleagues undertook a meta-analysis of "randomised controlled trials testing behavioural interventions to promote household action on climate change"

(Nisa et al., 2019, p. 1). They found that the average effect size of these interventions, measured in terms of Cohen's *d*, was 0.093, *half* the typical "small" effect size for Cohen's *d* of $d = 0.2$ (Cohen, 1988). They concluded that, "the available field experiments do not give rise to hopeful predictions about relying on behavioural interventions alone to tackle climate change," (Nisa et al., 2019, p. 8).

This conclusion would prompt an exchange of perspectives within the journal. Responding to Nisa et al. (2019), van der Linden and Goldberg (2020) expressed concern regarding the meta-analytical approach originally undertaken. In replicating the result, van der Linden and Goldberg (2020) argued the true average effect size may be, "at least twice as large as initially reported" (p. 1), concluding that, "although the average effects of behavioural interventions are still not particularly large in absolute terms ($d = 0.20$), they are not alarmingly small and could be consequential when scaled at population level" (p. 2).

This elicited a response from Nisa et al. (2020). While they critiqued van der Linden and Goldberg's (2020) meta-analytical replication on statistical grounds, Nisa et al. (2020) also argued that even *accepting* van der Linden and Goldberg's (2020) higher estimated effect size should not change the conclusion of the paper: "our conclusions hold because most overall effect sizes produced by different estimators are still interpretable as very small or, at best, borderline small" (Nisa et al., 2020, p. 2). They also refuted the claim that even a small effect-sized intervention could be impactful when scaled up because, "[as] interventions [are] scaled up to the general population… these will target a more heterogeneous set of individuals… This suggests that effect sizes would likely approach our estimates for naïve subjects" (p. 1). Finally, Nisa et al. (2020, p. 2, emphasis added) reaffirmed their position:

> "The U.N. declared the next 10 years as the Decade of Action for climate change… motivated by the awareness that action is not advancing at the speed or scale required, and calling for interventions to step-up their impact. Stating that effect-sizes in psychology are known to be small should not be used as a justification to inflate the meaningfulness of (very) small effects. *A thoughtful debate beyond statistical significance is long overdue to make psychological and behavioural science more relevant to intervention and policymaking*."

This exchange illustrates an important question in policymaking: *what makes a policy effective?* For van der Linden and Goldberg (2020), *statistical significance* seems to have been the benchmark for effectiveness. While the Cohen's *d* estimate remained small, this seems only to have influenced how van der Linden and Goldberg (2020) advocated these interventions to be *used*—through scaling—rather than whether these interventions *should be used*. For Nisa et al. (2020), statistical significance does not appear to have been sufficient to regard these interventions as *effective*. Set against a backdrop where interventions must, "step-up their impact," Nisa et al. (2020) seem to reject the idea that "effectiveness" derives from a *p*-value alone, and instead suggest this should be only one of several factors governing whether an intervention is effective.

Behavioral public policy, just as policymaking in general, should be concerned with the question: what makes a policy effective? This is especially important as the most popular and famous behavioral policy approach, nudging, rejects alternative approaches such as mandates and bans, while *simultaneously* advancing a thesis which suggests people may not always choose what is best for themselves (Sunstein, 2014; Sunstein et al., 2017; Thaler & Sunstein, 2003, 2008). This creates a delicate policy environment where it is difficult to know if a nudge has been effective in terms of, say, compliance or welfare (Tor, 2020). Furthermore, nudge policies—originating from psychological experiments usually based in laboratories and employing randomized controlled trials (RCTs)—is perhaps peculiar within the policy space insofar as its means of gathering policy evidence, and its intellectual outlook, has tended to focus on individual behavior (Chater & Loewenstein, 2022; Ewert, 2020). This has spurred recent commentary criticizing nudging as a policy tool given significant policy challenges (Chater & Loewenstein, 2022; Ewert, 2020; Harford, 2022; for a counter-perspective, see Hallsworth, 2022b), such as climate change (Maki, 2019; Nisa et al., 2020) and criminal justice reform (Kohler-Hausmann, 2020). Finally, the political success of programs such as "nudge," driven to a large degree by the apparent cost-effectiveness of nudging (de Ridder et al., 2022;

Eaglesham, 2008; List et al., 2022), has led some to hypothesize that these interventions may "crowd out," better policy alternatives (Hagman et al., 2019; Loewenstein & Chater, 2017; Maki, 2019) by steering political attention and willpower away from those alternatives (Lades & Delaney, 2022; again, a counter-perspective is provided by Hallsworth, 2022b).

This article takes up the challenge established by Nisa et al. (2020) to go "beyond statistical significance" when evaluating behavioral interventions, and nudges specifically. We are not the first to interrogate nudging from perspectives other than statistical significance. For instance, several authors (Entwistle, 2021; John et al., 2022; Sunstein, 2017; Tor, 2020) have dissected the policy suitability of nudging from the perspectives of *why* nudges fail? There is also a growing body of work incorporating welfare evaluations into nudges when reflecting upon effectiveness (Brown et al., 2022; Bulte et al., 2020; Laffan et al., 2021; List et al., 2022; Thunström, 2019; Thunström et al., 2018), as well as population heterogeneity (Mills, 2022; Sunstein, 2022). Furthermore, as behavioral public policy has matured and the "low hanging fruit" has been picked, there has been increasing interest in scaling interventions to achieve greater impact (Al-Ubaydli et al., 2017, 2021; Castleman, 2021; Sanders et al., 2018; van der Linden & Goldberg, 2020). Finally, and beyond behavioral policy *specifically*, astute criticisms of RCTs—which have been utilized extensively in developmental economics as well as behavioral economics (e.g., Banerjee, Chassang, Montero, et al., 2017; Banerjee, Chassang, & Snowberg, 2017; Banerjee & Duflo, 2009)—have been put forth to encourage researchers and policymakers to reflect on the usefulness of these methods in terms of policymaking (Deaton & Cartwright, 2018; Jamal et al., 2015).

We focus specifically on four elements of policy evaluation which are relevant to nudging, which we present in our "4S" framework. These elements are: (statistical) *significance*; *sufficiency*; *scalability*; and *subjectivity*. Through the 4S framework, we demonstrate how each element should contribute to the "effectiveness" of a nudge, and how effectiveness is often better understood as a trade-off between elements, rather than an exercise in satisfying each.

We recognize that this framework may be applicable beyond nudges, such as to the wider behavioral policy literature and to the literature on (experimental) evidence-based policymaking. Indeed, several components of the 4S framework draw on these literatures. For instance, our discussion of *significance* builds from various critiques of RCTs in experimental economics (e.g., Deaton & Cartwright, 2018), which themselves are aligned with various criticisms found in the policy evaluation literature (e.g., Pawson & Tilly, 1994, 1997). Another example is *scalability*, with several criticisms of scaling nudges also found—and frequently discussed—in the education policy literature (e.g., Castleman, 2021).

We focus on nudging in particular for two reasons. *First*, nudging has become increasingly prominent in policymaking over the past decade (Della Vigna & Linos, 2022; de Ridder et al., 2022; Sanders et al., 2018). The OECD (2018) reports that over 200 "nudge" units had been established as of 2018. Yet, current debates have begun to challenge the effectiveness of nudging (Chater & Loewenstein, 2022; Hallsworth, 2022a, 2022b; Maier et al., 2022). Given the growing prominence of nudging as a policy tool, but also rising concerns about the effectiveness of nudges *as* policy tools, an evaluative framework is of timely importance. The 4S framework is a contribution to this literature.

Second, because focusing on nudges demonstrates a specific use-case for the framework, which should serve as a model for using the framework in other areas of behavioral and experimental policymaking. There is certainly wider opportunity for using the 4S framework. For instance, many debates within developmental economics are juxtaposed between smaller interventions at individual and community levels built from experimental evidence, and more substantive economic investment and macroeconomic policy, typically at the state and international level (Banerjee & Duflo, 2010; Duflo, 2011; Duflo & Kremer, 2008). Challenges in this space clearly touch on elements of the 4S framework, specifically, *sufficiency* and *scalability*.

The structure of this article is as follows. *First*, we introduce the 4S framework and briefly outline how it fits in the existing "behavioral framework" infrastructure. *Second*, we discuss each element in more detail. As statistical significance is often the most prominent barometer of "effectiveness," we begin by discussing this element of the framework to front-run various critiques which are found in the remaining three elements. Some threads established

in the discussion of *significance* are thus further elaborated in the remaining section discussions. *Third*, we provide guidance on using the 4S framework, noting the importance of trade-offs when evaluating the effectiveness of a nudge, before concluding.

2 | THE 4S FRAMEWORK: AN OVERVIEW

The 4S framework consists of four elements of policy evaluation, summarized in Table 1.

Significance concerns whether interventions work *statistically* speaking (i.e., not due to random chance). *Sufficiency* concerns the ultimate objective of the policy, or the policy challenge which any potential policy is designed to solve. *Scalability* concerns the applicability of an intervention when it is applied to larger and more diverse populations than it was tested in. Finally, *subjectivity* concerns the welfare effects of interventions, as well as questions which arise when one considers heterogeneous populations and distributional effects.

We situate our framework within the "behavioral framework" infrastructure which already exists (see Figure 2), such as MINDSPACE (Dolan et al., 2012), COM-B (Michie et al., 2011), EAST (Service et al., 2015), and FORGOOD (Lades & Delaney, 2022). The 4S framework is distinct from these frameworks, as it is an *evaluative* framework, rather than a *design* framework. For instance, EAST argues effective interventions are typically *Easy*, *Attractive*, *Social*, and *Timely*. Yet, these are all *design* features, and offer little critique of *what effective means*? The four Ss of the 4S framework complement these design frameworks as part of a wider process of *discovering*, *applying*, and *evaluating* (e.g., Hallsworth & Kirkman, 2020; Haynes et al., 2012; Ruggeri, 2021; Sunstein, 2020). These frameworks do not all perform the same function. For instance, MINDSPACE is a broad behavioral intervention design framework, while COM-B is more grounded in understanding and applying psychological processes, and FORGOOD is an ethical

TABLE 1 4S framework overview.

Element	Motivating question(s)	Example
(Statistical) Significance	Is the effect of an intervention due to random chance, and if not, is there a causal effect between the intervention and the observed behavior?	A randomized controlled trial (RCT) where a treatment group is asked to choose between options with a default option selected, and a control group is asked to choose without any default being selected (i.e., an active choice).
Sufficiency	Is the intervention and/or policy program in which the intervention is used adequate to resolve whatever policy challenge prompted the need for a policy response?	Using a nudge to encourage vaccine uptake in response to a sudden viral epidemic, where herd immunity requires x% uptake, and current uptake is y%, where x > y.
Scalability	Does the causal effect of an intervention and/or policy program in which the intervention is used broadly hold as the intervention is applied to larger and more heterogeneous populations than were used to identify the effect?	An automatic reminder text message program to support low-income high school students into higher education, which is supported by a one-on-one mentoring service if a student engages with the nudge.
Subjectivity	How does the intervention and/or policy program in which the intervention is used interact with different sub-groups in the population, and how are the effects of the intervention felt by different sub-groups?	A calorie label nudge on snack food designed to encourage healthy eating, which is experienced differently by health-conscious and non-health-conscious groups in terms of (a) observed behaviors; and (b) welfare outcomes.

framework which likely has applicability in both the design of nudges and their evaluation. We elaborate more on this "infrastructure," and how the 4S framework contributes to it, in our discussion section.

3 | (STATISTICAL) SIGNIFICANCE

Statistical significance is often regarded as the minimum standard for an effective intervention. While several statistical approaches have been adopted in the literature, a popular approach is the randomized control trial (RCT; Della Vigna & Linos, 2022). The UK Cabinet Office has described RCTs as, "the best way of determining whether a policy is working" (Haynes et al., 2012, p. 4), and advocated RCTs be used throughout government. Given such praise, RCTs are often considered the "Gold Standard," of policy research. They are also a useful case study in this article, being both a popular approach to testing behavioral interventions (and in other areas, such as developmental economics; Banerjee & Duflo, 2009; Duflo & Kremer, 2008), and a relatively simple approach for discursive purposes.

Deaton and Cartwright (2018) offer an authoritative critique of RCTs. Their discussion focuses on two points which are relevant to this article. *First*, they argue that researchers and policymakers often over emphasize the advantages of randomization, which may lead to imprecise estimates of average treatment effects, or the erroneous attribution of causality to a treatment, rather than an unobserved factor. This is known as "random confounding" (Deaton & Cartwright, 2018). The advantage of randomization is that one can, *theoretically*, control for unobserved factors by assuming equal distribution of these factors across the control and the treatment group(s) post-randomization (Banerjee, Chassang, Montero, et al., 2017; Banerjee, Chassang, & Snowberg, 2017). However, such an assumption can rarely be confirmed or rejected (Dawid, 2000). Deaton and Cartwright (2018) note that attempts are often made to validate this assumption through comparison of the means of *observed* factors across control and treatment groups. Yet, they also argue that—owing to the many *unobservable* factors—such an approach cannot confirm this assumption of "balanced factors." As such, "the causality that is being attributed to the treatment might, in fact, be coming from an imbalance in some other cause in our particular trial" (Deaton & Cartwright, 2018, p. 5). Jamal et al. (2015, p. 1) call this the "black box" problem of RCTs.

Second, even accepting the balance of the experiment; this does not necessarily mean the result is applicable *beyond* the RCT setting. In the first instance, the sample used may not be reflective of the population. This, Deaton and Cartwright (2018) note, is a well-traced criticism relating to *external validity*, but is still important to recognize when reflecting upon areas such as *scalability* (Al-Ubaydli et al., 2017, 2021). More crucially, however, is *how* an RCT result is used, and understood to be *useable*. Deaton and Cartwright (2018) note that even an "internally valid" RCT (i.e., that which does not suffer from random confounding and is sufficiently powerful) does not produce results which will necessarily generalize as contexts change. They (p. 3) succinctly write, "extrapolating or generalizing RCT results requires a great deal of additional information that cannot come from RCTs… credibility in estimation can lead to incredibility in use." This is a view shared by al-Ubaydli et al. (2021, p. 4): "[T]he question of how to actually use… experimental insights for policymaking remains poorly understood." Therefore, it is unwise to conclude that a statistically significant RCT is the "true effect," and thus that statistical significance is sufficient to determine that a policy "works," beyond the RCT setting. Earlier contributions in the evaluation literature also point to this criticism. As Pawson and Tilley (1994) argue, statistical relationships which are not evaluated in relation to their context are liable to lead to policy which is ineffective, or which results in unexpected outcomes.

van der Linden and Goldberg's (2020) assertion that small effect-sized interventions may simply need to be scaled (or "extrapolated") may be evidence of the perspective which Deaton and Cartwright (2018) criticize, while this same argument—that RCT results, and statistically significant results more generally (Deaton & Cartwright, 2018; Pawson & Tilley, 1994) do not cleanly map onto the "real-world"—can be seen in the retort given by Nisa et al. (2020). Disparities in academic and applied RCT nudge results reported by Della Vigna and Linos (2022) may also be reasonably explained by this criticism of RCTs (also see our discussion of *scalability*), as might diverges found by

Dai et al. (2021) between nudges to increase vaccination intention implemented as part of an RCT, and those implemented in the real-world.

Deaton and Cartwright (2018) offer two examples of where difficulties in mapping RCT results come from. *First*, they (Deaton & Cartwright, 2018, p. 11, emphasis added) quote Drèze, who states, "when a foreign agency comes in with its heavy boots and deep pockets to administer a "treatment"… there tends to be a lot going on *other than the treatment*." *Second*, they note that behaviors may change "because of the presence of the 'treators'" (p. 11). Both examples resonate with arguments relating to the importance of behavioral spillovers (Al-Ubaydli et al., 2021; Banerjee, Banerji, et al., 2017; Dolan & Galizzi, 2015; Maki et al., 2019) and spillunders (Krpan et al., 2019).

Broadly, these spillovers and spillunders describe unanticipated behaviors which occur in conjunction with the behaviors targeted by an intervention. For instance, Drèze describes a *spillunder* effect: prior to receiving any treatment, the presence of infrastructure to give the treatment creates anticipatory effects. These may cause the treated to behave differently to participants in the RCT, thus leading to different results.

Typical examples of *spillover* effects include licensing effects (Mazar & Zhong, 2010; Merritt et al., 2010). For instance, an intervention to encourage healthy behaviors may also give participants a "license," to later—and outside the gaze of the researcher—over-indulge in unhealthy behaviors as a "reward" for their previous "good" behavior (Chiou et al., 2011). If such spillovers are not seen by the experimenter, the effect the experimenter observes may be substantially larger than the true effect post-experiment.

In their review of 174 nudge studies, Beshears and Kosowsky (2021) find only 12 studies measured additional outcome behaviors. Bryan et al. (2021) report comparable figures for behavioral science interventions more broadly. Without adjustment to an experimental design to investigate such spillover behaviors, an RCT (and other experimental approaches) *cannot* adjust the estimated effect size to account for these behaviors. A result may therefore be statistically significant *in isolation*, but may be *less significant*, or *insignificant*, when applied in a different context or setting (Al-Ubaydli et al., 2021; also see Maki et al., 2019).

de Ridder et al. (2022) also express concern at the lack of investigation of spilling effects, contextual factors, and the mechanisms behind nudges more generally (for a perspective on RCTs more broadly, see Jamal et al., 2015; van Belle et al., 2016). Likewise, Banerjee and John (2023) argue the broad lack of understanding concerning the mechanisms behind nudge effects contributes to current debates about nudge effectiveness, and represents an important future challenge (also see Bryan et al., 2021; Chapman et al., 2023; Hecht et al., 2022). From the perspective of Pawson and Tilly (1994, 1997), writing in the evaluation literature, without an understanding of the mechanisms behind a policy, isolated statistical results are inadequate to evaluate the effectiveness of a policy intervention (also see Jamal et al., 2015). Indeed, an apt quote from Pawson and Tilly (1994, p. 292, original emphasis) links this criticism back to criticisms of RCTs and other experimental methods more broadly: "the quest for control and certainty which is the *raison d'être* of the experimental approach is, in fact, the very factor which obliges that method to overlook the importance of those mechanisms and contexts which constitute the programs under investigation."

Despite these criticisms, statistical significance remains a crucial element of nudge design, and where statistical significance is sought, experimental designs such as RCTs may prove desirable (Banerjee & Duflo, 2009). Yet, there is an additional perspective on statistical significance, which might be regarded as a *political* perspective. While RCTs have done much for advancing notions of "evidence-based," and "data-driven," policymaking, and contributed to what has been called the "what works," literature (Deaton & Cartwright, 2018; Jamal et al., 2015) in areas such as developmental economics (e.g., Duflo & Kremer, 2008; for a broader perspective on "evidence-based" economics, see Brice & Montesinos-Yufa, 2019; Hamermesh, 2013), these notions may ultimately come to undermine policymaking when policies which cannot be "evidence-based," (either by design, or through *restriction*) must still be considered (Chater & Loewenstein, 2022). Deaton and Cartwright (2018) express this concern to a degree insofar as they worry the "special status" of RCTs (p. 2) in policymaking frequently leads alternative evidence to be marginalized or ignored. The historic shift from "theory-based economics" to "evidence-based economics," driven by the popularization of experimental and quasi-experimental techniques in the field since 2000 (Brice & Montesinos-Yufa, 2019;

Hamermesh, 2013), maybe indicative of such concerns (Deaton & Cartwright, 2018; de Ridder et al., 2022; Jamal et al., 2015; Pawson & Tilley, 1994).

Yet, our argument here also draws on arguments found within philosophy of science, particularly those of Feyerabend (2010) and Kuhn (2012), that what knowledge (e.g., evidence) exists, and *is allowed to exist*, is subject to a suite of political, economic, and social factors (on economic constraints, see Al-Ubaydli et al., 2017). Furthermore, the research question or policy area to be investigated may not lend itself to be "evidence-based" in the sense of, say, an RCT (Chater & Loewenstein, 2022). For instance, macroeconomic policy cannot meet this "Gold Standard" of evidence because there *cannot* be a control group. Thus, while evidence is frequently sought in terms of macroeconomic data; conjecture, opinion, debate, and historical interpretation all factor into macroeconomic policymaking.

Prioritizing statistical significance, or the *mere presence* of positive evidence more generally, can ignore the subtle critique that said evidence and significance only exists because of fortuitous circumstances. As such, unevidenced, or *under*-evidenced policy, may *still* be worth considering because *a lack of evidence does not necessarily mean an ineffective policy*. Thus, the importance of statistical significance is not *diminished*, but the potential importance of *additional* considerations is enhanced. It is to these additional considerations which we now turn.

4 | SUFFICIENCY

For a behavioral intervention to be *sufficient* within the 4S framework, it must achieve a pre-specified policy objective. Sufficiency begins with the question, "what outcome would lead us to conclude that this policy challenge is now resolved?" and the 4S framework advocates evaluating the effectiveness of a nudge based on whether the nudge has led to resolution of that policy challenge.

We are not the first to highlight the role of sufficiency in nudging. In their respective discussions of nudges that fail, both Sunstein (2017, p. 7) and Tor (2020, p. 317) write of "inadequate" nudges. Lades and Delaney (2022) also suggest that policymakers should always consider alternative policies when determining whether a nudge is *good* (i.e., ethical). In their critique of individual-level policy, Chater and Loewenstein (2022) reflect on several behavioral policy failings which could be described as critiques of sufficiency. These perspectives are valuable, but do not wholly align with our perspective on sufficiency. For instance, Sunstein et al. (2017) and Tor (2020) define inadequacy as occurring when a policy objective is best achieved by an alternative policy approach, such as a mandate or ban. Likewise, Lades and Delaney's (2022) framing of alternative options could lead one to view inadequacy as a false choice between *only nudging* or *doing something else*. Such a perspective would itself be an inadequate account of *insufficiency*, for nudges and behavioral policy more generally are often used in conjunction with traditional policy tools (Nisa et al., 2019; Stern, 2020). Tor (2020, p. 317) calls nudges that fail, despite being used in conjunction with traditional policy tools, "deficient" nudges. For our purposes, we regard *insufficient* nudges as both those that do not meet desired policy ends when *used alone* (inadequate nudging) *or* when used *in conjunction* with alternative policies (deficient nudging).

Sufficiency can appear monolithic insofar as one assumes there exists an "objective" percentage of policy compliance for the policy to be successful. Yet, what is considered "sufficient," will greatly vary. Consider automatic pension enrolment introduced in the United Kingdom in 2012. This policy changed pension enrolment from being opt-in (i.e., employees actively choose to be enrolled) to opt-out (i.e., employees actively choose to *not* be enrolled). The scheme—known as NEST—has been hailed as a broad success by the UK's Behavioral Insights Team (Service, 2015). The National Audit Office (NAO, 2015) reported only around 8%–14% of employees opting out of the scheme. Bourquin et al. (2020) report that the nudge has increased participation rates among 22–25-year-olds from 20% to 88%, and for 51–55-year-olds from 55% to 93%; figures which broadly align with the NAO. If *sufficiency* here means increasing the percentage of employees in the United Kingdom who are saving *something* for retirement, automatic enrolment is likely sufficient.

Yet, an effective pension program should ensure sufficient income in retirement. With relatively few assumptions, it is possible to estimate weekly retirement income under the current automatic enrolment program. As any individual who is auto-enrolled must, *by definition*, be contributing to national insurance and building state pension eligibility, we should consider that the outcome of the automatic enrolment is to increase pension savings *on top of* state pension provision.

We shall briefly explore the projected income in retirement of an individual who was auto-enrolled into NEST at the age of 22 (the specified enrolment age under the policy; UK Government, n.d.) and who retires at the age of 68 (the pensionable age of those born after April 6th, 1978). We make the same assumptions of investment growth as NEST (2.5% above inflation) and avoid any complicating factors such as spousal benefits, income guarantees of inflation protections, and so on. Forecasts are produced for 2 individuals contributing throughout their working lives at the standard auto-enrolled rate of 5% of income, and with the standard employer contributions (3%) and tax relief (25% of contribution). Finally, we present income figures as current values to allow for comparison against contemporary benchmarks. An individual currently earning the 2022 minimum wage of £19,760 per annum is expected to receive a total weekly income of £232.81. This consists of an auto-enrolled pension income of £90.96 and a state pension of £141.85. An individual earning the 2022 UK median wage of £31,280 receives a weekly income of £286.08, with an auto-enrolled pension of £144.23 and a state pension of £141.85.

Comparing our estimates to one potential standard of *sufficiency*—the After Housing Costs (AHC) relative poverty threshold of £166pw for a single adult with no children, given by the Joseph Rowntree Foundation (2022)—we can see that nudging retirement saving via automatic pension enrolment is sufficient to lift the retirement income of a worker *with no previous provision for their retirement aside from the state pension* above this relative poverty threshold. However, the expected retirement income for both minimum and medium wage individuals is less than the Pensions and Lifetime Savings Association's (PLSA, 2021) estimates for a comfortable retirement (£388.46 per week). In the case of retirement savings nudging may be sufficient to alleviate poverty but is likely insufficient in providing comfort (see Figure 1).

Our model is overly simplistic. It assumes consistency in the long-term behavior of our modeled worker, of the political landscape in which policy is made, and of the macroeconomy. Furthermore, the nudge itself may induce spillover behaviors if an employee *feels* like they need not save in other areas (Beshears et al., 2021; Chater & Loewenstein, 2022; Madrian & Shea, 2001). Our model is simply meant to illustrate that *what is considered* sufficient—which often has a political dimension (Sunstein, 2017)—can change *how* a policy is evaluated.

Often, sufficiency will be a matter of debate. This extends beyond the question of, "what is sufficient to solve the problem," to the question of, "what is the problem to be solved?" Chater and Loewenstein (2022) have been critical of behavioral policy viewing too many policy challenges, and thus solutions, as *individual*-level challenges, ultimately ignoring *systemic*-level policies. For instance, in the example of automatic pension enrolment, sufficiency is debated based on increasing individual saving, on the one hand, and ensuring adequate retirement income, on the other. Neither perspective challenges the notion of *individuated* retirement responsibilities.

Other examples abound. For instance, rather than nudging individuals to become organ donors, and thus addressing the supply of organs, nudging individuals to, say, drink less, or drive more safely, addresses the *demand* for organs. This is assuming nudging *should* be used; as Sunstein et al. (2017), Tor (2020), and Lades and Delaney (2022) each note, more coercive behavioral policy, or simply more coercive policy (Conly, 2013, 2017), may sometimes be worthwhile. Many arguments concerning nudging "crowding out" alternative policy approaches can be seen as intertwined with this political question of *what is to be achieved?*, which informs the question of *what is sufficient*? (Chater & Loewenstein, 2022; Hagman et al., 2019; Lades & Delaney, 2022; Loewenstein & Chater, 2017; Maki, 2019; Sunstein, 2017).

We will conclude our discussion of sufficiency with two important policy examples. These are vaccine uptake, and anthropogenic global warming. Both are worthwhile to discuss as they have reasonably objective standards of sufficiency, *relative* to other policies. For vaccine uptake, *herd immunity* is a common standard (WHO, 2020). For anthropogenic global warming, reducing emissions to prevent global temperatures rising 1.5°C is a common ambition (IPCC, 2022a).

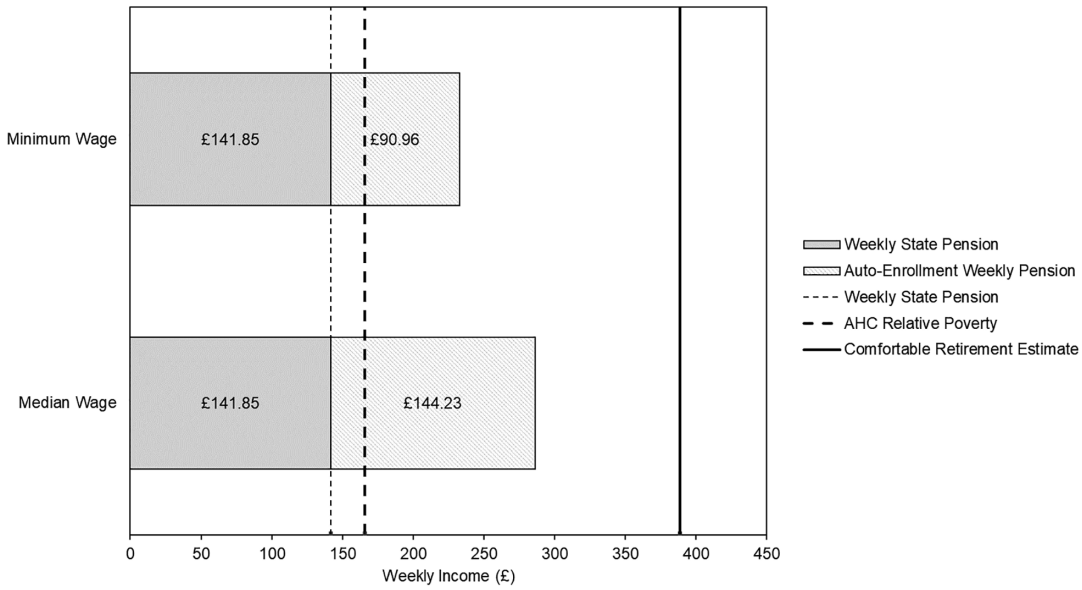

FIGURE 1 Estimated weekly retirement income figures for a median wage and minimum wage worker. Calculations are authors' own.

4.1 | Vaccine uptake

Herd immunity describes a scenario where enough members of a population possess disease antibodies to prevent the proliferation of the disease. Herd immunity requirements, measured as a percentage of the population vaccinated, varies depending on the disease. For instance, "herd immunity against measles requires about 95% of a population to be vaccinated… For polio, the threshold is about 80%" (WHO, 2020, para. 5). These thresholds can be very sensitive, and meeting them may be the difference between safeguarding public health, and risking it. As Plans-Rubió (2012, p. 72) writes, "The objectives of [influenza] vaccination coverage proposed in the United States—80% in healthy persons and 90% in high-risk persons—are sufficient to establish herd immunity, while those proposed in Europe—only 75% in elderly and high-risk persons—are not sufficient."

Uptake of the influenza vaccine, for instance, has been "consistently low" (Oakley et al., 2021, p. 1) in the United Kingdom, as well as in the United States. Oakley et al. (2021) report only 52.4% of at-risk adults (aged 16–64) received the seasonal influenza vaccine in the United Kingdom in 2019. Reducing this differential between *actual* uptake and *sufficient* uptake may represent an opportunity for nudging (Reñosa et al., 2021). For instance, Korn et al. (2018) report social norm nudges significantly increase vaccination intentions.

Examining the impact of nudging on *actual* uptake, Dai et al. (2021) report an increase in uptake of 26% for the COVID-19 vaccine following the use of an appointment reminder nudge, while Kim et al. (2018) reported a more modest increase in uptake of 10% for the influenza vaccine, following an intervention where medical staff prompted patients to book their vaccination appointment at the same time as other medical bookings. Returning to intentions, Kantorowicz-Reznichencko et al. (2022, p. 19) find statistically *insignificant* evidence of saliency nudges changing COVID-19 vaccination intentions among unvaccinated individuals in (relatively) highly vaccinated countries, and where a significant effect is found, they conclude, "the small effect [size of the nudge] does not seem to be promising."

Irrespective of these results—which, following Reñosa et al. (2021, p. 1) suggest, "Nudging-based interventions show potential to increase vaccine confidence and uptake, but further evidence is needed for the development of clear recommendations"—we would emphasize that in this policy area, *statistical significance* (while still important) matters less than whether nudging (in isolation, or in conjunction with other policies) is *sufficient* to achieve herd immunity.

4.2 | Anthropogenic global warming

Global warming is the most substantial policy challenge of the 21st century. The United Nation's Intergovernmental Panel on Climate Change (IPCC) is the foremost authority on the current state of the climate, and global warming-induced environmental harms (IPCC, 2022a). The IPCC produces forecasts and policy recommendations around various warming scenarios, which describe (a) the ecological harm to be expected from a given level of warming, and (b) the rate of carbon reduction required to limit temperature rises to hypothetical figures. The main IPCC targets are 1.5 and 2°C of warming, as these temperatures represent minimum adaptable temperature increases, given an already "locked in" increase of 1°C (IPCC, 2022a). The IPCC (2022b) suggest such targets would necessitate between 70% and 85% of electricity coming from renewables by 2050 (p. 15), with dramatic falls in energy demand from industry and individuals required, owing to time constraints.

It is difficult to convert such reductions in energy demand and carbon emissions to a figure such as a Cohen's *d*, nor would it necessarily be appropriate. Many behavioral interventions to promote pro-environmental behaviors target *individual* behavior, while the IPCC (2022b) scenario for limiting global warming to 1.5°C places a substantial demand on *industry* to reduce energy demand and emissions, while advocating extensive *systemic* investment in energy infrastructure. As various authors note (Hagman et al., 2019; Loewenstein & Chater, 2017; Maki, 2019), the policy solutions which are *sufficient* to meet these objectives are not ones which focus on individual behaviors, such as consumption behaviors.

This is not to say that climate policy does not have a behavioral component to it (Fischhoff, 2021), nor is it to say that changing individual behaviors, such as consumption behaviors, will not be *necessary* to tackling global warming (van der Linden et al., 2021). As the IPCC (2022b) note, *demand* for energy must fall *generally* across sectors in this time-constrained scenario. However, this example is worthwhile because it reveals the importance of *context* when reflecting on sufficiency. For instance, as Nisa et al. (2020) note, even scaling up small pro-environmental interventions is insufficient for the task at hand. Even if the intervention is *statistically significant*, it may not be sufficient given the context of, say, serious time constraints.

Another aspect to consider is whether the intervention is *only changing appearances*. For instance, nudging an individual to switch from a combustion engine car to an electric car does not change the demand for *energy*, only the *way in which that energy is demanded* (also see Chater & Loewenstein, 2022). Another example is an infamous social norm nudge investigated by Schultz et al. (2007) to reduce household energy usage. While informing high-usage households of the *average* energy usage saw a reduction in energy usage from these households, the same nudge saw low-usage households *increase* their usage by almost the same proportion, simply *shifting* energy usage, rather than reducing it—the so-called "magnet" effect (also see Chen et al., 2010).

Interventions may be statistically significant, and *could* have a large effect size, but still ultimately be *insufficient* because only the *appearance* of the problem, not *the problem itself*, is tackled. Indeed, as Werfel (2017) finds, interventions which reframe issues such as global warming as individual issues crowd out support for systemic policies which, to reiterate, the IPCC (2022b) highlight as the *vanguard* policy requirement.

5 | SCALABILITY

We have already seen that making nudges work at-scale is important for behavioral policy (Al-Ubaydli et al., 2021). As van der Linden and Goldberg (2020) argue, successfully scaling a nudge may be key to transforming a small effect-sized intervention into a large, *absolute* effect. This reveals the connections between *scale* and *significance* and *sufficiency*, and as we will show, scale is an important factor in *subjectivity* also. This section covers several broad challenges of scaling interventions—behavioral or otherwise—but we also offer some specific challenges associated with scaling behavioral interventions.

Scaling generally means taking an intervention tested and designed on a (relatively small) sample of the population and applying it to the whole population (Al-Ubaydli et al., 2021). Scaling "what works," can be difficult (List, 2022). While some interventions, based on experimental results (e.g., RCTs), successfully scale—insofar as the effect size of the intervention is preserved—"these are in the minority" (Al-Ubaydli et al., 2021, p. 10). In many instances, "scaling," an intervention results in a diminished effect, compared to that observed in an experimental setting—the so-called *voltage effect* (Al-Ubaydli et al., 2017, 2021; List, 2022). Al-Ubaydli et al. (2021) distinguish between two types of "representativeness" to explain this effect.

First, the sample examined under, say, an RCT study may not be representative of the population which subsequently experiences the scaled intervention. This relates to the above problem of *external validity* in RCTs (Deaton & Cartwright, 2018). Sampling problems can arise for several reasons, as Al-Ubaydli et al. (2021) note: (i) researchers may be biased in some way (e.g., publication bias;Della Vigna & Linos, 2022; Maier et al., 2022); (ii) there may be some unforeseen selection bias occurring, or a problem of "confounding randomness" (Deaton & Cartwright, 2018; also see Al-Ubaydli et al., 2017); (iii) participants in any experiment are necessarily those who *choose* to participate, and in participating, may come with prior beliefs which bias results (e.g., "adverse heterogeneity;" Al-Ubaydli et al., 2017, p. 283); (iv) samples are often constrained by resources, for instance time, funding, or serendipity (e.g., natural experiments; Al-Ubaydli et al., 2017).

Population challenges in scaling have interesting consequences for nudges. On the one hand, nudges assume systemic biases in human cognition (Thaler & Sunstein, 2008). In this regard, failures to scale nudge interventions may not be due to a failure of the *nudge* to scale, but of a failure of the logistical support which surrounds the nudge (e.g., Castleman, 2021; see below). Yet, nudges are unlikely *immune* to these population problems. Various authors (Della Vigna & Linos, 2022; Maier et al., 2022) find evidence of publication bias in nudge studies, suggesting that even if the assumption of the universality of biases holds for *some results*, it is highly doubtful this assumption holds for *all* results (and perhaps even *most* results; Maier et al., 2022). One should also be wary of the potentially confounding effects of the population. For instance, a population—with a specific set of cultural characteristics—may be more susceptible to social norm messengers than other populations with a different set of characteristics (Dolan et al., 2012; Schimmelpfennig & Muthukrishna, 2022). Failure to scale this intervention *may* be due to the design of the nudge (e.g., choosing the right messenger), but may also come from the relatively lacking importance of social influence within a target population.

Recent calls for more investigation into the mechanisms behind nudges align with this debate (Banerjee & John, 2023; de Ridder et al., 2022). Without understanding the mechanism behind a nudge, it may be difficult to evaluate whether the sample that a nudge was tested on is generalizable to the population (Bryan et al., 2021). For instance, a sample may be evenly split by gender—and so broadly generalizable to most populations—but if the psychological mechanism behind the nudge is not tied to gender, arguing that the nudge effect will scale is liable to result in error. Work by Chapman et al. (2023) has sought to find a more parsimonious model of behavioral biases from which to build consistent theories of human decision-making, which should support investigations of nudge-mechanisms, and thus aid scalability (also de Ridder et al., 2022). Note that these critiques are not wholly detrimental to nudging. As Soman and Hossain (2021) argue, even interventions which fail to scale can be valuable to behavioral policymakers for informing hypotheses, theory, and future experimental practices.

Second, the context or situation in which the intervention is initially tested and found to be significant should be maintained as the intervention is scaled. Al-Ubaydli et al. (2021) call this notion *situational representativeness*. We will focus on this aspect of scaling. Situational representativeness reveals several aspects of nudge design which may be missed through a mere focus on, say, statistical significance. For instance, Castleman (2021) discusses why various reminder nudges to encourage high school students in the United States to enter higher education have generally failed to scale effectively, despite promising experimental evidence. Discussing a collaborative experiment (Castleman & Page, 2015) utilizing a text message intervention to increase higher education uptake among low-income high school graduates, Castleman (2021) argues that this intervention failed to scale because the components of the intervention failed to scale *proportionately*. While the nudge component (and the *headline* component)

scaled easily, other elements did not. *Specifically*, in the experiment, students could respond to the text message to enter a one-on-one conversation with an educational mentor. When scaled, this mentor component did not (Castleman, 2021). This reveals two further considerations.

First, that the headline mechanism driving behavior change—in this instance, the text message—may be overstated in its function, compared to secondary and tertiary mechanisms (Chater & Loewenstein, 2022). Seeing a nudge as the primary factor for changing behavior, rather than an *enabling* intervention for more substantive behavior change, could lead policymakers to misunderstanding the intervention, while creating the potential for misuse. For instance, if a cheap nudge is offered as the policy solution, rather than a cheap nudge *and* further investment, a cost–benefit analysis may *accidently* attribute benefits to a scaled nudge, and subsequently reject the superior (though seemingly more expensive) policy (Ewert, 2020). This is to say nothing of the potential *political capital* to be gained through *purposely* emphasizing nudging to understate costs and overstate benefits (Chater & Loewenstein, 2022).

Second, that scaling is a monolithic term which may not appreciate that some elements of an intervention will scale easily, and others will not. For instance, the marginal cost of scaling a nudge is often close to zero, be it switching a default option, changing a social norm, or sending 10,000 (rather than 100) push notifications. However, the marginal cost of procuring employees to support an intervention, and *training* those employees to be proficient (up to the standard of the original intervention) may be high, and these costs may scale *linearly*, rather than with a *diminishing* cost (Al-Ubaydli et al., 2021). For instance, an attempt to scale an experimental intervention regarding class size in California was found to have "dampened" benefits as scaling required the hiring of less-experienced and undertrained staff, compared to those available in the original intervention (Al-Ubaydli et al., 2017, 2021; Jepsen & Rivkin, 2009, p. 223). For a nudge-specific example, we might consider transport. The United Kingdom and other countries have sought to nudge electric vehicle ownership using "green" number plates on low- and zero-emission vehicles (Costa et al., 2018). This is an immediately scalable nudge, and potentially a worthwhile one, with the Automobile Association (2020) finding that 18% of survey respondents *may be influenced* to purchase an electric car because of the nudge. Yet, even if the *nudge* is scalable, the behavior change is not, owing to the cost of electric vehicles, the provision of electric vehicle infrastructure, and the supply of raw materials such as lithium (Marx, 2022). An effective nudge will not successfully scale if the behavior change it is promoting cannot itself scale.

A final—though we acknowledge more *general*—element to consider is *by what factor is an intervention being scaled*? The discussion thus far has mostly focused on two scales: "small scale" experiments, and "large scale," national policies. However, in many countries, intermediary legislative regions exist. There may be a risk that, when scaling interventions, these intermediaries appear "small" in comparison to the national-level, and may thus receive inadequate resources for scaling, leading to incidents such as those described by Castleman (2021) and Jepsen and Rivkin (2009). A "national rollout" may more easily galvanize national—and thus *political*—support than a "regional rollout" which remains framed as a "small, local" policy. This would be unfortunate. For instance, scaling an RCT experiment conducted on 1000 people to a municipality the size of, say, Berlin (with a population of around 3.5 m) induces a scaling factor of 3500, while scaling this regional-level policy to the national level induces a scaling factor of only around 24 (based on a German population of around 84 m). In both instances, these are substantial figures, but where the success of an intervention is often predicated on political will to provide resources (Chater & Loewenstein, 2022; Lades & Delaney, 2022; Sunstein, 2017), and where political will can often be shaped by framing (Maki, 2019; Werfel, 2017), appreciating relative scaling factors may be important to the scaling discussion.

6 | SUBJECTIVITY

The final component of the 4S framework is *subjectivity*. *Subjectivity* broadly captures the role of heterogeneity in behavioral interventions, and more specifically, focuses on the question of *welfare* and *wellbeing* effects arising from

these interventions. In some instances, the emphasis within the *subjectivity* element will focus on the sample being analyzed, and in particular, on how the sample is *stratified* to reveal different (or more complicated) behavioral findings, with different (or more complicated) welfare implications. Because there are many ways in which a sample could be stratified, or *welfare* and *wellbeing* could be analyzed, we have chosen to use the broad term *subjectivity*. However, owing to the importance of stratification, this may be a "rule-of-thumb" alternative name for this element of the framework.

Reflections on subjectivity in nudging have emerged from two, related perspectives. *First*, the personalized nudging literature has highlighted the so-called "problem of heterogeneity," (Sunstein, 2012, p. 6) which argues that different people cannot be assumed to respond to the same nudge in the same way, and that making this assumption may harm individuals (Bryan et al., 2021; Mills, 2022; Peer et al., 2020; Sunstein, 2012, 2022). *Second*, a behavioral welfare literature has emerged to explore the question of whether nudges *actually* benefit individuals, and in probing such a question, has begun to uncover interesting results about *who* benefits and *who* suffers as a result of nudge interventions (Brown et al., 2022; Bulte et al., 2020; Lades & Delaney, 2022; Laffan et al., 2021; List et al., 2022; Thunström, 2019; Thuström et al., 2018; Tor, 2020).

Ultimately, both literatures deal with *heterogeneity*, and investigate different vectors by which a sample may be stratified to reveal novel results, beyond initially statistically significant findings. Indeed, in instances where two subsamples "cancel" each other out, the effect of an intervention on the whole sample may appear *insignificant* (e.g., Laffan et al., 2021; Schultz et al., 2007); but this may mask a more complex story with important consequences for welfare.

Thunström (2019) investigates the welfare effects of calorie label nudges on food decisions. In her sample, she distinguishes between those who report having low self-control regarding decisions about food, and those with high self-control. Thunström (2019) finds the nudge (i.e., food labels) functions as an *emotional tax*, eliciting a greater emotional response from low self-control individuals compared to high self-control individuals, who actually receive a hedonic benefit. Given this, Thunström (2019, p. 11) concludes, "[the nudge] therefore emotionally taxes the "right" people." However, when evaluating *behavior* and food decisions, Thunström (2019) finds that low self-control individuals adjust their consumption habits *less* than high self-control individuals. Thus, even though the nudge taxes the "right" sub-sample, the "wrong" sub-sample ultimately benefits more from the nudge. Indeed, the high self-control sub-sample *doubly* benefit, *first* from the hedonic benefit of reacting positively to the nudge, and *second* from the (albeit assumed) health benefit of *following* the nudge. When such an intervention is evaluated without considering heterogeneity, it is likely the nudge would be considered effective on the grounds of statistical significance. But the distributional effects of the nudge point to a more complex, and less immediately *positive*, conclusion (Sunstein, 2022).

Thunström et al. (2018) investigate the use of a reminder nudge to reduce spending behavior. The nudge is designed to remind participants of the opportunity cost of purchasing a product, in the hopes of discouraging purchasing. Indeed, when examining the sample overall, they find this nudge does significantly reduce spending for a treatment group, compared to the control. However, Thunström et al. (2018) *also* ask participants to self-report whether they identify as someone who spends too much ("spendthrifts") or someone who spends too little ("tightwads"). Comparing the effect of the nudge on these subsamples, they find *no significant effect* on spendthrifts—possibly because these people have a tolerance for overcoming opportunity cost—but a *significant effect* on tightwads—possibly because these people have a high sensitivity to opportunity cost. Thunström, Gilbert and Jones-Ritten (2018, p. 267) describe this intervention as a "[nudge] that hurt[s] those already hurting" insofar as tightwads likely will not benefit from *not spending*, while *already* reporting feelings that they do not spend enough. Furthermore, this intervention is statistically significant across the sample; yet, when *subjectivity* is considered, the importance of significance is called into question (also see Bryan et al., 2021).

Subjectivity need not always reveal hidden costs or negatives arising from interventions. Arulsamy and Delaney (2020) investigate the interaction of the United Kingdom automatic enrolment pension program and mental health, finding—prior to the 2012 introduction of the policy—there was a significant difference in saving rates between

those reporting some "psychological distress" and those reporting no "psychological distress," with pension participation rates being 4.4% higher for the latter than the former. By 2016, the gap had closed to only 1.4%, and was no longer statistically significant, while *both* groups saw an increase in participation rates of around 20%. Thus, *subjectivity* in this instance reveals a hidden *benefit* of the automatic enrolment nudge: not only did the nudge increase pension participation across the whole sample, but it also had an outsized benefit on a disadvantaged subsample (for another example, see Brown et al., 2022).

Without factoring in subjectivity, assessments of whether a nudge is "effective" or not may be inaccurate. Yet, this example also reveals an important challenge of subjectivity, namely, that there are many ways to stratify a sample. For instance, evaluating the *same* policy as Arulsamy and Delaney (2020), Bourquin et al. (2020) note that the policy may have benefited higher earners more than lower earners, in two ways. *First*, higher earners can save more, and thus exploit various tax benefits of the policy which lower earners cannot exploit. *Second*, lower earners may not be able to afford to save anything, and thus face the burden of opting-out of the scheme, as well as the potential financial shock of a 5% pay "reduction." As such, subjectivity is not a panacea for questions of welfare, inequality, and distributional effects.

One approach to subjectivity may be to personalize nudges to target different sub-populations (Mills, 2022; Sunstein, 2022, 2012)—what Sunstein (2013, p. 1871, original emphasis) has broadly called, "*personalized paternalism*." Personalization has been shown to be broadly advantageous in some areas of nudging, such as cybersecurity (Peer et al., 2020) and water consumption (Schultz et al., 2016), building on a more established literature of personalization and personality targeting in the marketing literature (e.g., Hirsh et al., 2012; Matz et al., 2017; Moon, 2002). For instance, Schultz et al. (2016) find that personalizing feedback regarding household water consumption significantly reduces the usage of water by high-usage households, but does not increase usage by low-usage households. Contrasting this result with Schultz et al. (2007), where a so-called "magnet" effect in energy consumption was observed, personalization appears to have overcome this challenge.

Yet, personalization is not *perfect*. As above, because there are many ways in which a population can be stratified, there remain important questions about *when* personalization is appropriate, *who* should be targeted, and *how* (Mills, 2022). Furthermore, personalization—and appreciation of subjectivity more broadly—complicates interventions, and adds to costs in terms of time, effort, and potential privacy costs (Sunstein, 2012, 2013). While not accounting for heterogeneity in interventions may lead to overzealous statistical conclusions (Bryan et al., 2021; Deaton & Cartwright, 2018; Della Vigna & Linos, 2022) and challenges in scaling (Al-Ubaydli et al., 2017, 2021)—both of which may undermine the *sufficiency* of any intervention—*incorporating* heterogeneity may also be difficult. At times, this may also be an *inappropriate* endeavor. For instance, the *publicity principle* describes the idea that policies should be sufficiently *public*, and publicly *understandable*, so that the general population can debate those policies, and reject them if deemed unacceptable (Rawls, 1971). As Mills (2022) argues, personalized policy may undermine the understandability of a policy, and thus undermine its democratic legitimacy.

7 | DISCUSSION

The 4S framework contributes to the existing behavioral policy "infrastructure" as an evaluative framework which complements various design frameworks, as shown in Figure 2.

When evaluated through the 4S framework, we suggest a nudge receives one of three outcomes (the 3Rs; a policy may *technically* face no criticisms, but this is highly unlikely). *First*, a nudge may require *redesigning*. For instance, a social norm nudge may be personalized to eliminate any "magnet" effect (e.g., Schultz et al., 2016). *Second*, a nudge may require *reformulation* through combining with other policy tools. For instance, nudging students to engage with higher educational resources, such as one-on-one mentors (Castleman & Page, 2015). *Third*, a nudge may require

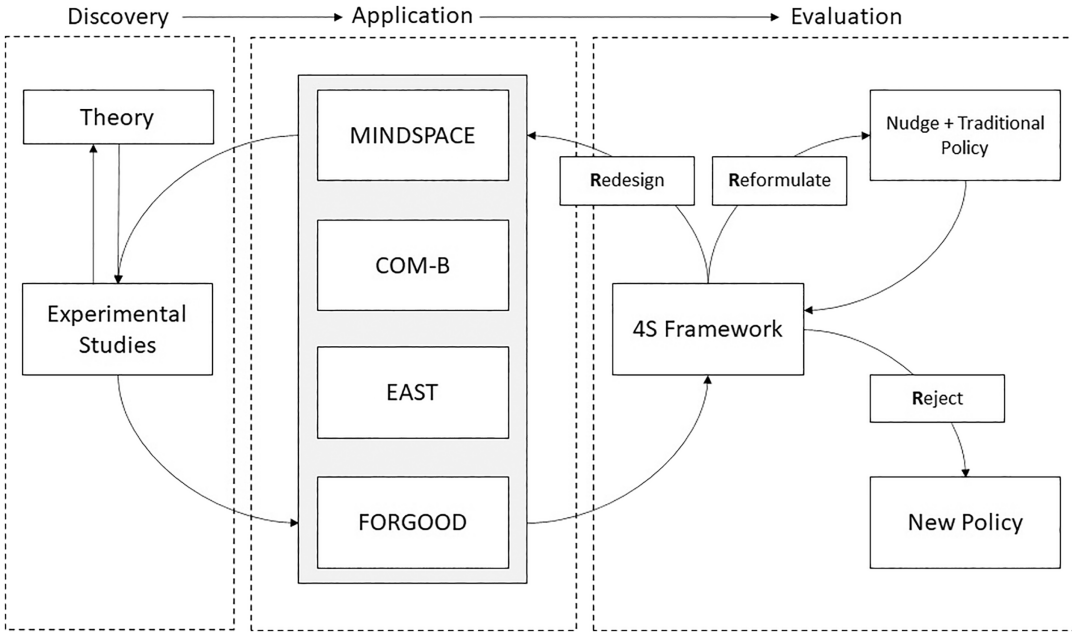

FIGURE 2 The 4S framework within the wider behavioral policy "Infrastructure."

rejection. For instance, abandoning a plan to nudge electric vehicle ownership, and instead investing in zero-emissions public transport (also see List et al., 2022).

These outcomes are to be informed through reflections by policymakers on each of the elements of the 4S framework. However, as we have intimated throughout, each element of the framework has *trade-offs* with *other elements* within the framework. A nudge may be statistically significant in an experimental setting, but the effect size may be *insufficient*, the intervention may not *scale*, and *subjective* factors may demand re-evaluation of result. A nudge may be sufficient, but only *significant* when used in conjunction with other policies, *scale* at a substantial cost, and be unequal or harmful for different *subjects*. A nudge may be scalable, but result in lessened *significance*, be costly compared to other *sufficient* policies, and be complicated when accounting for *subjectivity*. Finally, a nudge may account for subjectivity, but undermine *statistical significance*, be difficult to evaluate in terms of *sufficiency*, and as above be complicated to *scale*.

Various examples discussed throughout this article illustrate some potential trade-offs. For instance, take the much-discussed example of automatic pension enrolment in the United Kingdom. This nudge has clearly scaled to a national level, given its high uptake in the United Kingdom. Yet, we also present evidence that the pension plan itself is likely insufficient to provide a comfortable retirement. This does not immediately seem relevant to the nudge component—changing a default option—but it is. The amount that a person saves under the scheme represents a "cost" in terms of deferred income for the saver (Laibson, 1997). Setting the "cost" too high would likely reduce the policy uptake, undermining the effective scaling of the nudge. For instance, the recent "cost of living" crisis in the United Kingdom has seen people leave their schemes as the cost of contributing has become unaffordable (Cumbo, 2022).

Therefore, there is a trade-off between ensuring the nudge can scale adequately, and ensuring the nudge leads to a sufficient policy outcome. In the case of UK automatic enrolment, sufficiency may have been compromised—as we have shown, the nudge may alleviate poverty in retirement, but likely does not provide comfort. Furthermore, we have already discussed trade-offs in terms of *subjectivity* for this nudge policy. The nudge seems to have positive distributional effects when the mental health of employees is examined (Arulsamy & Delaney, 2020), though negative

distributional effects when the socioeconomic status of employees is examined (Bourquin et al., 2020). As above, there are many (perhaps endless) ways of stratifying a sample to analyze a nudge. As such, questions such as "is this nudge sufficient?" or, "is this nudge significant?" are contingent on what degree *subjectivity* is prioritized (e.g., sufficient, or significant, *for whom*?).

As such, the 4S framework is not offered as a checklist to be used to outright determine an "effective" nudge from an "ineffective" one. This would be impractical, if not impossible. The framework instead allows policymakers to determine what element matters most within the policy context, and choose the appropriate path (i.e., redesign, reformulate, reject) to follow. For instance, treating *subjectivity* as focusing on different stakeholder groups, rather than more broadly "thinking about welfare." Such a pragmatic, realist approach to nudge evaluation has forbears in previous policy evaluation literature (e.g., Pawson & Tilley, 1994).

In this same spirit, the 4S framework does not address all criticisms of nudging, or behavioral policy. Substantial debates remain in the nudge literature about the efficacy of nudging (Bovens, 2009; Hansen & Jespersen, 2013; Henderson, 2014; Oliver, 2019; Rebonato, 2012, 2014; Rizzo & Whitman, 2020; Ryan, 2018; Selinger & Whyte, 2010, 2011, 2012; Veetil, 2011). These debates are valid, and often engage with questions important to policymaking, such as, "what are the values of our society?" The 4S framework does not engage with these debates. While the framework does suggest that a nudge which fails against the 4 S's could be rejected in favor of a different policy (one of the 3 R's is *reject*), such rejection—within the framework—is because of some critical weakness (e.g., the nudge does not scale) rather than some value-based critique (e.g., people should be free from interference). Where such debates are important, existing frameworks—such as the FORGOOD ethical framework (Lades & Delaney, 2022)—become important companions to the 4S framework.

A further limitation of the framework is while it acknowledges the limitations concerning "evidence"—in terms of both the weaknesses of collecting evidence and the debate about what evidence is—the framework does not advocate a new method or direction for determining and collecting evidence. For instance, we have re-emphasized various criticisms of RCTs, but we have not offered an alternative means of determining effect sizes. Some may regard this as a substantial weakness of our framework. Yet, the purpose of the 4S framework is to offer criteria *in addition* to statistical data which are important to incorporate into an evaluation of a nudge. These additional criteria are necessary precisely because of the criticisms of RCTs expressed by authors such as Deaton and Cartwright (2018).

The 4S framework represents an important synthesis of emerging threads in the nudge and behavioral policy literatures, and a potential bridge between these literatures and the wider policy evaluation literature. The framework speaks to important criticisms of nudges which have been raised in recent years (e.g., Chater & Loewenstein, 2022; de Ridder et al., 2022; List et al., 2022; Maier et al., 2022; Nisa et al., 2020), and offers an approach for the field to meet the challenge posed by Nisa et al. (2020) to go "beyond statistical significance."

8 | CONCLUSION

Nudge interventions face growing criticism. These criticisms come from various angles and undermine the apparent efficacy of several nudges. Behavioral policymakers have overwhelmingly regarded statistical significance as the benchmark for determining that an intervention "works." However, within a wider policy context, statistical significance alone is insufficient to judge whether a nudge is an effective policy response.

We offer the 4S framework as a more comprehensive evaluative framework for nudges, which when applied encourages behavioral policymakers to evaluate whether a nudge intervention "works," from several important perspectives in addition to statistical significance. We argue nudges must be judged against their *sufficiency* to resolve policy challenges, on their ability to *scale* to policy audiences, and on their *subjective* characteristics. As nudging matures and comes to face more complex and challenging policy environments, the 4S framework will be an important evaluative tool for effective policymaking.

ACKNOWLEDGMENTS

The authors are grateful to the editors of *Public Administration* for their handling of this article, and to the two anonymous reviewers for their insightful and constructive comments which have substantially improved this article. All errors are our own.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Stuart Mills https://orcid.org/0000-0002-6698-0983

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/padm.12941.

DATA AVAILABILITY STATEMENT

All data and figures used or otherwise discussed within this article are within the public domain.

REFERENCES

Al-Ubaydli, O., Lee, M.S., List, J.A., MacKevicius, C.L. & Suskind, D.L. (2021) How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, *5*(1), 2–49.

Al-Ubaydli, O., List, J.A. & Suskind, D.L. (2017) What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review*, *107*(5), 282–286.

Arulsamy, K. & Delaney, L. (2020) *The impact of automatic enrolment on the mental health gap in pension participation: evidence from the UK*. UCD Geary Institute for Public Policy Discussion Paper Series WP2020/04. [Online] Available at: https://www.ucd.ie/geary/static/publications/workingpapers/gearywp202004.pdf [Accessed 11th July 2022]

Automobile Association. (2020) *Green light for EV plates*. [Online] Available at: https://www.theaa.com/about-us/newsroom/first-green-number-plates-for-electric-vehicles [Accessed 10th July 2022]

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S. et al. (2017) From proof of concept to scalable policies: challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73–102.

Banerjee, A., Chassang, S., Montero, S. & Snowberg, E. (2017) *A theory of experiments*. NBER Working Paper No. 23867. [Online] Available at: https://www.nber.org/papers/w23867 [Accessed 30th June 2022]

Banerjee, A., Chassang, S. & Snowberg, E. (2017) Decision theoretic approaches to experiment design and external validity. NBER Working Paper No. 22167. [Online] Available at: https://www.nber.org/system/files/working_papers/w22167/w22167.pdf [Accessed 30th June 2022]

Banerjee, A. & Duflo, E. (2009) The experimental approach to development economics. *Annual Review of Economics*, *1*, 151–178.

Banerjee, A. & Duflo, E. (2010) Giving credit where it is due. *Journal of Economic Perspectives*, *24*(3), 61–80.

Banerjee, S. & John, P. (2023) Nudge and nudging in public policy. In: van Gerven, M., Allison, C.R. & Schubert, K. (Eds.) *Springer encyclopedia of public policy*. USA: Springer. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314881 [Accessed 4th January 2023]

Beshears, J., Dai, H., Milkman, K.L. & Benartzi, S. (2021) Using fresh starts to nudge increased retirement savings. *Organizational Behavior and Human Decision Processes*, *167*, 72–87.

Beshears, J. & Kosowsky, H. (2021) Nudging: Progress to date and future directions. *Organizational Behavior and Human Decision Processes*, *161*, 3–19.

Bourquin, P., Cribb, J. & Emmerson, C. (2020) *Who leaves their pension after being automatically enrolled?* Institute for Fiscal Studies. [Online] Available at: https://ifs.org.uk/uploads/Who-leaves-their-pension-after-being-automatically-enrolled-BN272.pdf [Accessed 07th July 2022]

Bovens, L. (2009) The ethics of Nugde. In: Hansson, M.J. & Grüne-Yanoff, T. (Eds.) *Preference change: approaches from philosophy, economics and psychology*. Berlin: Springer.

Brice, B.D. & Montesinos-Yufa, H.M. (2019) *The era of empirical evidence*. [Online] Available at: https://www.researchgate.net/publication/318600096 [Accessed 23rd January 2023]

Brown, A.L., Grodzicki, D. & Medina, P.C. (2022) *When nudges spill over: student loan use under the CARD act*. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4129413 [Accessed 30th July 2022]
Bryan, C.J., Tipton, E. & Yeager, D.S. (2021) Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*, 980–989.
Bulte, E., List, J.A. & van Soest, D. (2020) Toward an understanding of the welfare effects of nudges: evidence from a field experiment in the workplace. *The Economic Journal*, *130*(632), 2329–2353.
Castleman, B.L. (2021) *Why aren't text message interventions designed to boost college success working at scale?* Behavioral Scientist. [Online] Available at: https://behavioralscientist.org/why-arent-text-message-interventions-designed-to-boost-college-success-working-at-scale/ [Accessed 06th July 2022]
Castleman, B.L. & Page, L.C. (2015) Summer nudging: can personalized text message and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior and Organization*, *115*, 144–160.
Chapman, J., Dean, M., Ortoleva, P., Snowberg, E. & Camerer, C. (2023) Econographics. *Journal of Political Economy: Microeconomics*, *1*, 115–161. Available from: https://doi.org/10.1086/723044
Chater, N. & Loewenstein, G. (2022) The i-frame and the s-frame: how focusing on individual-level solutions has led behavioral public policy astray. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4046264 [Accessed 30th June 2022]
Chen, Y., Harper, F.M., Konstan, J. & Li, S.X. (2010) Social comparisons and contributions to online communities: a field experiment on MovieLens. *American Economic Review*, *100*(4), 1358–1398.
Chiou, W., Yang, C. & Wan, C. (2011) Ironic effects of dietary supplementation: illusory invulnerability created by taking dietary supplements licenses health-risk behaviors. *Psychological Science*, *22*(8), 1081–1086.
Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. USA: Lawrence Erlbaum.
Conly, S. (2013) *Against autonomy: justifying coercive paternalism*. UK: Cambridge University Press.
Conly, S. (2017) Paternalism, coercion and the unimportance of (some) liberties. *Behavioural Public Policy*, *1*(2), 207–218.
Costa, E., Reiner, C. & Fitzhugh, E. (2018) *How new number plates could green Britain's roads*. BIT. [Online] Available at: https://www.bi.team/blogs/how-new-number-plates-could-green-britains-roads/ [Accessed 20th August 2022]
Cumbo, J. (2022) *Workers halt pension savings as cost of living crisis bites*. The Financial Times. [Online] Available at: https://www.ft.com/content/779cfb06-d21f-4d87-be55-65b3ef9780c6 [Accessed 20th August 2022]
Dai, H., Saccardo, S., Han, M.A., Roh, L., Raja, N., Vangala, S. et al. (2021) Behavioural nudges increase COVID-19 vaccinations. *Nature*, *597*, 404–409.
Dawid, A.P. (2000) Causal inference without counterfactuals. *Journal of the American Statistical Association*, *95*(450), 407–424.
de Ridder, D., Kroese, F. & van Gestel, L. (2022) Nudgeability: mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science*, *17*(2), 346–359.
Deaton, A. & Cartwright, N. (2018) Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, *210*, 2–21.
Della Vigna, S. & Linos, E. (2022) RCTs to scale: comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116.
Dolan, P. & Galizzi, M.M. (2015) Likes ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, *47*, 1–16.
Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R. & Vlaev, I. (2012) Influencing behaviour: the mindspace way. *Journal of Economic Psychology*, *33*(1), 264–277.
Duflo, E. (2011) *A research agenda for development economics*. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1888605 [Accessed 08th January 2023]
Duflo, E. & Kremer, M. (2008) Use of randomization in the evaluation of development effectiveness. In: Easterly, W. (Ed.) *Reinventing foreign aid*. USA: Brookings.
Eaglesham, J. (2008) *Nudging not nannying to achieve social goals*. The Financial Times. [Online] Available at: https://www.ft.com/content/0e6f1088-6280-11dd-9a1e-000077b07658 [Accessed 31st January 2022]
Entwistle, T. (2021) Why nudge sometimes fails: fatalism and the problem of behaviour change. *Policy and Politics*, *49*(1), 87–103.
Ewert, B. (2020) Moving beyond the obsession with nudging individual behaviour: towards a broader understanding of Behavioural public policy. *Public Policy and Administration*, *35*(3), 337–360.
Feyerabend, P. (2010) *Against method*. UK: Verso Books.
Fischhoff, B. (2021) Making behavioral science integral to climate and action. *Behavioural Public Policy*, *5*(4), 439–453.
Hagman, D., Ho, E.H. & Loewenstein, G. (2019) Nudging out support for a carbon tax. *Nature Climate Change*, *9*, 484–489.
Hallsworth, M. (2022a) Making sense of the "do nudges work?". Debate Behavioral Scientist. [Online] Available at: https://behavioralscientist.org/making-sense-of-the-do-nudges-work-debate/ [Accessed 04th January 2023]

Hallsworth, M. (2022b) *Misconceptions about the practice of behavioral public policy*. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4328659 [Accessed 26th January 2023]
Hallsworth, M. & Kirkman, E. (2020) *Behavioral insights*. USA: MIT Press.
Hamermesh, D.S. (2013) Six decades of top economics publishing: who and how? *Journal of Economic Literature*, *51*(1), 162–172.
Hansen, P.G. & Jespersen, A.M. (2013) Nudge and the manipulation of choice: a framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, *4*(1), 3–28.
Harford, T. (2022) *What nudge theory got wrong*. The Financial Times. [Online] Available at: https://www.ft.com/content/a23e808b-e293-4cc0-b077-9168cff135e4 [Accessed 30th June 2022]
Haynes, L., Service, O., Goldacre, B. & Torgerson, D. (2012) *Test, learn, adapt: developing public policy with randomised controlled trials*. UK Cabinet Office. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf [Accessed 14th March 2022]
Hecht, C.A., Dweck, C.S., Murphy, M.C. & Yeager, D.S. (2022) Efficiently exploring the causal role of contextual moderators in behavioral science. *Proceedings of the National Academy of Science*, *120*(1), e.2216315120.
Henderson, D.R. (2014) Libertarian paternalism: leviathan in Sheep's clothing? *Society*, *51*, 268–273.
Hirsh, J.B., Kang, S.K. & Bodenhausen, G.V. (2012) Personalized persuasion: tailoring persuasive appeals to recipients' personality traits. *Psychological Science*, *23*(6), 578–581.
IPCC. (2022a) *Climate change 2022: impacts, adaption and vulnerability*. IPCC. [Online] Available at: https://www.ipcc.ch/report/ar6/wg2/downloads/report/IPCC_AR6_WGII_FinalDraft_FullReport.pdf [Accessed 09th July 2022]
IPCC. (2022b) *Special report on global warming of 1.5°C*. IPCC. [Online] Available at: https://www.ipcc.ch/site/assets/uploads/sites/2/2022/06/SPM_version_report_LR.pdf [Accessed 09th July 2022]
Jamal, F., Fletcher, A., Shackleton, N., Elbourne, D., Viner, R. & Bonell, C. (2015) The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-exampke. *Trials*, *16*(466), 466. Available from: https://doi.org/10.1186/s13063-015-0980-y
Jepsen, C. & Rivkin, S. (2009) Class size reduction and student achievement: the potential tradeoff between teacher quality and class size. *Journal of Human Resources*, *44*(1), 223–250.
John, P., Martin, A. & Mikołajczak, G. (2022) Support for behavioral nudges versus alternative policy instruments and their perceived fairness and efficacy. *Regulation & Governance*, *17*, 363–371. Available from: https://doi.org/10.1111/rego.12460
Joseph Rowntree Foundation. (2022) *UK poverty 2022: the essential guide to understanding poverty in the UK*. Josehp Rowntree Foundation. [Online] Available at: https://www.jrf.org.uk/report/uk-poverty-2022 [Accessed 07th July 2022]
Kantorowicz-Reznichencko, E., Kantorowicz, J. & Wells, L. (2022) Can vaccination intentions against COVID-19 be nudged? *Behavioural Public Policy*, 1–25. Available from: https://doi.org/10.1017/bpp.2022.20
Kim, R.H., Day, S.C., Small, D.S., Snider, C.K., Rareshide, C.A.L. & Patel, M.S. (2018) Variations in influenza vaccination by clinic appointment time and an active choice intervention in the electronic health record to increase influenza vaccination. *JAMA Network Open*, *1*(5), e.181770.
Kohler-Hausmann, I. (2020) Nudging people to court. *Science*, *370*(6517), 658–659.
Korn, L., Betsch, C., Böhm, R. & Meier, N.W. (2018) Social nudging: the effect of social feedback interventions on vaccine uptake. *Health Psychology*, *37*(11), 1045–1054.
Krpan, D., Galizzi, M.M. & Dolan, P. (2019) Looking at spillovers in the mirror: making a case for "Behavioral Spillunders". *Frontiers in Psychology*, *10*. Available from: https://doi.org/10.3389/fpsyg/2019/01142
Kuhn, T.S. (2012) *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.
Lades, L.K. & Delaney, L. (2022) Nudge FORGOOD. *Behavioural Public Policy*, *6*(1), 75–94.
Laffan, K., Sunstein, C.R. & Dolan, P. (2021) Facing it: assessing the immediate emotional impacts of calorie labelling using automatic facial coding. *Behavioural Public Policy*, 1–18. Available from: https://doi.org/10.1017/bpp.2021.32
Laibson, D. (1997) Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.
List, J.A. (2022) *The voltage effect: how to make good ideas great and great ideas scale*. UK: Penguin Books.
List, J.A., Rodemeier, M., Roy, S. & Sun, G. (2022) *Judging nudging: toward an understanding of the welfare effects of nudges versus taxes*. Working Paper. [Online] Available at: https://s3.amazonaws.com/fieldexperiments-papers2/papers/00765.pdf [Accessed 04th January 2023]
Loewenstein, G. & Chater, N. (2017) Putting nudges in perspective. *Behavioural Public Policy*, *1*(1), 26–53.
Madrian, B.C. & Shea, D.F. (2001) The power of suggestion: inertia in 401(k) participation and savings behavior. *The Quarterly Journal of Economics*, *116*(4), 1149–1187.
Maier, M., Bartoš, F., Stanley, T.D. & Wagenmakers, E. (2022) No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, *119*(31), e.2200300119.
Maki, A. (2019) The potential cost of nudges. *Nature Climate Change*, *9*(6), 435–439.
Maki, A., Carrico, A.R., Raimi, K.T., Truelove, H.B., Araujo, B. & Yeung, K.L. (2019) Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, *2*, 307–315.

Marx, P. (2022) *Road to nowhere: what Silicon Valley gets wrong about the future of transportation*. UK: Verso.
Matz, S.C., Kosinski, M., Nave, G. & Stillwell, D.J. (2017) Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 12714–12719.
Mazar, N. & Zhong, C. (2010) Do green products make us better people? *Psychological Science*, *21*(4), 494–498.
Merritt, A.C., Effron, D.A. & Monin, B. (2010) Moral self-licensing: when being good frees us to Be bad. *Social and Personality Psychology Compass*, *4*(5), 344–357.
Michie, S., van Stralen, M.M. & West, R. (2011) The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*, 1–11.
Mills, S. (2022) Personalized nudging. *Behavioural Public Policy*, *6*(1), 150–159.
Moon, Y. (2002) Personalization and personality: some effects of customizing message style based on consumer personality. *Journal of Consumer Psychology*, *12*(40), 313–325.
National Audit Office. (2015) *Automatic enrolment to workplace pensions*. National Audit Office. [Online] Available at: https://www.nao.org.uk/wp-content/uploads/2015/11/Automatic-enrolment-to-workplace-pensions.pdf [Accessed 07th July 2022]
Nisa, C.F., Bélanger, J.J., Schumpe, B.M. & Faller, D.G. (2019) Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nature Communications*, *10*(4545), 1–13. Available from: https://doi.org/10.1038/s41467-019-12457-2
Nisa, C.F., Sasin, E.M., Faller, D.G., Schumpe, B.M. & Bélanger, J.J. (2020) Reply to: alternative meta-analysis of behavioural interventions to promote action on climate change yields different conclusions. *Nature Communications*, *11*(3901), 1–3. Available from: https://doi.org/10.1038/s41467-020-17614-6
Oakley, S., Bouchet, J., Costello, P. & Parker, J. (2021) Influenza vaccine uptake among at-risk adults (aged 16-64 years) in the UK: a retrospective database analysis. *BMC Public Health*, *21*(1734), 1734. Available from: https://doi.org/10.1186/s12889-021-11736-2
OECD. (2018) *Behavioural insights*. OECD. [Online] Available at: https://www.oecd.org/gov/regulatory-policy/behavioural-insights.htm [Accessed 02nd March 2022]
Oliver, A. (2019) Towards a new political economy of behavioral public policy. *Public Administration Review*, *79*(6), 917–924.
Pawson, R. & Tilley, N. (1994) What works in evaluation research? *The British Journal of Criminology*, *34*(3), 291–306.
Pawson, R. & Tilley, N. (1997) *Realistic evaluation*. UK: Sage.
Peer, E., Egelman, S., Harbach, M., Malkin, N., Mathur, A. & Frik, A. (2020) Nudge me right: personalizing online security nudges to people's decision-making styles. *Computers in Human Behavior*, *109*, e.106347.
Plans-Rubió, P. (2012) The vaccination coverage required to establish herd immunity against influenza viruses. *Preventive Medicine*, *55*, 72–77.
Rawls, J. (1971) *A theory of justice*. USA: Belknap Press.
Rebonato, R. (2012) *Taking liberties: a critical examination of libertarian paternalism*. UK: Palgrave Macmillan.
Rebonato, R. (2014) A critical assessment of libertarian paternalism. *Journal of Consumer Policy*, *37*, 357–396.
Reñosa, M.D.C., Landicho, J., Wachinger, J., Dalglish, S.L., Bärnighausen, K., Bärnighausen, T. et al. (2021) Nudging toward vaccination: a systematic review. *BMJ Globalization and Health*, *6*, e.006237.
Rizzo, M.J. & Whitman, G. (2020) *Escaping paternalism: rationality, behavioral economics, and public policy*. UK: Cambridge University Press.
Ruggeri, K. (2021) *Psychology and behavioral economics: applications for public policy*. UK: Routledge.
Ryan, S. (2018) Libertarian paternalism is hard paternalism. *Analysis*, *78*(1), 65–73.
Sanders, M., Snijders, V. & Hallsworth, M. (2018) Behavioural science and policy: where are we now and where are we going? *Behavioural Public Policy*, *2*(2), 144–167.
Schimmelpfennig, R. & Muthukrishna, M. (2022) *Cultural evolutionary behavioural science in public policy*. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4057679 [Accessed 10th July 2022]
Schultz, P.W., Messina, A., Tronu, G., Limas, E.F., Gupta, R. & Estrada, M. (2016) Personalized normative feedback and the moderating role of personal norms: a field experiment to reduce residential water consumption. *Environment and Behavior*, *48*(5), 686–710.
Schultz, P.W., Nolan, J.M., Cialdini, R.B., Goldstein, N.J. & Griskevicius, V. (2007) The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434.
Selinger, E. & Whyte, K.P. (2010) Competence and trust in choice architecture. *Knowledge, Technology and Policy*, *23*, 461–482.
Selinger, E. & Whyte, K.P. (2011) Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass*, *5*(10), 923–935.
Selinger, E. & Whyte, K.P. (2012) Nudging cannot solve complex policy problems. *European Journal of Risk Regulation*, *3*(1), 26–31.
Service, O. (2015) *Automatic enrolment and pensions: a behavioural success story*. BIT. [Online] Available at: https://www.bi.team/blogs/automatic-enrolment-and-pensions-a-behavioural-success-story/ [Accessed 07th July 2022]

Service, O., Hallsworth, M., Halpern, D., Algate, F., Gallagher, R., Nguyen, S. et al. (2015) *EAST: four simple ways to apply behavioural insights*. BIT. [Online] Available at: https://www.bi.team/wp-content/uploads/2015/07/BIT-Publication-EAST_FA_WEB.pdf [Accessed 07th July 2022]

Soman, D. & Hossain, T. (2021) Successfully scaled solutions need not be homogeneous. *Behavioural Public Policy*, *5*(1), 80–89.

Stern, P.C. (2020) A reexaminiation on how behavioral interventions can promote household action to limit climate change. *Nature Communications*, *11*(918), 1–3. Available from: https://doi.org/10.1038/s41467-020-14653-x

Sunstein, C.R. (2012) *Impersonal default rules vs. active choices vs. personalized default rules: a triptych*. SSRN. [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2171343 [Accessed 10th July 2022]

Sunstein, C.R. (2013) The Storrs lectures: behavioral economics and paternalism. *The Yale Law Journal*, *122*, 1826–1899.

Sunstein, C.R. (2014) *Why nudge? The politics of libertarian paternalism*. USA: Yale University Press.

Sunstein, C.R. (2017) Nudges that fail. *Behavioural Public Policy*, *1*(1), 4–25.

Sunstein, C.R. (2020) *Behavioral science and public policy*. UK: Cambridge University Press.

Sunstein, C.R. (2022) The distributional effects of nudges. *Nature Human Behaviour*, *6*, 9–10.

Sunstein, C.R., Reisch, L.A. & Rauber, J. (2017) A worldwide consensus on nudging? Not quite, but almost. *Regulation and Governance*, *12*(1), 3–22.

Thaler, R.H. & Sunstein, C.R. (2003) Libertarian paternalism. *American Economic Review*, *93*(2), 175–179.

Thaler, R.H. & Sunstein, C.R. (2008) *Nudge: improving decisions about health, wealth and happiness*. UK: Penguin Books.

Thunström, L. (2019) Welfare effects of nudges: the emotional tax of calorie menu labelling. *Judgment and Decision Making*, *14*(1), 11–25.

Thunström, L., Gilbert, B. & Jones-Ritten, C. (2018) Nudges that hurt those already hurting—distributional and unintended effects of salience nudges. *Journal of Economic Behavior and Organization*, *153*, 267–282.

Tor, A. (2020) Nudges that should fail? *Behavioural Public Policy*, *4*(3), 316–342.

UK Government. (n.d.) *Workplace pensions*. [Online] Available at: https://www.gov.uk/workplace-pensions/what-you-your-employer-and-the-government-pay [Accessed 07th July 2022]

van Belle, S., Wong, G., Westhorp, G., Pearson, M., Emmel, N., Manzano, A. et al. (2016) Can "realist" randomised controlled trials be genuinely realist? *Trials*, *17*(313), 313. Available from: https://doi.org/10.1186/s13063-016-1407-0

van der Linden, S. & Goldberg, M.H. (2020) Alternative meta-analysis of behavioral interventions to promote action on climate change yields different conclusions. *Nature Communications*, *11*(3915), 1–2.

van der Linden, S., Pearson, A.R. & van Boven, L. (2021) Behavioural climate policy. *Behavioural Public Policy*, *5*(4), 430–438. Available from: https://doi.org/10.1038/s41467-020-17613-7

Veetil, V.P. (2011) Libertarian paternalism is an oxymoron: an essay in defence of liberty. *European Journal of Law and Economics*, *31*, 321–334.

Werfel, S.H. (2017) Household behaviour crowds out support for climate change policy when sufficient progress is perceived. *Nature Climate Change*, *7*(7), 512–515.

World Health Organization. (2020) *Coronavirus disease (COVID-19): herd immunity, lockdowns and COVID-19*. World Health Organization. [Online] Available at: https://www.who.int/news-room/questions-and-answers/item/herd-immunity-lockdowns-and-covid-19 [Accessed 08th July 2022]

How to cite this article: Mills, S., & Whittle, R. (2023). Seeing the nudge from the trees: The 4S framework for evaluating nudges. *Public Administration*, 1–21. https://doi.org/10.1111/padm.12941