Behavioural Public Policy (2023), 7: 1, 25–54
© *Cambridge University Press* doi:10.1017/bpp.2020.6 First published online 07 May 2020

'Better off, as judged by themselves': do people support nudges as a method to change their own behavior?

NATALIE GOLD

Faculty of Philosophy, University of Oxford, Oxford, UK

YILING LIN

Biological and Experimental Psychology Group, School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

RICHARD ASHCROFT

City Law School, City, University of London, London, UK

MAGDA OSMAN*

Biological and Experimental Psychology Group, School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Abstract: In this study, we investigated how people evaluate behavioral interventions (BIs) that are targeted at themselves, aiming to promote their own health and wellbeing. We compared the impact on people's assessments of the acceptability of using BIs to change their own behavior of: the transparency of the BI (transparent or opaque); the designer of the BI (researchers, government policy-makers, advertisers); and three types of arguments regarding their efficacy (positive, positive + negative, negative). Our target BIs were actual interventions that have been used in a range of policy domains (diet, exercise, alcohol consumption, smoking, personal finances). We found that transparent BIs were considered more acceptable than opaque BIs. On average, all BIs were considered acceptable for changing participants' own behavior, except for the opaque BI in the finance context; there was differential acceptability of BIs across contexts, with finance clearly least acceptable. However, the perceived effectiveness of the BIs was at least as influential a predictor of acceptability ratings as the ease of identification of the behavior change mechanism across the five contexts. Furthermore, effectiveness was partially mediated by desire to

* Correspondence to: Magda Osman, Reader in Experimental Psychology, Head of Centre for Mind in Society (LSI), Turing Research Fellow, Research Fellow to the Government, Biological and Experimental Psychology Group, School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK. E-mail: m.osman@qmul.ac.uk

change, suggesting that people do think BIs make them better off, ‘as judged by themselves’.

Submitted 10 July 2019; revised 3 February 2019; accepted 19 February 2020

Introduction

Behavioral interventions (BIs) – sometimes called nudges – use behavioral science to generate a change in behavior without fundamentally changing the incentive structure of the context in which decisions are made (see Osman *et al.*, 2018; see also Oliver, 2013). BIs can be used for many ends (e.g., to conserve the environment, to get people to pay their taxes on time or to promote health and wellbeing). Examples of BIs in health and wellbeing include changing the default on pensions, so that a portion of an employee’s salary is put into retirement saving unless they opt out, and changing the size of glassware in pubs to encourage people to drink less.[1] All over the world, BIs are being used in public policy in domains including health, finance, consumer protection, education, energy, the environment, transport, taxation, telecommunications, public service delivery and the labor market (World Bank, 2015; OECD, 2017).

In some situations where BIs are used, people have a clear interest in the behavior of others. For instance, when we face a problem of social cooperation such as conserving the environment or a ‘negative externality’ (a cost incurred by a third party) such as second-hand smoke, then BIs can encourage people to behave pro-socially. But in some cases, BIs are supposed to promote the self-interest of the recipients, such as when implemented in the context of health behaviors. One prominent argument for using BIs that promote health and wellbeing is that they “make choosers better off, as judged by themselves” (Thaler & Sunstein, 2008, p. 5). This is an empirical claim that needs to be judged in the light of the evidence.

Recent work suggests that the majority of the public find BIs acceptable (Hagman *et al.*, 2015; Jung & Mellers, 2016; Petrescu *et al.*, 2016; Reisch & Sunstein, 2016; Reisch *et al.*, 2016; Osman *et al.*, 2018; Venema *et al.*, 2018). However, people may approve of BIs because they hope that they will change other people’s behavior. Studies show that people’s support for BIs is higher when they are given a justification of the policy in terms of its

[1] BIs are part of a broader behavioral insights approach, which integrates insights and methodologies from the behavioral and social sciences (including decision-making, psychology, cognitive science, neuroscience, organizational and group behavior) in order to deliver evidence-based public policy (OECD, 2017).

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

effects on 'people' in general rather than when they are given a justification in terms of its effects on 'you' (Cornwell & Krantz, 2014). People also think that BIs will be more effective for others than for themselves and their judgments of the acceptability of BIs are predicted by how effective they anticipate BIs will be on others' behavior, whereas the evidence is mixed as to whether acceptability judgments are predicted by how effective BIs will be at changing their own behavior (Bang *et al.*, 2018). In these cases, people may regard the ill health of others as imposing an externality on them through the economic costs of ill health, which may increase insurance premiums or may require increased government spending, especially in countries with socialized medicine (Gold, 2018). Alternatively, it may be that people have 'meddlesome preferences' – preferences about how other people behave in domains where everyone should be free to make their own decisions (Sen, 1970; Blau, 1975). Indeed, a systematic review of the acceptability of government interventions to change health-related behaviors found that support for the interventions was highest among those not engaging in the targeted behavior (Diepeveen, 2013).

In order to judge whether BIs make choosers better off, as judged by themselves, we need evidence that directly targets that claim. There is debate about how exactly to cash out the claim (Sugden, 2017, 2018; Sunstein, 2018), but a first start is to investigate whether people support BIs as a method to change their own behavior. Previous studies, which have asked in general terms whether BIs are acceptable, cannot distinguish whether people support them because they want to change their own behavior or because they want other people's behavior to be changed. The studies cited above (Diepeveen, 2013; Bang *et al.*, 2018) suggest that support is at least partly driven by a desire to change other people's behavior. Therefore, in this study, we investigate how people evaluate BIs that are targeted at promoting *their own* health and wellbeing, asking them how acceptable it is for BIs to be used to change *their own* behavior.

Previous studies

We build on previous empirical work on the factors that affect the acceptability of BIs.

Transparency matters

Previous work has consistently shown that people evaluate BIs more favorably when they are aware of the process that leads to behavioral change (Diepeveen *et al.*, 2013; Felsen, *et al.*, 2013; Jung & Mellers, 2016; Petrescu *et al.*, 2016; Reisch & Sunstein, 2016; Reisch *et al.*, 2016; Sunstein, 2016; Osman *et al.*, 2018). They prefer transparent BIs, where they can identify the mechanism

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

that is being used to influence their behavior, as opposed to opaque BIs, where they cannot identify the mechanism of behavioral change. We define transparency in terms of ease of identification of the mechanism underpinning the BI, which has been used by other researchers (e.g., Hansen & Jespersen, 2013; Bang *et al.*, 2018).[2] Another way of achieving transparency is via disclosure – telling people at the point of decision that BIs are being used to change their behavior. The two sorts of transparency are related because, as well as revealing the intended effect of the BI, full disclosure can include revealing the mechanism of behavior change. One explanation of the preference for transparency is that it enables people to maintain a sense of agency over the behavior being targeted by the BI (Osman, 2014). Free choice is underpinned by a sense of agency and so, relative to opaque interventions, if people know how a behavior change is achieved, then they feel that they can more easily choose to do otherwise, thus preserving their autonomy (Lin *et al.*, 2017; Osman *et al.*, 2017).

Designer matters

Previous research shows that people trust BIs that are developed and proposed by researchers more than those that are developed and proposed by government (Osman *et al.*, 2018). We also know that trust in government affects the acceptability of government interventions (Branson *et al.*, 2012), and it has been suggested that negative attitudes to BIs stem from mistrust in government (Jung & Mellors, 2016). In support of this conjecture, Bang *et al.* (2018) found that the acceptability of BIs depends on who designs and implements them (a friend being more acceptable than a government or corporate designer) and that these differences in acceptability were explained by perceived differences in the intention of the designer. Consistent with this story, Tannenbaum *et al.* (2017) found that people's support for a BI depended on whether they were told that it had been enforced by a policy-maker they supported or one they opposed (the Bush administration versus the Obama administration).

Expectations about effectiveness matter

Previous research shows that the acceptability of government interventions and BIs strongly depends on their expected effectiveness (Pechey *et al.*, 2014;

[2] Others have couched this distinction in terms of System 1 versus System 2 BIs (e.g., Jung & Mellors, 2016; Sunstein, 2016), but we prefer the terminology of 'transparent' and 'opaque' because it cuts up the BI space in the same way, but without any implication that there are two separate systems in the brain (Lin *et al.*, 2017).

Petrescu *et al.*, 2016) and that directly manipulating the effectiveness of BIs by quantifying the resulting change in behavior affects acceptability (Sunstein, 2016; Arad & Rubinstein, 2018). Therefore, it is not surprising that giving people positive arguments – telling them that BIs are likely to be effective – affects their evaluations of the BIs. Sunstein (2016) found that, although people prefer transparent BIs to opaque ones, telling people that opaque BIs were more effective shifted their preferences toward opaque BIs by approximately 12% from baseline. However, to date, there has been no work investigating the impact of negative arguments – telling people about the possible backfire effects of the intervention – even though, outside of the laboratory, discussions about the effectiveness of BIs are more likely to be put in terms of general arguments for and against than to have precise quantifications attached.

Present study

In the present study, we compared the impacts on people's assessments of the acceptability of using BIs to change their own behavior of: the transparency of the BI (Transparent, Opaque); the designer of the BI (Researchers, Government, Advertisers); and three types of arguments regarding their efficacy (Positive, Positive + Negative, Negative). We tested the following hypotheses:

> *Hypothesis 1*: Transparent BIs will be more acceptable than opaque BIs.
>
> *Hypothesis 2*: The designer of the BI will affect the acceptability of the BI.
>
> *Hypothesis 3*: The type of argument given will affect the acceptability of the BI.
>
> *Hypothesis 4*: There will be an interaction effect between the designer of the BI and the type of argument given.

The rationale for Hypothesis 4 is that the more ambiguous the outcome of the BI, or the more salient the possible backfire effects, the more important the trustworthiness of the designer will be. Information about possible negative effects could cause people to doubt either the expertise or the intentions of the designer.

We expect that the transparency of a BI and its perceived likelihood of being effective are both factors that explain the acceptability of a BI and that these are mediated by a desire to change one's behavior through transparent and effective methods. Therefore, in order to discover the relative weight given to

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

transparency and effectiveness and to test for mediation by desire to change behavior, we asked participants to rate the perceived transparency and effectiveness of each BI and their desire to have their behavior changed by that method. We also used the transparency and effectiveness ratings as manipulation checks.

In order to establish the generalizability of any results, we used five different contexts in which BIs have been implemented to promote health and wellbeing – exercise, diet, smoking, alcohol and finance – and compared acceptability across contexts. All of the interventions we showed participants were genuine interventions that have been implemented by policy-makers.

Methods

Design

We used a mixed factorial design with one within-subject factor and three between-subject factors to give a 5 × 2 × 3 × 3 design. The within-subject manipulation was 5 different contexts in which a BI was implemented (Exercise, Diet, Smoking, Alcohol, Finance). The between-subject manipulations were: 2 Transparency of the BI (Transparent, Opaque) × 3 Designer of the BI (Researcher, Government, Advertiser) × 3 Argument about the likely effectiveness of the BI (Positive [Experiment 1a], Positive + Negative [Experiment 1b], Negative [Experiment 1c]). Experiments 1a, 1b and 1c were run serially (no one participated in more than one experiment); in each experiment, participants were randomly allocated to one of the 6 between-subject conditions.

For each context, there were four probative questions regarding BIs. After responding to the probes in all five contexts, participants were asked five demographic questions – about their age, sex, education, political affiliation and religion – and whether they were a smoker. At the end of the experiment, they were also asked some questions about their attitudes to BIs, including to indicate which contexts should not involve psychological methods designed to change behavior.

All experiments were presented via Qualtrics, which is an online platform for running experiments, and launched via Prolific Academic, a crowdsourcing system for participant recruitment of those who have university email addresses, including large pools in the USA and UK, both of which we used. All participants were financially compensated for their time (90 cents), calculated according to Prolific Academic's rates.

The Queen Mary University of London college ethics board granted ethical approval for the experiments under the project titled 'Ethical concerns around nudges' (QMERC2014/54).

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Participants

Each experiment included US (total $n = 872$) and UK samples (total $n = 843$) (see Table 1). Although Experiments 1a, 1b and 1c were run serially, they drew from the same population and there were no differences in demographics (see Supplementary Appendix S1, available online, for details), so we have combined them for the analysis.[3] Participants who took much less or more time to complete the task than the allocated time (less than 8 or more than 30 minutes) were excluded. Four further participants failed an attention check question and were removed from the analysis.

Procedure

After consenting to take part in the experiment, all participants were told that they were going to be asked questions about psychological methods that have been used to bring about behavior change and that "All of these methods are designed to help guide people to make the best decision for their own health and wellbeing." For full instructions, see Supplementary Appendix S2.

Manipulation of designer

Participants were told the following

> The [**Top Advertising Company, Government, Top Researchers in Laboratories**] in this country **is/are** using psychological research to help develop a set of simple methods that adjust the way information is presented, so that it can help people to make better decisions. The reason for using psychological methods is to help improve people's behavior, because in many day-to-day contexts people may not make a decision that is best for their own health, wellbeing and their happiness.

BIs and manipulation of transparency

Participants were then presented with a description of a BI. For each context, there were two BIs – one that was transparent and one that was opaque – all based on genuine interventions that have been implemented. Participants were randomly assigned to receive descriptions of either five transparent BIs or five opaque BIs. (The ten BIs are given in Table 2.) For each context, first, they were told what the context was in which the method would be used

[3] Specifically, there were no statistically significant differences between experiments for age, sex, religion and education. There were differences in political affiliation between experiments, but those differences were not significant predictors of acceptability ratings, including as interaction terms in a multivariate analysis of variance (ANOVA). Full details are in Supplementary Appendix S1.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Table 1. Participant profiles from Experiments 1a, 2b and 3c combined.

Sample	USA	UK
Total participants	*n* = 872 (all US residents, US nationals, first language English)	*n* = 843 (all UK residents, UK nationals, first language English)
Females	471 (54%)	413 (49%)
Age (years)	Mean 35 (*SD* = 12.24) ranging from 18 to 74	Mean 32.36 (*SD* = 11.32) ranging from 18 to 71
Educational background	Mixed, 56.7% qualified with a degree (at least a bachelor's degree, maybe postgraduate qualification as well)	Mixed, 57.1% qualified with a degree (at least a bachelor's degree, maybe postgraduate qualification as well)
Political affiliation	51.6% identifying as left, 8.6% as center, 16.7% as right and 23.1% as other	47.0% identifying as left, 16.1% as center, 17% as right and 19.9% as other
Religion	38.9% reported that they did not have one, 5.4% reported that they were not sure and 55.7% reported that they were religious	37.2% reported that they did not have one, 9.5% reported that they were not sure and 53.3% reported that they were religious
Smokers	135 (15.5%) smokers, 10 (1.4%) prefer not to say	112 (13.3%) smokers, 12 (1.3%) prefer not to say

(e.g., 'Smoking'), and what the Recommended Psychological Method was (e.g., "Design cigarette packaging so that it incorporates graphic pictures of damaged lungs and warnings such as 'Smoking seriously harms you and others around you', 'Smoking harms your unborn baby'"). The order of presentation of the five contexts was randomized for each participant.

Arguments and manipulation of effectiveness

In Experiments 1a and 1b, participants were then presented with:

> **Argument for method to work:** By highlighting the negative physical and moral issues concerning smoking, the negative experiences will become more obviously associated with smoking, and this will encourage smokers to reduce or even stop smoking.

In Experiments 1b and 1c, participants were presented with:

> **Argument for method NOT to work:** By highlighting the negative physical and moral issues concerning smoking, the negative experiences will become so obviously associated with smoking that smokers will feel more defensive of their smoking habit, as a result, smokers will end up smoking more, meaning that the method will lead to increases in smoking.

A full list of the arguments for each context can be found in Table 3.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Table 2. Recommended Psychological Method: description of transparent and opaque behavioral interventions.

Context	Transparent version	Opaque version
Exercise	Design stairwells with 'point-of-choice' signage that displays messages about the health advantages of taking the stairs, such as "Stair climbing burns more calories per minute than tennis," "7 minutes of stair climbing per day protects your heart," etc.	Design stairwells by hanging artworks. Pictures are changed periodically to keep stair users interested in order to prolong effectiveness.
Diet	Design packaging on food so that the front label includes nutritional information by using a simple traffic light system (red, amber, green) to indicate how much saturated fat, salt and sugar, and calories are in food products.	Design the size of plates so that the quantity of food on them is adjusted. Large plates and bowls can make servings of food appear smaller, whereas smaller plates can lead people to misjudge that very same quantity of food as being significantly larger.
Smoking	Design cigarette packaging so that it incorporates graphic pictures of damaged lungs and warnings such as "Smoking seriously harms you and others around you," "Smoking harms your unborn baby."	Increasing the length of the filter by 10 mm and at the same time reducing the length of the cigarette to 60 mm.
Alcohol	Design signage in pubs and restaurants so that they include messages such as the following: "Men and women are advised not to regularly drink more than 14 units a week," and "Spread your drinking over three days or more if you drink as much as 14 units a week."	Design the glassware used in pubs and restaurants in such a way so that straight glasses are used because, relative to curvy glasses, it is easier to judge and pace the amount of alcohol consumed.
Finance	Design investment schemes in such a way so that customers can evaluate the associated riskiness of each product based on a traffic light system; red indicates highly risky, green indicates low risk.	Design investment schemes with an automatic enrollment system so that the bank/building society will decide on an individual's behalf exactly how the money will be allocated to investment schemes. Although if the individual did not want it, they could opt out of the scheme, but this would involve filling in relevant paperwork.

Explanation of transparency

Before the first probe, all participants were provided with the following definition of transparent and opaque BIs:

> There are two types of psychological methods: transparent and non-transparent. A transparent psychological method works in such a way that anyone

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Table 3. Positive and negative arguments for each behavioral intervention.

	Argument for it to work		Argument for it not to work	
Context	Transparent version	Opaque version	Transparent version	Opaque version
Exercise	By presenting messages at strategic positions, people will be encouraged to use stairs instead of lifts or escalators/elevators, and this in turn will encourage people to value being more active and, in turn, exercise more in general.	By presenting artworks along the stairwell, this is more likely to encourage people to use stairs instead of lifts or escalators/elevators, and this in turn will encourage people to value being more active and, in turn, exercise more in general.	By presenting messages at strategic positions, people will avoid the stairs and hence the messages for the reason that they do not want to feel guilty about not exercising enough, meaning that this method will lead people to be less active.	The artworks along the stairwell are not changed regularly enough, and people get bored looking at them, and so to avoid looking at them most people end up taking the lift and, in turn, get less exercise overall.
Diet	By making it easier for people to interpret the nutritional content of food items through a traffic light labeling system, people will be more aware of which foods are healthier than others, and in turn adopt/maintain a healthier diet.	By making the plates smaller, people would be better able to adjust the amount of food they put on their plate and avoid overconsumption of food, and in turn adopt/maintain a healthier diet.	By making it easier for people to interpret the nutritional content of food items, people change their eating habits and as a result consume more food to compensate for eating healthily, meaning that this method increases people's overall daily calorie intake.	By making the plates smaller, people change their eating patterns and as a result consume more food to compensate for the smaller plate, meaning that this method will lead to increases in people's overall daily calorie intake.
Smoking	By highlighting the negative physical and moral issues concerning smoking, the negative experiences will become more obviously associated with smoking, and this will encourage smokers to reduce or even stop smoking.	If a standard cigarette is 70 mm, then cutting back on the harmful chemicals and replacing them with more filter would make it seem as if the size of cigarette had not changed, but the amount of harmful chemicals would be reduced. By reducing nicotine content adequately, this method helps smokers gradually adapt to lower nicotine levels, and this will encourage smokers to reduce or even stop smoking.	By highlighting the negative physical and moral issues concerning smoking, smokers will feel more defensive of their smoking habit, and as a result, smokers will end up smoking more, meaning that the method will lead to increases in smoking.	By reducing the length of the cigarette and cutting back on the nicotine content, people will change their smoking habits and as a result smoke more to compensate for the shorter cigarette, meaning that this method will lead to people increasing their consumption of cigarettes.

Alcohol	By informing people about the actual appropriate amount of alcohol consumption that is reasonable to be consumed in a typical week, people will be more aware of exceeding the limit, and this should in turn reduce alcohol overconsumption.	By changing the shape of the containers that are used to serve alcohol, this will in turn reduce the actual amount of alcohol consumed at any one sitting, and this should in turn reduce alcohol overconsumption.	By informing people about the actual appropriate amount of alcohol consumption that is reasonable to be consumed in a typical week, those who drink lightly will consume more alcohol, as they will believe it is safe to drink 14 units a week, meaning that this method will increase overall alcohol consumption.	By changing the shape of the containers that are used to serve alcohol, people will change their drinking habits, and as a result consume more alcohol to compensate for the smaller container, meaning that this method will increase people's overall alcohol consumption.
Finance	By making it easier for people to interpret the riskiness of an investment scheme through a traffic light labeling system, people will be more aware of which financial products are risker than others, and in turn this will help them to make better financial decisions.	Because people find it difficult to think about their future financial status, making the investment schemes a default would encourage people to invest their savings, and in turn this would help them make better financial decisions.	By making it easier for people to interpret the riskiness of an investment scheme through a traffic light labeling system, the method highlights the potential financial gains through risky choices, meaning that it will lead people to taking more gambles with their money and be worse off in the long run.	The default investment scheme does not take into account the fact that people have different needs because their lifestyles are different, and as a result the scheme means that, in the long run, people will end up saving less overall.

can easily identify the actual psychological method used to change their behavior, as well as easily identify how their behavior is changed by it. A non-transparent psychological method works in such a way that no one can identify the actual psychological method used to change their behavior, and no one can identify how their behavior is changed by it.

Probes

For each of the five BIs, all participants were required to respond to four questions concerning:

(1) **Ease of identification**
To what extent is it easy for you to identify HOW your behavior is going to be changed by the psychological method? (Scale **0** = I cannot easily identify how my behavior is changed by the psychological method *to* **100** = I can easily identify how my behavior is changed by the psychological method)

(2) **Desire to change behavior**
To what extent do you want to change your behavior through the psychological method in this particular situation? (Scale 1 = Not at all likely *to* 9 = Very likely)

(3) **Effectiveness**
To what extent do you think the psychological method used in this particular situation would positively change YOUR behavior? (Scale 1 = Not at all likely *to* 9 = Very likely)

(4) **Acceptability**
To what extent do you think it is acceptable to use the psychological method described in this context to change your behavior? (Scale 1 = Unacceptable *to* 9 = Acceptable)

Once participants had responded to all four questions for each of the five scenarios, they were presented with five demographic questions about their age, sex, education level, political affiliation and religion, and asked whether they smoked or not (or preferred not to say).

In addition, we asked several exploratory questions, most of which we did not analyze since they do not bear directly on our hypotheses. Participants were asked: about the extent to which each BI would lead to positive changes in behavior in the population; whether they think that there are ethical issues concerning each BI and, if so, what they are; and some questions about how they value their health and wellbeing. Full instructions are available in Supplementary Appendix S2.

We have analyzed one of our exploratory questions because it was relevant given our results. At the end of the experiment, participants were asked to indicate which contexts *should not* involve psychological methods designed to change behavior, with the following possible responses (they could choose as

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

many or as few as they wanted, so a participant may have chosen multiple options):

- □ Food and nutrition decisions concerning food, drink and nutritional intake
- □ Smoking decisions to quit smoking
- □ Alcohol decisions to reduce drinking
- □ Exercise decisions to increase levels of physical activity
- □ Financial decisions concerning investment
- □ Financial decisions concerning savings
- □ NONE of the five contexts should have psychological methods used to influence decision-making behavior
- □ I cannot decide
- □ All of the contexts should have psychological methods implemented to influence decision-making behavior

We analyzed a second of our exploratory questions at the request of an anonymous reviewer. For each of the five BIs, we asked:

> *Based on the psychological method used in this particular situation, to what extent do you think it would lead to positive changes in behavior IN THE POPULATION?* (Scale 1 = Not at all likely *to* 9 = Very likely)

Results

Manipulation check: Ease of Identification of behavior change method ratings

In order to test that our manipulation of transparency worked, we ran a four-way mixed multivariate analysis of variance (ANOVA) with the Ease of Identification ratings in the five contexts as the dependent variables, Context as the within-subject variable and the elements of the factorial design supplying the between-subject independent variables (including interaction effects).

Tests of between-subject effects showed that our manipulation affected Ease of Identification across all contexts combined. As we expected, given that we manipulated transparency, there was a small to medium main effect of Ease of Identification, $F(1, 1697) = 163.70$, $p < 0.001$, partial $\eta^2 = 0.88$, with the Ease of Identification of Transparent BIs ($M = 69.21$, $SE = 0.57$) being higher than Opaque ones ($M = 58.85$, $SE = 0.57$). There were no other statistically significant effects. The full between-subjects model is given in Supplementary Appendix S3.

Multivariate tests showed that the difference in Ease of Identification between Opaque and Transparent BIs was also found in most of the individual contexts. There was a small to medium interaction effect between Ease of

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Identification and Transparency, *F*(4, 1694) = 38.31, $p < 0.001$, Wilks' $\Lambda = 0.114$, partial $\eta^2 = 0.012$. *Post hoc* pairwise comparisons revealed that the Transparent BIs all had higher Ease of Identification ratings than the Opaque ones in all contexts except Diet (all $p < 0.001$, except Diet, which was $p = 0.144$; see Figure 1). The full set of multivariate tests are given in Supplementary Appendix S3, as are the results of the less powerful within-subject tests, which show the same pattern of effects.

Manipulation check: Effectiveness ratings

In order to test whether giving different arguments affected the perceived effectiveness of the BIs, we ran a four-way mixed multivariate ANOVA with the Effectiveness ratings in the five contexts as the dependent variables, Context as the within-subject variable and the elements of the factorial design supplying the between-subject independent variables (including interaction effects).

The between-subjects tests showed that our manipulations failed to have the desired effect on Effectiveness across all contexts combined. There was only a very small main effect of Argument *F*(1, 1697) = 61.66, $p = 0.006$, partial $\eta^2 = 0.006$, and a very small effect of Designer, *F*(1, 1697) = 3.66, $p = 0.026$, partial $\eta^2 = 0.004$. Surprisingly, the largest main effect was the small main effect of Transparency, *F*(1, 1697) = 86.25, $p < 0.001$, partial $\eta^2 = 0.048$, on Effectiveness ratings. There were no significant interaction effects. See Supplementary Appendix S3 for the full model.

Multivariate tests showed that the effect of Transparency on Effectiveness was also seen in the individual contexts. There was a medium-sized interaction effect between Effectiveness and Transparency, *F*(4, 1694) = 38.98, $p < 0.001$, Wilks' $\Lambda = 0.916$, partial $\eta^2 = 0.084$. *Post hoc* pairwise comparisons revealed that the Transparent BIs were rated as more likely to be effective than the Opaque ones in all contexts except Alcohol (all $p < 0.001$, except Diet, which was $p = 0.02$, and Alcohol, which was $p = 0.267$; see Figure 2). The full set of multivariate tests are given in Supplementary Appendix S3, as are the results of the less powerful within-subject tests, which showed the same pattern of effects.

Hypothesis testing: Acceptability ratings

We ran a four-way mixed multivariate ANOVA with the participants' Acceptability ratings in the five contexts as the dependent variables, Context as the within-subject variable and the elements of the factorial design supplying the between-subject independent variables (including interaction effects). We used between-subject tests to examine our hypotheses across the five contexts

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

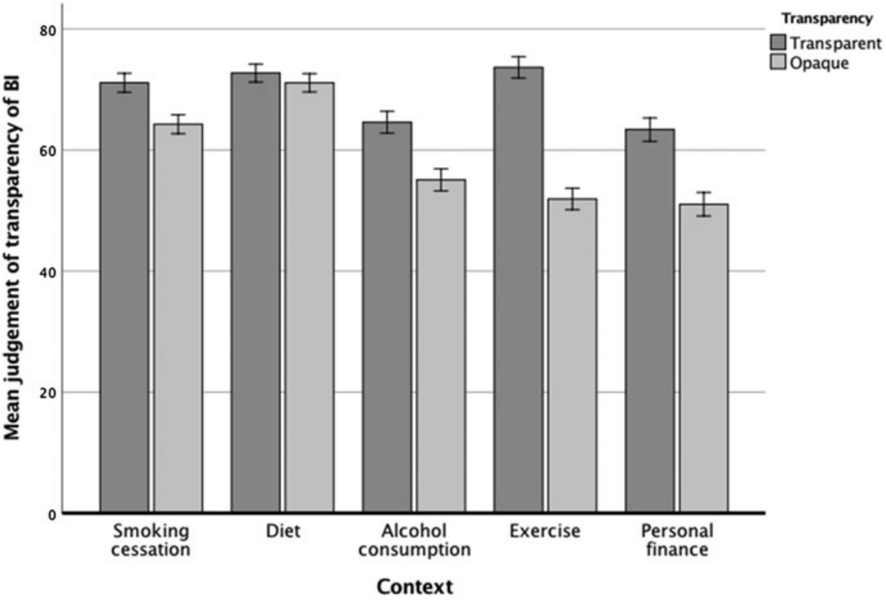

Figure 1. Comparison of the Ease of Identification ratings of transparent and opaque behavioral interventions (BIs) in the five Contexts.

combined and multivariate tests to investigate whether the results held in each context considered individually.

Between-subject tests

There was a medium-sized main effect of Transparency, $F(1, 1697) = 248.08$, $p < 0.001$, partial $\eta^2 = 0.128$, with the acceptability of Transparent BIs ($M = 6.96$, $SE = 0.05$) being higher than Opaque BIs ($M = 5.86$, $SE = 0.05$). This supports Hypothesis 1 that Transparent BIs will be more acceptable than Opaque BIs.

There was a significant but negligible main effect of Designer $F(1, 1697) = 3.60$, $p = 0.028$, partial $\eta^2 = 0.004$: Researchers $M = 6.53$, $SE = 0.06$; Advertisers $M = 6.38$, $SE = 0.06$; Government $M = 6.31$, $SE = 0.06$. This technically supports Hypothesis 2 that the designer of the BI will affect its acceptability, but the effect size is not meaningful.

There was a small main effect of Argument, $F(2, 1697) = 10.52$, $p < 0.001$, partial $\eta^2 = 0.012$. *Post hoc* pairwise comparisons revealed that the effect of Argument was due to the mean acceptability being higher for Positive ($M = 6.64$, $SE = 0.064$) than for Positive + Negative ($M = 6.33$, $SE = 0.058$) and Negative ($M = 6.26$, $SE = 0.058$; both $p < 0.001$ and well under the Bonferroni-adjusted

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

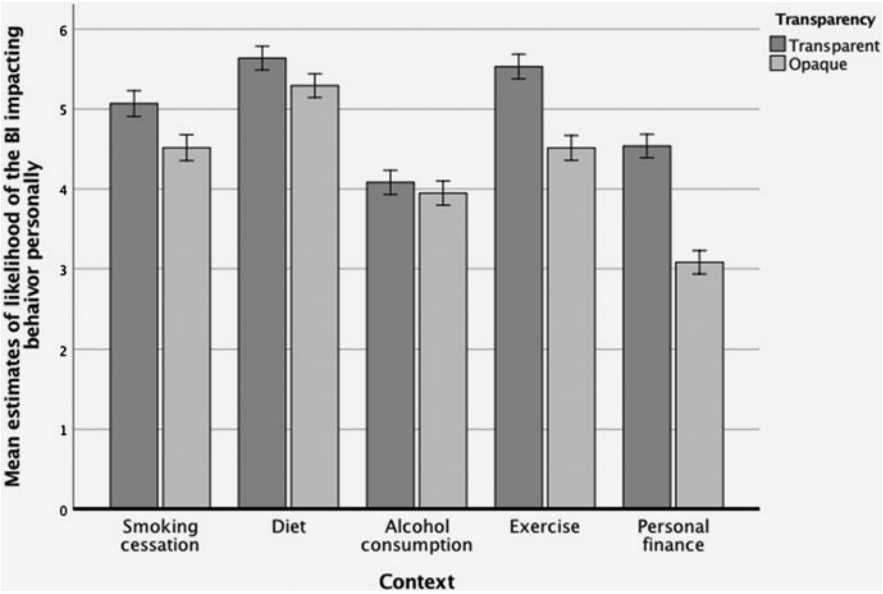

Figure 2. Comparison of the Effectiveness ratings of transparent and opaque behavioral interventions (BIs) in the five Contexts.

significance level of p = 0.017), but there was no significant difference between acceptability for Positive + Negative and Negative (p = 0.37). This supports Hypothesis 3 that the arguments will affect the acceptability of the BI.

There were no significant interaction effects, so Hypothesis 4 – that there will be an interaction effect between the designer of the BI and the type of argument – was not supported. See Supplementary Appendix S3 for the full model.

Multivariate tests

Multivariate tests, exploring our within-subject variable, showed the differential acceptability of BIs in the five contexts. There was a large main effect of Context on Acceptability, $F(4, 1694) = 423.40$, $p < 0.001$, Wilks' $\Lambda = 0.495$, partial $\eta^2 = 0.505$, suggesting that there were differences in Acceptability between at least one pair of contexts. *Post hoc* pairwise comparisons revealed that the means of the Acceptability ratings in each context differed (all $p < 0.001$, lower than the Bonferroni-adjusted significance level of $p = 0.005$), except for the means of Smoking and Alcohol ($p = 0.71$). The BIs were most acceptable in the context of Exercise ($M = 7.28$, $SE = 0.048$), followed by Diet ($M = 6.97$, $SE = 0.046$), Smoking ($M = 6.55$, $SE = 0.051$), Alcohol ($M = 6.53$, $SE = 0.052$) and Finance ($M = 4.72$, $SE = 0.054$).

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

There was a large interaction effect between Acceptability and Transparency, $F(4, 1694) = 4122.94$, $p < 0.001$, Wilks' $\Lambda = 0.777$, partial $\eta^2 = 0.225$. *Post hoc* pairwise comparisons revealed that the Transparent BIs were rated as more acceptable than the Opaque BIs in all contexts except Exercise (all $p < 0.001$, except Exercise, which, at $p = 0.027$, was more than the Bonferroni-adjusted significance level of $p = 0.01$; see Figure 3).

There was also a small interaction effect between Acceptability and Argument, $F(8, 3388) = 4.44$, $p < 0.001$, Wilks' $\Lambda = 0.979$, partial $\eta^2 = 0.010$, and a very small three-way interaction between Acceptability, Argument and Transparency, $F(8, 3388) = 2.91$, $p < 0.003$, Wilks' $\Lambda = 0.986$, partial $\eta^2 = 0.007$ (see Figure 4). *Post hoc* tests and means can be found in Supplementary Appendix S3.

There were no significant effects of the Designer of the BI (see Figure 5).

The results of within-subject effects tests confirmed the results of the multivariate tests. The full set of multivariate tests and within-subject tests can be found in Supplementary Appendix S3.

Predictors of Acceptability ratings

We ran regressions to discover the best predictors of Acceptability judgments in each context using standardized coefficients in order to be able to meaningfully compare effect sizes. For the Smoking context, we also ran a set of regressions that were limited to the participants who said they were smokers, since smokers are a small proportion of the population, less than 20% in the UK (ONS, 2018). The models are given in Table 4. For each context, Model 1 had Ease of Identification ratings as the sole predictor, Model 2 had Effectiveness as the sole predictor, Model 3 had Desire to Change Behavior as the sole predictor, Model 4 had both Ease of Identification and Effectiveness as predictors and Model 5 had all three of Ease of Identification, Effectiveness and Desire to Change Behavior as predictors.

There are clear patterns that held across the five contexts. All three ratings were significant predictors of Acceptability ratings when entered into the model separately (Models 1–3, Table 4), except that Ease of Identification was not a significant predictor amongst smokers for the acceptability of BIs for Smoking in any of the models.

Comparing the predictive power of Ease of Identification and Effectiveness by entering them both in Model 4, we can see that, across contexts, Effectiveness had a bigger effect on Acceptability than Ease of Identification (except for Exercise, where they were approximately equal with $\beta = 0.230$ for Ease of Identification and $\beta = 0.228$ for Effectiveness, both $p < 0.001$). The

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

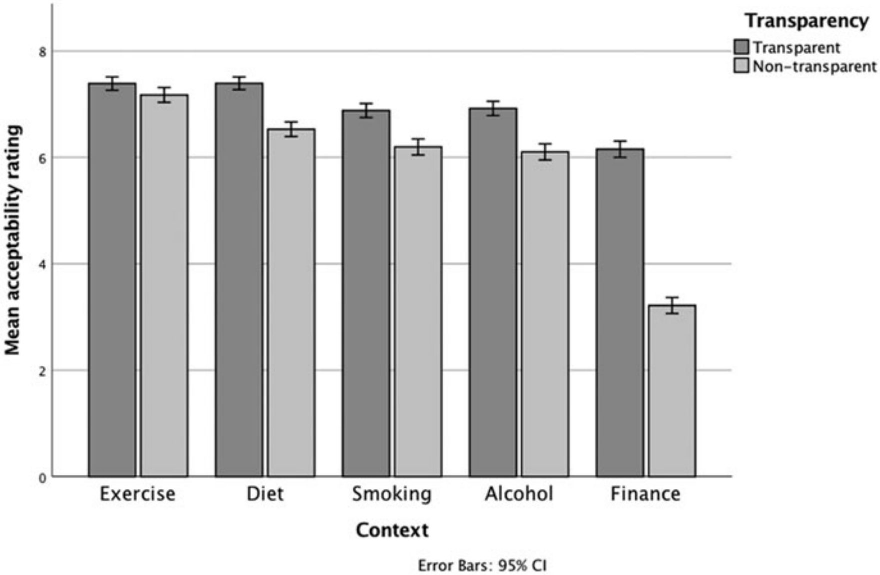

Figure 3. Comparison of the Acceptability ratings of transparent and opaque behavioral interventions (BIs) in the five Contexts.

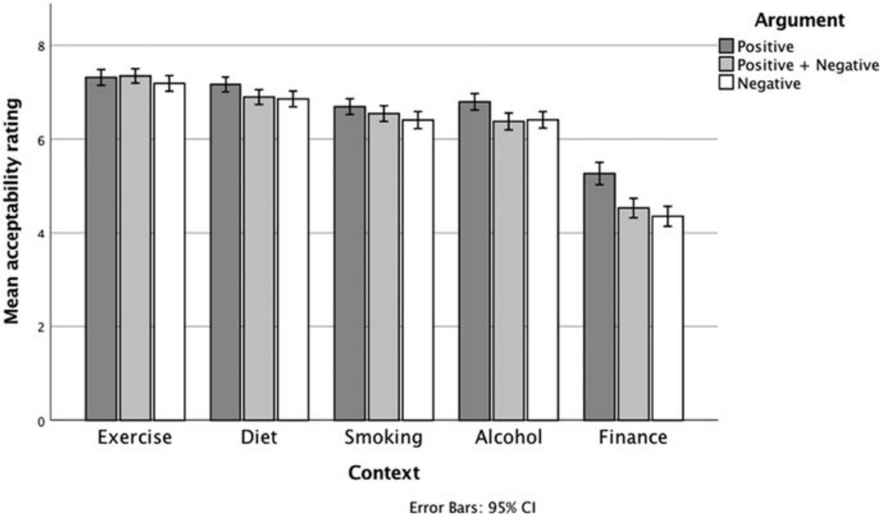

Figure 4. Comparison of the Acceptability ratings depending on which Arguments were given in each of the five Contexts.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

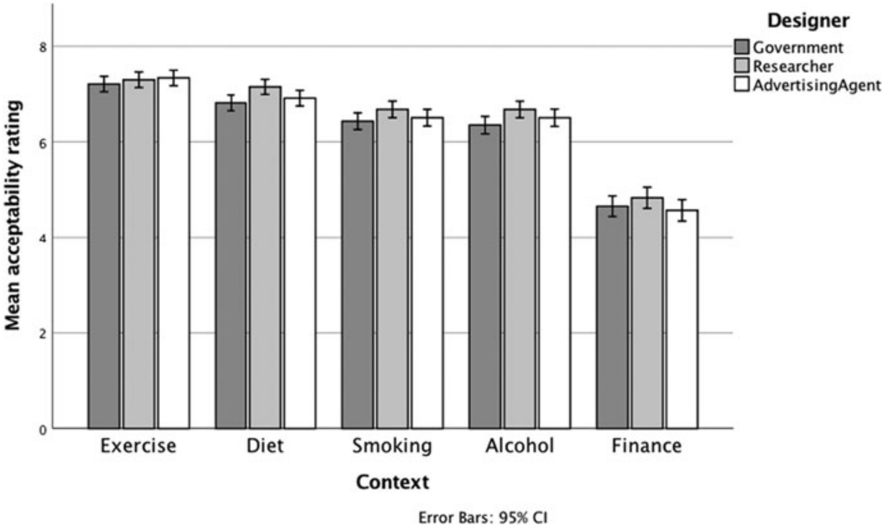

Figure 5. Comparison of the Acceptability ratings depending on the Designer of the behavioral intervention in each of the five Contexts.

largest differences were found in the Finance context, where the coefficient on Effectiveness was more than four times larger than that on Ease of Identification ($\beta = 0.140$ for Ease of Identification, $\beta = 0.597$ for Effectiveness, both $p < 0.001$), and amongst smokers in the Smoking context ($\beta = 0.003$, $p = 0.963$ for Ease of Identification, $\beta = 0.463$, $p < 0.001$ for Effectiveness).

However, when we added Desire to Change Behavior into the models (Model 5), the coefficient on Effectiveness clearly decreased. In the case of Exercise, Effectiveness even became non-significant, $\beta = 0.037$, $p = 0.295$. This suggests that Desire to Change Behavior wholly mediates Effectiveness for Exercise and partially mediates Effectiveness in the other four contexts. We confirmed this by testing the remaining step for mediation (Baron & Kenny, 1986): regressing Effectiveness on Desire to Change Behavior. The models in Table 5 show that Effectiveness was a significant predictor of Desire to Change Behavior in all five contexts, confirming that Desire to Change Behavior partially mediated Effectiveness.

Multiple-choice question on the use of BIs in different contexts

BIs were clearly less acceptable in financial contexts, with 63.0% of participants saying that they should not be used for financial decisions involving investment and 53.4% saying that they should not be used for financial

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Table 4. Regressions showing the predictors of Acceptability ratings in each context. Five models for each context, with subscript indicating the context.

Context	Ease of Identification β (standardized)	Effectiveness β (standardized)	Desire β (standardized)	Adjusted R^2
Exercise				
Model 1_E	0.325, $p < 0.001$			0.105
Model 2_E		0.324, $p < 0.001$		0.105
Model 3_E			0.372, $p < 0.001$	0.138
Model 4_E	0.230, $p < 0.001$	0.228, $p < 0.001$		0.148
Model 5_E	0.203, $p < 0.001$	0.037, $p = 0.295$	0.259, $p < 0.001$	0.173
Diet				
Model 1_D	0.280, $p < 0.001$			0.078
Model 2_D		0.463, $p < 0.001$		0.213
Model 3_D			0.473, $p < 0.001$	0.223
Model 4_D	0.161, $p < 0.001$	0.414, $p < 0.001$		0.237
Model 5_D	0.141, $p < 0.001$	0.218, $p < 0.001$	0.259, $p < 0.001$	0.262
Smoking				
Model 1_S	0.154, $p < 0.001$			0.023
Model 2_S		0.375, $p < 0.001$		0.140
Model 3_S			0.281, $p < 0.001$	0.079
Model 4_S	0.112, $p < 0.001$	0.362, $p < 0.001$		0.152
Model 5_S	0.111, $p < 0.001$	0.375, $p < 0.001$	0.076, $p = 0.008$	0.155
Smoking – smokers only				
Model 1_{SS}	0.072, $p = 0.261$			0.001
Model 2_{SS}		0.463, $p < 0.001$		0.212
Model 3_{SS}			0.416, $p < 0.001$	0.170
Model 4_{SS}	0.003, $p = 0.963$	0.463, $p < 0.001$		0.208
Model 5_{SS}	0.002, $p = 0.978$	0.347, $p < 0.001$	0.155, $p = 0.071$	0.216
Alcohol				
Model 1_A	0.270, $p < 0.001$			0.072
Model 2_A		0.285, $p < 0.001$		0.081
Model 3_A			0.282, $p < 0.001$	0.079
Model 4_A	0.208, $p < 0.001$	0.229, $p < 0.001$		0.120
Model 5_A	0.198, $p < 0.001$	0.138, $p < 0.001$	0.125, $p < 0.001$	0.127
Finance				
Model 1_F	0.330, $p < 0.001$			0.108
Model 2_F		0.642, $p < 0.001$		0.411
Model 3_F			0.623, $p < 0.001$	0.388
Model 4_F	0.140, $p < 0.001$	0.597, $p < 0.001$		0.429
Model 5_F	0.124, $p < 0.001$	0.384, $p < 0.001$	0.258, $p < 0.001$	0.447

Table 5. Effectiveness as a predictor of Desire to Change Behavior.

Context	β (standardized)	P-value	Adjusted R^2
Exercise	0.781	<0.001	0.610
Diet	0.787	<0.001	0.619
Smoking	0.621	<0.001	0.385
Alcohol	0.745	<0.001	0.555
Finance	0.846	<0.001	0.716

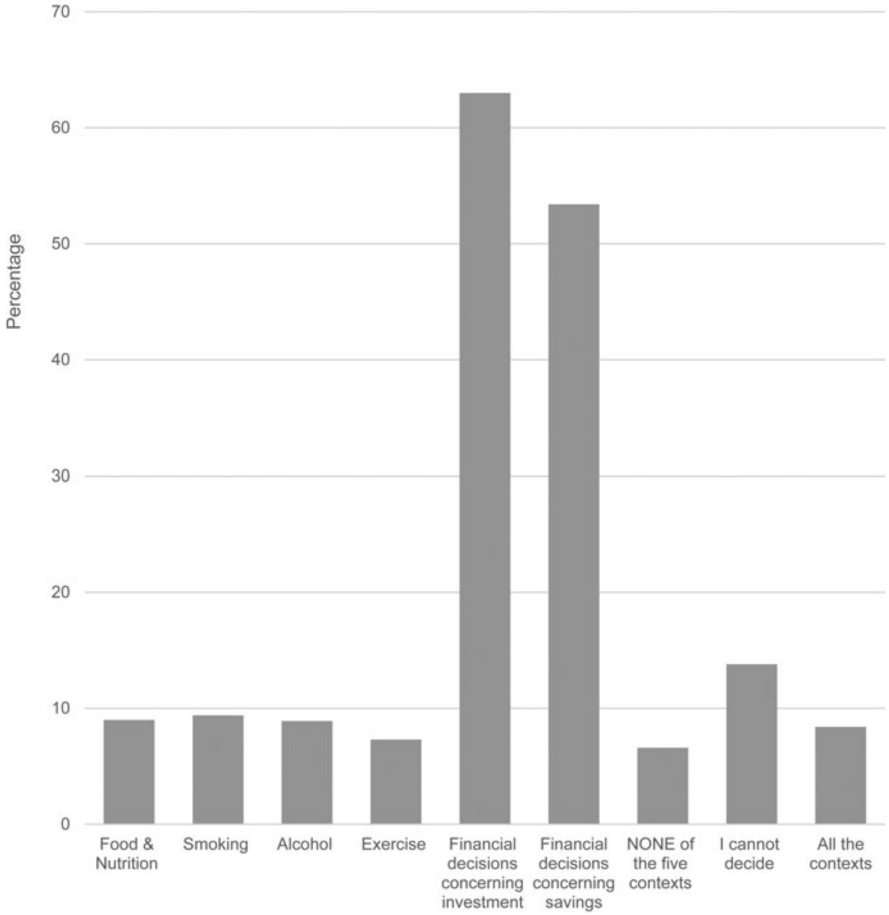

Figure 6. Percentages who indicated that context(s) *should not* involve psychological methods designed to change people's behavior.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

decisions involving savings. The next largest group was the 13.8% who could not decide, and all the other answers were chosen by less than 10% (see Figure 6).

Perceived effectiveness of BIs on self compared to effectiveness of BIs on the population

The BI was judged likely to be more effective on the population's behavior than on the participant's own behavior for four out of our five BIs, as shown by paired *t*-tests: Diet, population behavior *M* = 5.64, *SD* = 1.8, own behavior *M* = 5.47, *SD* = 2.23, *t*(1714) = 3.9, p < 0.001; Smoking, population behavior *M* = 5.24, *SD* = 1.93, own behavior *M* = 4.79, *SD* = 2.44, *t*(1714) = 8.7, p < 0.001; Alcohol, population behavior *M* = 4.27, *SD* = 1.95, own behavior *M* = 4.01, *SD* = 2.26, *t*(1714) = 5.8, p < 0.001; Finance, population behavior *M* = 4.20, *SD* = 2.07, own behavior *M* = 4.79, *SD* = 2.32, *t*(1714) = 10.4, p < 0.001. For Exercise, the difference was in the other direction, with participants judging the BI as less likely to affect population behavior (*M* = 4.78, *SD* = 2.00) than their own behavior (*M* = 5.02, *SD* = 2.38), *t*(1714) = –6.2, p < 0.001. We investigated this difference further by running multivariate ANOVAs for each domain with the Effectiveness on own behavior and on population behavior as the dependent variables (so own–population behavior was the within-subject variable) and Transparency as the between-subject dependent variable. For all five areas, there was a statistically significant interaction effect between Transparency and own–population behavior (all p < 0.001). For Diet, Smoking, Alcohol and Finance, there was an increased discrepancy in effectiveness on own behavior versus on population behavior in opaque BIs compared to the transparent BIs. However, for Exercise, the difference was the other way around: participants rated the transparent BI as more likely to be effective on their own behavior than on population behavior, and this discrepancy decreased in the opaque BI (see Table 6).

General discussion and conclusions

We found that transparent BIs were more acceptable than opaque BIs (Hypothesis 1), and this result held across all contexts taken individually, except Exercise. The type of argument given affected the assessment of the BI, with BIs presented alongside positive arguments rated as more acceptable than those presented alongside negative or a mix of positive and negative arguments (Hypothesis 3), but this was a small effect, and the only two contexts that showed this effect individually were Alcohol and Finance. BIs that are implemented by researchers were judged as being slightly more acceptable

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Table 6. Likelihood of effectiveness of the behavioral intervention on own and population behavior and the discrepancy between them (effectiveness on own behavior – effectiveness on population behavior) for transparent and opaque behavioral interventions.

Policy domain	Transparency	Own behavior, mean (*SD*)	Population behavior, mean (*SD*)	Discrepancy (own behavior – population behavior) (*SE*)	P-value
Diet	Transparent	5.64 (2.1)	5.57 (1.7)	0.07 (0.06)	0.254
	Opaque	5.29 (2.3)	5.70 (2.0)	–0.41 (0.06)	<0.001
Exercise	Transparent	5.53 (2.3)	5.13 (1.8)	0.41[a] (0.06)	<0.001
	Opaque	4.52 (2.4)	4.44 (2.1)	0.07 (0.06)	0.194
Alcohol	Transparent	3.95 (2.2)	4.50 (1.8)	–0.04 (0.06)	0.485
	Opaque	4.08 (2.3)	4.04 (2.1)	–0.55 (0.06)	<0.001
Smoking	Transparent	4.52 (2.5)	5.15 (1.9)	–0.25 (0.07)	<0.001
	Opaque	5.07 (2.4)	5.32 (1.9)	–0.64[a] (0.07)	<0.001
Finance	Transparent	3.08 (2.3)	3.71 (1.9)	–0.14 (0.05)	0.006
	Opaque	4.54 (2.1)	4.68 (2.1)	–0.63 (0.05)	<0.001

[a] These are the correct mean differences, even though they are not the differences between the numbers in the preceding columns; the discrepancies are caused by rounding to two decimal places.

than BIs implemented by governments (Hypothesis 2), but this was such a small effect that it is not meaningful. There was no interaction effect between the designer of the BI and the type of argument given, contra Hypothesis 4. On average, all of the BIs were considered acceptable for changing participants' own behavior (with mean acceptability ratings above the mid-point of the scale), except for the opaque BI in the Finance context; there was differential acceptability of BIs across contexts, with Finance clearly least acceptable.

As well as finding transparent BIs more acceptable than opaque BIs, our participants regarded them as more likely to result in positive behavior change. Furthermore, the effectiveness of the BIs was at least as influential a predictor of acceptability ratings as the ease of identification of the behavior change mechanism across the five contexts (and considerably more influential in some, especially Finance and, interestingly, amongst smokers in the context of smoking cessation). There was a direct effect of ease of identification on acceptability – except, notably, for smokers when asked about BIs that discourage smoking – which we had expected given H1. This is consistent with arguments that people care about having a sense of agency over their actions (Osman, 2014) and past findings that people view opaque BIs as more autonomy-threatening than transparent ones (Jung & Mellors, 2016). However, the likelihood that the BI would result in positive behavior change had more predictive power than Ease of Identification in all contexts except Exercise. Bang *et al.* (2018) found mixed results on this point, with their study 1 finding no relationship between effectiveness for self and acceptability, but their study 2 finding that the expected effectiveness of a change in choice architecture on one's own behavior predicted the acceptability of the change. Our results are consistent with those of their study 2. Our finding that people predict that the BI will be more likely to be effective on the population as a whole than on their own behavior is also consistent with the Bang *et al.* (2018) finding that BIs will be more effective for others than for themselves.

It is not surprising that people care about both transparency and effectiveness – this is an obvious prediction that is also consistent with previous results (e.g., Sunstein, 2016; Arad & Rubinstein, 2018). However, the relative importance of effectiveness is at odds with the theoretical focus on the acceptability of transparency. For example, a parliamentary report in the UK identified the extent to which a BI is covert as one of two criteria that should bear on its acceptability (House of Lords, Science and Technology Select Committee, 2011). (The other being the extent to which the BI is popular with the public.) Our results are consistent with another survey study whose authors also drew conclusions from average ratings: Petrescu *et al.* (2016) tested the hypothesis that stating that interventions work via non-conscious processes decreases their acceptability. They found no evidence to support the hypothesis, but they did find that the effectiveness of the BI was a predictor of acceptability (Petrescu *et al.*, 2016). The authors of a qualitative study also reported that interviewees had very limited concerns regarding the manipulative aspects of BIs (Junghans *et al.*, 2015). It is possible that transparency is a strong concern for a minority of people; for instance, Arad and Rubinstein (2018) found that a minority of their subjects reported an opposition to BIs, and this was driven by concerns about manipulation and the fear of a 'slippery slope' to non-consensual interventions. It is also possible that the use of survey methods decreases the impact of transparency, and if we had conducted a vignette study, then transparency would have been a more influential predictor

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

of acceptability, since getting participants to imagine being in the situation would have simulated the feeling of being manipulated.

Our finding that people rated transparent BIs as more effective than opaque BIs is also surprising. In academic debates on the acceptability of BIs, it has been assumed that transparency and effectiveness pull in different directions (Bovens, 2009; House of Lords, Science and Technology Select Committee, 2011). Bovens (2009, pp. 209, 217) says that these techniques "work best in the dark." However, the participants in our sample did not seem to agree with that. (Ditto the participants of Jung & Mellers, 2016, who found that transparent, System 2 BIs were viewed as more effective for changing behavior.) It could be that being able to identify a mechanism made it seem more likely to our participants that the BI would be effective, or there could be a halo effect whereby a more acceptable BI is generally judged to have more of other desirable properties as well. However, it seems that our participants' folk psychology is right, since there are now several studies showing that disclosure does not affect effectiveness, most of which concerned defaults (Loewenstein *et al.*, 2015; Steffel *et al.*, 2016; Bruns *et al.*, 2018), but one of which concerned the placement of food items in a snack shop (Kroese *et al.*, 2015).

In our study, effectiveness was partially mediated by the desire to change behavior through the BI. In other words, when people believed that a BI would be effective, then they wanted to use it to change their behavior. This supports the contention that people do want to achieve positive behavior change and they support BIs that will help them to do that; there is a sense in which people find BIs acceptable because the interventions will make them better off as judged by themselves.

We found a general lack of support for Hypotheses 2–4. There was only a very small effect of Argument, a negligible effect of Designer and there was no evidence of the predicted interaction effect between Argument and Designer. The low impact of Argument (and lack of interaction effect) is less surprising when we consider that the Argument manipulation did not have much impact on effectiveness ratings. The lack of substantial impact of the designer of the BI is more surprising given that scientists are more trusted than governments: an Ipsos Mori (2018) survey found that 85% of British adults trust scientists to tell the truth compared to 19% for politicians and 16% for advertising executives, and a previous study showed that people trust BIs that are developed and proposed by scientists more than those that are developed and proposed by governments (Osman *et al.*, 2018). Osman *et al.* (2018) explained their result with reference to research on 'source credibility' in the psychology of communication, which is the idea that people are more receptive when there is a good fit between the area of expertise of

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

the communicator (in this case, the proposer of the BI) and the topic of communication (in this case, the BI being proposed); for a review of the literature on source credibility, see Pornpitakpan (2004). It may be that our subjects did not see any differential in expertise, since we stressed that the advertisers and researchers were 'top' of their field, and there were only negligible effects of the designer on ratings of whether the BI would positively change behavior. The idea that we would see an interaction effect between Designer and Argument was predicated on creating ambiguity, but as well as being unsuccessful at creating ambiguity about the likely effectiveness of the BI, we probably did not create ambiguity around the intention of the designer, since we had also stated that the aim of the BI was to promote positive behavior change. If participants thought that all three designers were equally effective and well-intentioned, then there would be no reason for there to be an effect of Designer on their judgments.

We found that using BIs in Finance was less acceptable than using BIs in other contexts. The mean acceptability of the health-related BIs ranged from 7.28 to 6.53, while the mean acceptability of Finance-related BIs was only 4.72. Our results in the four health-related contexts are consistent with evidence that people approve of BIs for health behavior (Junghans *et al.*, 2015; Reisch *et al.*, 2017). We can only speculate about why our Finance-related BIs were less acceptable, since we had only a single pair of transparent and opaque BIs in each context, and the BIs were not matched (matching was not possible given that we used BIs that had actually been implemented). The opaque Finance-related BI was also the only BI that used a default. However, we do not think that it was the default alone that caused the low ratings, since there are field experiments whose results show that people approve of having their own behavior changed by BIs using defaults. For instance, after being exposed to a BI that presented the vegetarian meal as a default when registering for a conference – increasing the number choosing the vegetarian meal from 13% to 89% – 90% of those exposed said that they approved of changing the default (Hansen *et al.*, 2019). In addition, after an intervention that changed the default positions of sit–stand desks in a workplace to the standing position, increasing the rate of standing from 1.8% to 13.1%, 56.5% of employees said that it was acceptable to be unconsciously influenced in this way (Venema *et al.*, 2018).[4]

[4] In contrast, Felsen *et al.* (2013) found a lack of approval for defaults in organ donation, but this does not affect the point that the acceptability of defaults is context-dependent and that in at least some contexts they are acceptable.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

In our multiple-choice follow-up question, we found that Finance was a clear outlier, with the majority of participants saying that psychological methods should not be used in that context. We suspect that there are features that differentiate health from finance, which made the finance-related BIs less acceptable. In health, everyone can agree that, for example, high-sugar and high-fat products are unhealthy. However, in finance, the best product for someone depends on their attitudes to risk. Even though a traffic-light rating marks the riskiest products as red, those may be the most appropriate products for some people; ditto a default product may not be the best choice for everyone.[5] Therefore, people may be more skeptical that BIs in finance will actually promote their wellbeing. In our study, people considered that BIs were least likely to have a positive effect in Finance (in fact, with a mean of 3.84, the rating of BIs in Finance was on the non-effective side of the scale), and the differential predictive power of Effectiveness and Ease of Identification on Acceptability ratings was particularly striking for Finance. In the regression models, the coefficient on Effectiveness was higher for Finance than for any other context. So our participants had a high level of concern regarding the Effectiveness of the Finance BI, but did not think that it would have positive effects.

Another reason why participants may have been dubious about the effectiveness of the Finance BI is that they might have been worried about whether the default would benefit them if it was influenced by industry, as the bank might wish to default them into an option that would be profitable for the bank. Since the financial crisis, attitudes to the financial services industry have become more negative (Bennett & Kottasz, 2012), and some authors have concluded that there is now a crisis of trust in that sector (Bachmann *et al.*, 2011; Sapienza, & Zingales, 2012). This lack of trust may also help explain why people tend to think that finance-related BIs are less likely to lead to positive changes than health-related interventions, and why they find finance-related BIs less acceptable. This is ironic because financial literacy around retirement saving and pension plans is low (Lusardi & Mitchell, 2011a, 2001b), suggesting that there is scope for using BIs to improve outcomes in this area.

[5] Of course, a default in health may not be best for absolutely everyone at all times. For instance, a diabetic with hypoglycemia does need high-sugar food, or an anorexic may need to eat more in general. However, it is no coincidence that these are both clinical conditions. The vast bulk of the population needs to eat well on balance, which can be done by following healthy eating guidelines. In contrast, in finance, all we can say is that the vast bulk of the population needs retirement products. But the best way of saving for retirement will show a lot of individual variation due to differing risk preferences and other factors, such as longevity risk.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/10.1017/bpp.2020.6

Financial support

This work was supported by Natalie Gold's funding from the European Research Council, under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 283849. This research was also funded by Queen Mary University of London Life Sciences Initiative Studentship (LSIPGRS).

References

Arad, A. and A. Rubinstein (2018), The people's perspective on libertarian-paternalistic policies, *The Journal of Law and Economics*, **61**(2): 311–333.

Bachmann, R., N. Gillespie and R. Kramer, 2011. Trust in crisis: Organizational and institutional trust, failures and repair, *Organization Studies*, **32**(9), 1311–1313.

Bang, H. M., S. B. Shu and E. U. Weber (2018), The role of perceived effectiveness on the acceptability of choice architecture, *Behavioural Public Policy*, 1–21.

Baron, R. M. and D. A. Kenny (1986), The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations, *Journal of personality and social psychology*, **51**(6): 1173.

Blau, J. H. (1975), Liberal values and independence, *The Review of Economic Studies*, **42**(3): 395–401.

Bennett, R. and R. Kottasz (2012), Public attitudes towards the UK banking industry following the global financial crisis, *International Journal of Bank Marketing*, **30**(2): 128–147.

Bovens, L. (2009), The ethics of nudge. In T. Grüne-Yanoff and S. O. Hansson. (Eds.). (2009), Preference change: Approaches from philosophy, economics and psychology (Vol. 42). Springer Science & Business Media (pp. 207–219).

Branson, C., B. Duffy, C. Perry and D. Wellings (2012), Acceptable behaviour: Public opinion on behaviour change policy. *Ipsos MORI, London.*

Bruns, H., E. Kantorowicz-Reznichenko, K. Klement, M. L. Jonsson and B. Rahali (2018), Can nudges be transparent and yet effective?, *Journal of Economic Psychology*, **65**, 41–59.

Cornwell, J. F. and D. H. Krantz (2014), Public policy for thee, but not for me: Varying the grammatical person of public policy justifications influences their support, *Judgment and Decision Making*, **9**(5): 433.

Diepeveen, S., T. Ling, M. Suhrcke, M. Roland and T. M. Marteau (2013), Public acceptability of government intervention to change health-related behaviours: a systematic review and narrative synthesis, *BMC public health*, **13**(1): 756.

Felsen, G., N. Castelo and P. B. Reiner (2013), Decisional enhancement and autonomy: public attitudes towards overt and covert nudges, *Judgement and Decision Making*, **8**(3): 202. https://doi.org/10.1177/2332858416674007

Gold, N. (2018), The Origins of Behavioural Public Policy, Adam Oliver. Cambridge University Press, 2017, 252 pages. *Economics & Philosophy*, **34**(2): 267–274.

Hagman, W., D. Andersson, D. Västfjäll and G. Tinghög (2015), Public views on policies involving nudges, *Review of Philosophy and Psychology*, **6**(3): 439–453.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Hansen, P. G. and A. M. Jespersen (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.

Hansen, P. G., M. Malthesen and M. Schilling (2019), Nudging healthy & sustainable food choices: Three field experiments using a vegetarian lunch-default as a normative signal. *Journal of Public Health*.

House of Lords, Science and Technology Select Committee (2011), Behaviour change (Second report). London, United Kingdom

Ipsos Mori (2018), *Ipsos-Mori Veracity Index*, https://www.ipsos.com/sites/default/files/ct/news/documents/2018-11/veracity_index_2018_v1_161118_public.pdf, accessed on 9th June 2019.

Jung, J. Y. and B. A. M. Mellers (2016), American attitudes toward nudges, Judgment and Decision Making. *Judgement and Decision Making*, **11**(1): 62–74.

Junghans, A. F., T. T. Cheung and D. D. De Ridder (2015), Under consumers' scrutiny-an investigation into consumers' attitudes and concerns about nudging in the realm of health behavior, *BMC public health*, **15**(1): 336.

Kroese, F. M., D. R. Marchiori and D. T. de Ridder (2015), Nudging healthy food choices: a field experiment at the train station, *Journal of Public Health*, **38**(2): e133-e137.

Lin, Y., M. Osman and R. Ashcroft (2017), Nudge: Concept, Effectiveness, and Ethics, *Basic and Applied Social Psychology*, **39**(6): 293–306. https://doi.org/10.1080/01973533.2017.1356304

Loewenstein, G., C. Bryce, D. Hagmann and S. Rajpal (2015), Warning: You are about to be nudged, *Behavioral Science & Policy*, **1**(1): 35–42.

Lusardi, A. and O. S. Mitchell (2011a), *Financial literacy and planning: Implications for retirement wellbeing* (No. 17078). National Bureau of Economic Research. http://www.nber.org/papers/w17078

Lusardi, A. and O. S. Mitchell (2011b), Financial literacy around the world: an overview, *Journal of Pension Economics and Finance*, **10**(04): 497–508.

OECD (2017), Behavioral Insights and Public Policy: Lessons from Around the World, OECD Publishing, Paris. https://dx.doi.org/10.1787/9789264270480-en

Oliver, A. (Ed.). (2013), *Behavioural public policy*, Cambridge University Press.

ONS (2018) Adult smoking habits in Great Britain. Office of National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2017 Accessed 11 June 2019.

Osman, M. (2014), *Future-minded: The Psychology of Agency and Control*, New York: Palgrave Macmillan.

Osman, M., N. Fenton, T. Pilditch, D. Lagnado and M. Neil (2018), Whom Do We Trust on Social Policy Interventions?, *Basic and Applied Social Psychology*, **40**(5): 249–268.

Osman, M., Y. Lin and R. Ashcroft (2017), Nudging: A lesson in the theatrics of choice, *Basic and Applied Social Psychology*, **39**(6): 311–316.

Pechey, R., P. Burge, E. Mentzakis, M. Suhrcke and T. M. Marteau (2014), Public acceptability of population-level interventions to reduce alcohol consumption: a discrete choice experiment, *Social science & medicine*, **113**, 104–109.

Petrescu, D. C., G. J. Hollands, D. L. Couturier, Y. L. Ng and T. M. Marteau (2016), Public acceptability in the UK and USA of nudging to reduce obesity: the example of reducing sugar-sweetened beverages consumption, *PLoS One*, **11**(6): e0155995.

Pornpitakpan, C. (2004), The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence, *Journal of Applied Social Psychology*, **34**(2), 243–281.

Reisch, L. A. and Sunstein, C. R. (2016), Do Europeans like nudges? *Judgement and Decision Making*, **11**(4): 310–325.

https://doi.org/10.1017/bpp.2020.6 Published online by Cambridge University Press

Reisch, L. A. Sunstein, C. R. and W. Gwozdz (2016), Better than a Whip? European Attitudes toward Health Nudges, *Food Policy*, **69**, 1–10. https://doi.org/10.1016/j.foodpol.2017.01.007

Reisch, L. A., C. R. Sunstein and W. Gwozdz (2017), Beyond carrots and sticks: Europeans support health nudges, *Food Policy*, **69**, 1–10.

Sapienza, P. and L. Zingales (2012), A trust crisis, *International Review of Finance*, **12**(2): 123–131.

Sen, A. (1970), The impossibility of a Paretian liberal, *Journal of political economy*, **78**(1): 152–157.

Steffel, M., E. F. Williams and R. Pogacar (2016), Ethically deployed defaults: Transparency and consumer protection through disclosure and preference articulation, *Journal of Marketing Research*, **53**(5): 865–880.

Sugden, R. (2017), Do people really want to be nudged towards healthy lifestyles?, *International Review of Economics*, **64**(2): 113–123.

Sugden, R. (2018), 'Better off, as judged by themselves': a reply to Cass Sunstein, *International Review of Economics*, **65**(1): 9–13.

Sunstein, C. R. (2016), People prefer system 2 nudges (kind of), *Duke Law Journal*, **66**, 121–168.

Sunstein, C. R. (2018), 'Better off, as judged by themselves': a comment on evaluating nudges, *International Review of Economics*, **65**(1): 1–8.

Tannenbaum, D., C. R. Fox and T. Rogers (2017), On the misplaced politics of behavioural policy interventions, *Nature Human Behaviour*, **1**(7): 0130.

Thaler, R. H. and C. R. Sunstein (2008), Nudge: improving decisions about health. *Wealth, and Happiness*. Penguin: New York.

Venema, T. A., F. M. Kroese and D. T. De Ridder (2018), I'm still standing: A longitudinal study on the effect of a default nudge, *Psychology & Health*, **33**(5): 669–681.

World Bank (2015), *World Development Report 2015: Mind, Society, and Behavior*, Washington, DC: World Bank. doi: 10.1596/978-1-4648-0342-0.